

Generación automática de interfaces software para el soporte a la toma de decisiones. Aplicación de ingeniería de dominio y machine learning

PLAN DE INVESTIGACIÓN
PROGRAMA DE DOCTORADO EN INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE SALAMANCA

Andrea Vázquez Ingelmo

DIRECTORES:

Francisco José García Peñalvo
Roberto Therón Sánchez

2 de junio de 2019

INTRODUCCIÓN

La introducción de los sistemas de información para dar soporte y automatizar gran diversidad de actividades ha provocado un incremento exponencial del volumen de datos generados. Sin embargo, la posesión de grandes volúmenes de datos no es siempre valiosa en sí misma, los datos adquieren un valor real cuando son analizados y explotados con el objetivo de extraer conocimiento de los mismos (Albright, Winston, & Zappe, 2010). En los últimos años el análisis de datos ha ganado importancia en diversos sectores (Patil & Mason, 2015) con la visión de lograr que la toma de decisiones estratégicas esté arropada por una consistente base informativa (Sharda, Delen, & Turban, 2013).

En el proceso de toma de decisiones pueden tomar parte sujetos con diversos perfiles, pudiendo existir brechas de comunicación entre ellos al no poseer los mismos niveles de conocimientos en lo referente a sus especialidades (Eppler, 2007). Para obtener un máximo rendimiento, es importante que todos los individuos que forman parte de la toma de decisiones comprendan qué implicaciones tienen los datos recolectados y analizados. Sin embargo, las variables influyentes en la toma de decisiones pueden llegar a ser complejas y necesitar de soporte adicional para alcanzar una plena comprensión de los resultados obtenidos.

Uno de los productos software más utilizados para la extracción de conocimiento y explotación de conjuntos de datos lo constituyen los cuadros de mandos o paneles de información (también conocidos como dashboards) (Few, 2006), compuestos por series de recursos gráficos (Keim et al., 2008) y métricas, con el objetivo de mostrar la información de una forma más comprensible (Tuft & Graves-Morris, 2014) y permitir la identificación de patrones o indicadores relevantes para la toma de decisiones.

Sin embargo, el diseño de un dashboard no es un proceso trivial (Few, 2006); es necesario tener en cuenta las necesidades de los usuarios para facilitar la tarea de analizar los resultados, siendo este proceso de búsqueda de necesidades (Patnaik & Becker, 1999), en muchos casos, difuso, debido a la imposibilidad de saber exactamente desde un primer momento qué tipo de métricas o visualizaciones serán más beneficiosas en función del dominio de los datos y los usuarios objetivo.

En estudios actuales, se han identificado retos en el diseño de estas herramientas, como la flexibilidad funcional de los mismos, la necesidad de tener en cuenta el conocimiento o experiencia de los usuarios respecto a las visualizaciones, el impacto social, etc. (Sarıkaya et al., 2018). En efecto, es necesario conocer que los procesos de extracción de información y generación de conocimiento pueden ser distintos para cada persona, añadiéndole el hecho de que los objetivos de cada usuario pueden ser completamente diferentes. Pero crear dashboards personalizados para cada uno de sus potenciales usuarios es inviable, teniendo en cuenta todas las variables que pueden influir en el diseño del mismo y suponiendo un gasto de recursos inconcebible.

El objeto de estudio de esta investigación es la generación automática de dashboards personalizados para el incremento de efectividad de dichas herramientas, así como el incremento de productividad del proceso de desarrollo de las mismas, estableciendo relaciones entre objetivos, preferencias y habilidades concretas de cada usuario con el conjunto de componentes que finalmente formará parte del dashboard.

Para abordar este problema, el uso del enfoque de las familias de software (Pohl, Böckle, & Linden, 2005), combinado con un enfoque dirigido por modelos (Kleppe, Warmer, & Bast, 2003), establece una buena estrategia para plantear la construcción de dashboards potentes, flexibles y mantenibles. La fase de ingeniería de dominio de este paradigma posibilita la identificación de las partes del sistema base que necesitarán ser configurables para adaptarse a los requisitos de un producto específico de la familia, siendo el modelado de características (feature models) una herramienta principal para representar las partes configurables de la familia (Kang, Cohen, Hess, Novak, & Peterson, 1990), identificando puntos comunes y puntos de variabilidad entre los productos específicos que serán parte de la línea de productos.

Por otra parte, ramas de la inteligencia artificial como Machine Learning (aprendizaje automático) (Nasrabadi, 2007) pueden ser aplicadas en conjunción a este paradigma para obtener ya no solo herramientas versátiles en cuanto a su desarrollo, sino herramientas personalizadas que pueden alimentarse de las características del usuario (Cruz-Benito et al., 2017) para ofrecerles en cada momento los componentes que mayores beneficios reportarían en sus procesos de toma de decisiones y en el alcance de sus objetivos particulares.

