I Workshop sobre Innovación en Centros Educativos y de Investigación (I WICEI)

Reporte Breve de Investigación

Explotación y visualización de información. Su aplicación a las industrias rionegrinas

Autores: Gabriela Anahí Cayu[1]; Patricio Nicolás Castro[1]; Gustavo Agüero[1]; Camila Carrera[1]; Federico Ezequiel Di Fabio[1]; Mauricio Tassara[1]; Mauro Germán Cambarieri[1]; Pablo Enrique Argañaras[2]; Martín René Vilugrón[2]; G. Balbarrey[1]; Paola Verónica Britos[1,2]

[1]Laboratorio de Informática Aplicada UNRN Sede Atlántica, Argentina

[2]Laboratorio de Informática Aplicada UNRN Sede Andina, Argentina

{gcayu,pcastro,gaguero,ccarrera,fdifabio,mtassara,mcambarieri,parganaras,mrvilugron,gbalbarrey,pbritos}@unrn.edu.ar

Eje temático: I + D + i

Cantidad de páginas: 7

Resumen

La Ingeniería de la Explotación de Información por una parte se enfoca en los métodos y procesos para encontrar patrones a partir de masas de información; un componente importante de éstos es la forma en la cual los patrones se visualizan. Es sobre este aspecto sobre el cual produce aportaciones la visualización de información (datos) y la visión de computador. Sin embargo, no existen métodos y/o procesos que formalmente articulen la forma en la cual la visualización de datos y la visión por computador permiten reconocer los patrones de conocimiento emergentes de los procesos de explotación. En este contexto, este proyecto busca explorar los procesos/métodos que faciliten la visualización de patrones de conocimientos usando visualización de datos en el marco de proyectos de Explotación de Información en las industrias rionegrinas. El proyecto será validado con diversos casos dentro del marco de trabajo en conjunto con organismos nacionales e internacionales.

Palabras clave: Big data; visualización; tratamiento de imágenes; minería de datos

Contexto

El presente proyecto de investigación, aprobado y acreditado por Resolución Rectoral Nº 0709-17 de la Universidad Nacional de Río Negro el 13 de julio de 2017, con código PI 40-C-542 y duración trienal, tiene como directora a la Dra. P.V. Britos, y como integrantes a G. P. Balbarrey, M. Tassara, G. A. Cayú, P. N. Castro, C. Carrera, F. E. Di Fabio, G. Agüero, P. E. Argañaras, M. G. Cambarieri y M. R. Vilugrón, donde 2 de ellos son investigadores formados, 3 investigadores en formación son alumnos avanzados de la carrera de Licenciatura en Sistemas, y 3 son becarios CIN. En este marco se desarrollan 1 tesis de Doctorado, 2 tesis de Maestría y 3 Trabajos Finales de Grado.

La evaluación del proyecto estuvo a cargo de pares del banco de Evaluadores con categoría I y II del Programa de Incentivos a los Docentes-Investigadores del Ministerio de Educación y Deportes de la Nación.

Introducción

Explotación de información

El término de minería de datos está fuertemente vinculado al concepto de grandes bases de datos y se remonta a la definición de algoritmos de búsqueda de patrones de conocimiento [Maimon, O., Rokach, L. (Eds.), 2005]. Sin embargo, hoy en día existen líneas de investigación en campos tales como minería de textos [Tan, A., 1999], minería de imágenes [Hsu, W., Lee, M., Zhang, J., 2002], minería de flujos de información [Gaber, M., Zaslavsky, A., Krishnaswamy, S., 2010], minería web [Kosala, R., Blockeel, H., 2000], entre otros. Por lo que nuestra línea de trabajo considera apropiado utilizar el término “Explotación de Información” [Gopal et al., 2011] como una referencia genérica a cualquiera de los tipos de minería antes mencionados.

Un Proceso de Explotación de Información se define [Britos, 2008; García Martínez et al., 2003] como un grupo de tareas relacionadas lógicamente [Curtis et al., 1992] que, a partir de un conjunto de información con un cierto grado de valor para la organización, se ejecuta para lograr otro, con un grado de valor mayor que el inicial [Ferreira et al., 2005; Kanungo, 2005].

Basado en el concepto de Ingeniería de Software, que se ha definido en SWEBOK [Abran et al., 2004] como "la aplicación de un enfoque sistemático, disciplinado y cuantificable para el desarrollo, operación y mantenimiento de software, y el estudio de estos enfoques, es decir, la aplicación de la ingeniería al software"; se acuerda definir el concepto de Ingeniería de Explotación de Información (IEI) [García-Martínez et al., 2011] como la aplicación de un enfoque sistemático, disciplinado y cuantificable para el desarrollo de Proyectos de Explotación de Información y el estudio de estos enfoques, es decir, la aplicación de la Ingeniería a la Explotación de Información.

Desde esta perspectiva, la Ingeniería de Explotación de Información [García-Martínez, et al., 2016] se centra en la definición de procedimientos para guiar el desarrollo de un proyecto de explotación de información. Su principal objetivo es identificar patrones interesantes y piezas de conocimiento relevantes para la organización, asegurar su correcta comprensión y proporcionar apoyo fiable para el proceso de toma de decisiones. Para lograr este objetivo es necesario establecer un conjunto de actividades que proporcionen la estructura general del proyecto, apoyando el proceso de desarrollo [Britos, P. & García-Martínez, R. 2009; Schiefer, J. et al., 2004].

Cabe aclarar que los proyectos de Ingeniería de Explotación de Información poseen características muy distintas a los proyectos de desarrollo de software tradicional [Martins et al., 2014a]. La diferencia se presenta en los procesos de desarrollo y mantenimiento en los cuales el ciclo de fases de un proyecto de software tradicional — inicio, requisitos, análisis y diseño, construcción, integración y pruebas — no resultan naturales en un proyecto de explotación de información [Vanrell et al., 2012]. Por otra parte, al evaluar las principales metodologías existentes para los proyectos de explotación de información [Chapman et al., 2000; SAS, 2008; Pyle, 2003], se ha observado la falta de herramientas que permitan soportar de forma completa la fase de administración de proyectos.

Visualización de información

Por su parte, la visión por computador, también conocida como visión artificial, es una rama de la inteligencia artificial que permite simular el sistema visual humano, extrayendo información

a partir de imágenes digitales [Szeliski, R. 2011]. La misma se utiliza en distintas áreas, como la industria, la medicina, la seguridad, la automotriz, el deporte. Existe una tendencia importante en los últimos años sobre su utilización en videojuegos.

Se pueden mencionar cuatro etapas principales que comprenden un sistema de visión por computador:

1) Captura o adquisición de imágenes digitales. [Alegre et al. 2003]
2) Procesamiento digital de las imágenes. [Gonzalez, R. C., Woods, R. E. 2008; Gonzalez, R. C., Wintz, P. 1996; Gonzalez, R. C. et al. 2009]
3) Segmentación. [Acharya, T., Ray, A. K. 2005]
4) Reconocimiento o clasificación.

Se destaca que estas etapas no se realizan siempre de forma secuencial, por lo que se puede volver a una etapa anterior, realimentando las mismas hacia atrás.

Por otra parte, la visualización es la formación de imágenes visuales. Como lo define J. Foley, es el mapeo de datos en representaciones que pueden ser percibidas. Los tipos de mapeo pueden ser visuales, auditivos, táctiles, etc. o una combinación de éstos [Foley, J., Van Dam, A. F. 1992]. La visualización no es un fenómeno nuevo. El hombre ha utilizado estas técnicas desde hace miles de años para entender mejor su medio ambiente. La visualización por computadora es un proceso de mapeo de las representaciones hechas por la computadora a representaciones perceptuales, eligiendo técnicas de codificación para maximizar el entendimiento y comunicación con los seres humanos [Foley, J., Van Dam, A, F, 1992]. Básicamente, la visualización nos permite interpretar datos que se obtienen de investigaciones matemáticas o científicas. Se utilizan los sistemas computacionales no para simular, sino para representar estos datos.