HIPÓTESIS DE TRABAJO Y PRINCIPALES OBJETIVOS

La hipótesis de trabajo es la siguiente: "la personalización de interfaces de usuario para el soporte a la toma de decisiones aumenta la eficiencia y efectividad a la hora de extraer información y conocimiento de los datos mostrados".

El principal objetivo del proyecto de tesis es obtener un marco generativo para la creación de paneles de control de manera automatizada y sistemática, de tal forma que sea posible generar interfaces software personalizadas no solo a nivel de estructura y diseño visual, sino también a nivel de datos y patrones de interacción para promover la usabilidad y flexibilidad de los mismos, así como la aplicación de técnicas de aprendizaje automático para ofrecer las configuraciones de la línea de productos más beneficiosas según el contexto.

Para alcanzar este objetivo global, se plantean una serie de sub-objetivos:

- Identificar (a nivel meta) las características de los paneles de control
- Identificar mecanismos de conexión para que la producción de productos específicos en la SPL esté guiada por los meta-modelos y modelos de características
- Implementar mecanismos que fomenten la interoperabilidad y que suavicen el impacto la heterogeneidad de la información para permitir la conexión de diversas fuentes de datos
- Definición e implementación de core assets reutilizables y configurables para la generación de productos específicos de la familia
- Evaluar el correcto comportamiento de la línea de productos a nivel generativo y funcional
- Evaluar la usabilidad y adaptación a requisitos dinámicos de la línea de productos
- Estudiar la adaptación dinámica de las características de los productos en función del comportamiento de los usuarios y su contexto a través de la aplicación de Machine Learning

Estudiar la integración de la línea de productos de paneles de control en diversos ecosistemas tecnológicos y casos de estudio

Este proyecto de investigación se plantea como un proceso iterativo donde se tendrán en cuenta los conocimientos adquiridos gracias a experiencias previas y los resultados obtenidos en los consecuentes ciclos. Para llevar a cabo este proceso se tomará como marco metodológico el método Investigación-Acción (Lewin, 1946).

Según Kemmis (1984) la Investigación-Acción puede plantearse como una forma de indagación autorreflexiva llevada a cabo por los participantes en las situaciones sociales con el fin de mejorar sus propias prácticas sociales, así como la comprensión de las mismas y de los contextos en los cuales se realizan dichas prácticas. Posteriormente McTaggart, junto al propio Kemmis, describe las características de esta metodología, de acuerdo con las cuales la metodología de Investigación-Acción se basa en una espiral de ciclos de investigación y acción compuesto por una serie de fases y secuencias (McTaggart & Kemmis, 1988). Por tanto, el proceso de Investigación-Acción es un proceso iterativo, donde cada ciclo proporciona una salida, que será a su vez la entrada para el siguiente ciclo, obteniendo eventualmente soluciones más refinadas y permitiendo abordar en el siguiente ciclo los problemas detectados en el ciclo recientemente finalizado.

Debido a su carácter iterativo y a su semejanza con ciertos modelos de desarrollo de software, el paradigma Investigación-Acción será el marco metodológico general para este proyecto de Tesis Doctoral. Sin embargo, para poder comenzar los ciclos de Investigación-Acción es necesario formalizar el problema a resolver. Para ello se estudiarán problemas similares y las soluciones alcanzadas en la literatura, utilizando como metodología la revisión sistemática de la literatura (Kitchenham, 2004).

Una vez formalizado el problema, se plantean dos ciclos de Investigación-Acción para el desarrollo de la propuesta de solución para la generación de paneles de control y su evaluación en contextos reales para obtener la correspondiente realimentación que permita mejorar la solución propuesta. Como marco de desarrollo software se apostará por un enfoque ágil basado en SCRUM (Schwaber, 1997). Este marco dispone el proceso, las reglas, las prácticas, los roles y los artefactos necesarios para aumentar la productividad de un equipo de desarrollo basado en un ciclo de desarrollo de software iterativo e incremental (Schwaber & Beedle, 2002).

MEDIOS Y RECURSOS MATERIALES DISPONIBLES

El presente proyecto de investigación está enmarcado en el grupo de investigación GRIAL (GRupo de Investigación en InterAcción y eLearning) (García-Peñalvo et al., 2012; Grupo GRIAL, 2019). Dicho grupo ofrecerá recursos para dar soporte al proyecto de investigación a través de la disponibilidad de espacios de trabajo, servidores y cualquier otro recurso material y/o tecnológico que se considere de necesidad.