La tecnología de visualización es una integración de las áreas de graficación, procesamiento de imágenes, visión computacional, modelado geométrico, diseño asistido por computadora, psicología perceptual, estudios de interfaces de usuarios, etc. Por lo tanto, las personas que se encargan del estudio de sistemas de visualización deben contar con conocimientos en áreas como diseño gráfico, ciencias, matemáticas, graficación por computadora y animación.

Hay tres partes importantes en un sistema de visualización:

- Construcción de un modelo empírico de los datos: Este modelo puede tener consideraciones sobre teoría del muestreo, como el teorema Nyquist, y esquemas de interpolación matemática. También debemos tomar en cuenta la probabilidad de que haya errores en los datos.
- Selección de esquemas: Significa tomar como modelo un objeto de visualización abstracta (un mapa, por ejemplo). Puede ser en 2 o 3 dimensiones, estáticos o interactivos.
- La representación de la imagen en un ambiente gráfico. Puede ser en 2 o 3 dimensiones, estáticos o interactivos.

Relevancia del problema

Una de las lecciones aprendidas sobre desarrollo de software en Informática derivada de los estadios tempranos de la disciplina, es que la ausencia de una ingeniería de software

conllevaba a un desarrollo artesanal de los artefactos software [Böehm, 1981; Pressman, 2004; Pfleeger y Atlee, 2005; Ochoa et al., 2008; Van Vliet, 2008]. El desarrollo artesanal implicaba la imposibilidad de establecer, dentro de valores racionales, parámetros tales como: [a] cantidad y calificación de los recursos humanos a emplear en el proyecto, [b] tiempos de desarrollo, [c] modelos de proceso que guiaran el desarrollo y permitieran establecer hitos de entrega, [d] formalismos de documentación que dieran cuenta de lo hecho en el proyecto de desarrollo del artefacto software y de las decisiones de diseño asumidas, constituyendo el punto de partida para futuras ampliaciones de funcionalidades, [e] modelos de costo de proyecto [Böehm, 1981], entre otros. De hecho, la estimación de estos parámetros se hacía en acuerdo a la experiencia de individuos sin ninguna base ingenieril y lo que la estimación para un grupo de desarrollo podía hacerse en meses, para otro podía hacerse en años. La explotación de información está en sus primeros estadios y, al igual que lo que ocurría con el desarrollo de artefactos software, adolece de una ingeniería que provea instrumentos para un proceso de decisión-estratégica / control gestión / desarrollo de proyectos de este tipo.

Metodología

El desarrollo de este proyecto utilizará la metodología propia de la investigación documental, del estudio de casos, de técnicas de análisis comparativo y de síntesis de comparaciones. Con base en que:

1. Se producirá un relevamiento en:
    1.1. Técnicas de educción y representación de requerimientos usuales en el marco de la Ingeniería de Requisitos e Ingeniería del Conocimiento.
    1.2. Técnicas de visión por computador.
    1.3. Técnicas de visualizaciones.
2. Se explorará la forma en la cual los requisitos educidos pueden ser representados en los formalismos para el uso en Big Data. A tal efecto se realizará:
    2.1 Identificación de casos de estudio de educción y representación de requisitos.
    2.2 Estudio comparado de la utilización de las diversas técnicas de representación de conocimientos.
    2.3 Formalización de los casos de estudio identificados utilizando las técnicas previamente seleccionadas de representación de conocimientos en diversos dominios.
    2.4 Análisis de las ventajas y desventajas de representar los requisitos educidos mediante técnicas aplicadas.
    2.5 Estudio comparativo de diversas técnicas de visualización de los mismos.
    2.6 Análisis de ventajas y desventajas de la representación de las visualizaciones de los requisitos educidos.
    2.7 Estudio comparativo de diversas técnicas de visión por computador.
    2.8 Análisis de ventajas y desventajas de las diversas técnicas de visión por computador.

3. Posteriormente a los pasos indicados, se procederá a la aplicación de los diversos formalismos identificados:

    3.1 Para la detección de patrones de conocimientos y reglas de negocios en diversos dominios de conocimiento.

    3.2 Para la visualización de grandes masas de datos.

    3.3 Para el tratamiento de datos a través de la visión por computador.

Desarrollo

El objetivo general que guía a este proyecto es *Articular integradamente mecanismos de proceso de interpretación de grandes masas de información a través de diversas técnicas de tratamiento de datos y su visualización*. Y los objetivos específicos asociados son:

- Articular, dentro del subproceso de control y gestión del Modelo de Proceso para Proyectos de Explotación de Información desarrollado por el grupo de trabajo, los artefactos: [a] Modelo de Proceso para Educción de Requerimientos de Proyectos de Explotación de Información. [b] Test de Viabilidad y Modelo de Estimación de Esfuerzo. [c] Métricas para Proyectos de Explotación de Información. [d] Construcción de las herramientas que permitan llevar adelante este proceso.
- Desarrollar una metodología de visualización de datos científicos que se apoye en la escritura de código y en el empleo de tecnología de cómputo para generación de visuales de punta para favorecer la comprensión y comunicación de datos complejos.
- Investigar, desarrollar, diseñar y producir nuevos modelos de visualización y simulación de datos científicos.

Justificación

En este momento nos encontramos en la primera etapa del proyecto, realizando el relevamiento de Casos de Estudio con investigación documental en las distintas áreas involucradas y con identificación de dominios de conocimiento de los que se disponen bases de datos y documentación para cubrir aspectos documentables, de técnicas de visualización y de técnicas de visión por computador.

El avance en un caso de estudio en particular dio como resultado una tesis de grado de Licenciatura en Sistemas [Cayú G., 2017] y publicaciones asociadas [Cayú G., et al., 2016], dejando abierta "la posibilidad de la identificación de los productores de los cuales provienen las muestras de mieles, y analizar conjuntos de datos obtenidos en otra época del año". (Cayú G., 2017, p.112)

Referencias

Abran, A., Moore, J. W., Bourque, P., Dupuis, R., Tripp, L. (2004). Guide to the Software Engineering Body of Knowledge (2004 version). IEEE. ISBN 0-7695-2330-7.

Acharya, T., Ray, A. K. (2005). Image processing: principles and applications. John Wiley & Sons.

Alegre, E., Sánchez, L., Fernández, R. Á., Mostaza, J. C. (2003). Procesamiento Digital de Imagen: fundamentos y prácticas con Matlab. Universidad de León. ISBN 84-9773-052-6.

Böehm, B. (1981). Software Engineering Economics. Prentice Hall. ISBN 0-13-822122-7

Britos, P. & García-Martínez, R. (2009). Propuesta de Procesos de Explotación de Información. Proceedings XV Congreso Argentino de Ciencias de la Computación. Workshop de Base de Datos y Minería de Datos, pp. 1041-1050. ISBN 978-897-24068-4-1

Britos, P. (2008). Procesos de Explotación de Información basados en Sistemas Inteligentes. Tesis Doctoral. Universidad Nacional de La Plata, Facultad de Informática. La Plata. Argentina. Disponible en: http://www.iidia.com.ar/rgm/tesistas/td-pb-fi-unlp.pdf. Último Acceso: Enero de 2016.

Cayú, G. A., Agüero, G. A., Balbarrey, G. P., Cabrera, M. M., Carrera, C., Britos, P., & Vivas, H. L. (2016). Building honey-based territorial identity for the Formosa Monte through information exploitation using intelligent systems. IEEE CACIDI 2016 - IEEE Conference on Computer Sciences, doi: 10.11097CACIDI.2016.7785980

Cayú, G. (2017). *Identificación de Origen en Mieles a través de una Metodología de Explotación de Información* (tesis de grado). Universidad Nacional de Río Negro, Viedma, Río Negro, Argentina.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Curtis, B., Kellner, M., Over, J. (1992). Process Modelling. Communications of the ACM, 35(9): 75-90.

Ferreira, J.; Takai, O; Pu, C. (2005). Integration of Business Processes with Autonomous Information Systems: A Case Study in Government Services. Proceedings Seventh IEEE International Conferenceon E-Commerce Technology. 471, 474.

Foley, J., Van Dam, A. (1992). Fundamentals of Interactive Computers Graphics, Addison-Wesley, Reading, Massachusetts, segunda edición, 1992.

Gaber, M., Zaslavsky, A. Krishnaswamy, S. (2010). Data stream mining. En Maimón, O. and Rokach, L. eds. Data mining and knowledge discovery handbook. Springer, pp. 759-787.

García Martínez, R., Servente, M. y Pasquini, D. (2003). Sistemas Inteligentes. Editorial Nueva Librería. ISBN 987-1104-05-7.

García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo, F., Rodríguez, D., Pytel, P., Vanrell. J. (2011). Towards an Information Mining Engineering. In Software Engineering, Methods, Modeling & Teaching. Medellín UniversityPress. ISBN 9789588692326. pp. 83-99.

García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo-Cattaneo, F., Rodríguez, D., Pytel, P. y Vanrell. J. (2011a). Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching. Sello Editorial Universidad de Medellín, pp. 83–99. ISBN 978–958–8692–32–6.

García-Martínez, R., Britos, P., Martins, S., Baldizzoni, E. (2016). Explotación de Información. Ingeniería de Proyectos. Editorial Nueva Librería ISBN 978-987-1871-34-6

Gonzalez, R.C. & Woods, R.E. (2008). Digital Image Processing 3rd edition. Prentice-Hall

Gonzalez, R.C., Wintz, P. (1996). Procesamiento digital de imágenes. Addison-Wesley.

Gopal, R., Marsden, J., Vanthienen, J. (2011). Information mining: Reflections on Recent Advancements and the Road Ahead in Data, Text, and Media Mining. Decision Support Systems, 51(4): 727-731.

Hsu, W., Lee, M., Zhang, J. (2002). Image mining: Trends and developments. Journal of Intelligent Information Systems, 19(1): 7-23.

Kanungo, S. (2005). Using Process Theory to Analyze Direct and Indirect Value-Drivers of Information Systems. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, pp. 231-240.

Kosala, R., Blockeel, H. (2000). Web mining research: A survey. ACM SIGKDD Explorations Newsletter, 2(1): 1-15.

Martins, S., Pesado, P., García-Martínez, R. (2014). Propuesta de Modelo de Procesos para una Ingeniería de Explotación de Información: MoProPEI Revista Latinoamericana de Ingeniería de Software, 2(5): 313-332, ISSN 2314-2642

Maimon, O., Rokach, L. (Eds.). (2005). Data mining and knowledge discovery handbook. Springer.

Ochoa, A. F. E., Britos, P., García-Martínez, R. (2008). Metodologías de Ingeniería Informática. Editorial Nueva Librería. ISBN 978-987-1104-54-3.

Pfleeger, S., Atlee, J. (2005). Software Engineering: Theory and Practice. 3rd Edition. Prentice Hall.

Pressman, R. (2004). Software Engineering: A Practitioner's Approach. Editorial McGraw Hill.

Pyle, D. (2003) Business Modeling and Business intelligence. Morgan Kaufmann.

Rafael C. González, Richard E. Woods, Steven L. Eddins (2009). Digital image processing using Matlab. González, Woods, & Eddins.

SAS, (2008). SAS Enterprise Miner: SEMMA.

http://www.sas.com/technologies/analytics/datamining/miner/semma.html. Último acceso Junio 2008.

Schiefer, J., Jeng, J., Kapoor, S. y Chowdhary, P. (2004). Process Information Factory: A Data Management Approach for Enhancing Business Process Intelligence. Proceedings 2004 IEEE International Conferenceon E-Commerce Technology, pp. 162-169.

Szeliski, R. (2011). Computer vision: Algorithms and applications. London: Springer.

Tan, A. (1999). Text mining: The state of the art and the challenges. In Proc. PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases. pp. 65-70.

Thomsen, E. (2003). BI's Promised Land. Intelligent Enterprise, 6(4), pp. 21-25.

Van Vliet, H. (2008). Software Engineering: Principles and Practice. Publisher, John Wiley and Sons Ltd.

Vanrell, J. A., Bertone, R., & García-Martínez, R. (2012). A Process Model for Data Mining Projects. Un Modelo de Procesos para Proyectos de Explotación de Información. In Proceedings Latin American Congress on Requirements Engineering & Software Testing LACREST (p. 53).