A su vez, la Universidad de Salamanca da acceso a los recursos bibliográficos necesarios para el desarrollo de la tesis doctoral a través de sus bibliotecas y los recursos electrónicos suscritos y disponibles para los miembros de la comunidad universitaria.

PLANIFICACIÓN TEMPORAL

El proyecto de Tesis Doctoral se circunscribe en el proyecto de investigación A Digital Ecosystem Framework for an Interoperable NEtwork-based Society (DEFINES) (García-Peñalvo, 2016). Las diferentes fases planteadas en la metodología se llevarán a cabo de manera secuencial, junto a las fases del proyecto que lo avala. Al final de cada fase se plantea un hito con una serie de resultados que permitirán continuar a la siguiente fase. Los hitos principales cuya entrega se ha planificado a lo largo de ese periodo de 36 meses (3 años) son los siguientes:

1. Systematic Literature Review (Mes 6)
 2. Primer ciclo action-research: Definición de la línea de productos (Mes 14)
 3. Segundo ciclo action-research: Creación del sistema generativo y casos de estudio (Mes 30)
- Memoria de Tesis Doctoral (Mes 36)

Todos los artículos relacionados con esta tesis estarán disponibles en acceso abierto (García-Peñalvo et al., 2010).

REFERENCIAS

- Albright, S. C., Winston, W., & Zappe, C. (2010). Data analysis and decision making: Cengage Learning.
- Chang, E., & West, M. (2006). Digital ecosystems a next generation of the collaborative environment. In G. Kotsis, D. Taniar, E. Pardede, & I. K. Ibrahim (Eds.), *Proceedings of iiWAS'2006 - The Eighth International Conference on Information Integration and Web-based Applications Services*, 4-6 December 2006, Yogyakarta, Indonesia (pp. 3-24): Austrian Computer Society.
- Chang, R. (2017). Using Machine Learning to Predict Value of Homes On Airbnb. Retrieved from <https://medium.com/airbnb-engineering/using-machine-learning-to-predict-value-of-homes-on-airbnb-9272d3d4739d>
- Clements, P., & Northrop, L. (2002). *Software product lines*: Addison-Wesley.
- Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*: Sage publications.
- Cruz-Benito, J., Vázquez-Ingelmo, A., Sánchez-Prieto, J. C., Therón, R., García-Peñalvo, F. J., & Martín-González, M. (2018). Enabling adaptability in web forms based on user characteristics detection through A/B testing and machine learning. *IEEE Access*, 6, 2251-2265. doi:10.1109/ACCESS.2017.2782678
- Eppler, M. J. (2007). Knowledge communication problems between experts and decision makers: An overview and classification. *Electronic Journal of Knowledge Management*, 5(3).
- Few, S. (2006). *Information dashboard design*.
- García-Peñalvo, F. J. (2016). A Digital Ecosystem Framework for an Interoperable Network-based Society (DEFINES). Retrieved from Salamanca, España: <https://goo.gl/FDbN5K>
- García-Peñalvo, F. J., García de Figuerola, C., & Merlo-Vega, J. A. (2010). Open knowledge: Challenges and facts. *Online Information Review*, 34(4), 520-539. doi:10.1108/14684521011072963
- García-Peñalvo, F. J., Rodríguez-Conde, M. J., Seoane-Pardo, A. M., Conde-González, M. Á., Zangrando, V., & García-Holgado, A. (2012). GRIAL (GRupo de investigación en InterAcción y eLearning), USAL. *IE Comunicaciones. Revista Iberoamericana de Informática Educativa*(15), 85-94.
- Gomez-Urbe, C. A., & Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4), 13.
- Grupo GRIAL. (2019). *Producción Científica del Grupo GRIAL de 2011 a 2019* (GRIAL-TR-2019-010). Salamanca, España: Grupo GRIAL, Universidad de Salamanca. Retrieved from <https://bit.ly/3019mLh>
- Kang, K. C., Cohen, S. G., Hess, J. A., Novak, W. E., & Peterson, A. S. (1990). Feature-oriented domain analysis (FODA) feasibility study. Retrieved from
- Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges *Information visualization* (pp. 154-175): Springer.
- Kemmis, S. (1984). *Point-by-point guide to action research*. Victoria: Deakin University.
- Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004), 1-26.
- Kleppe, A. G., Warmer, J., & Bast, W. (2003). *MDA Explained. The Model Driven Architecture: Practice and Promise*: Addison-Wesley Longman Publishing Co., Inc., Boston, MA.
- Lewin, K. (1946). Action research and minority problems. *Journal of social issues*, 2(4), 34-46.
- McTaggart, R., & Kemmis, S. (1988). *The action research planner*: Deakin university.
- Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4), 049901.
- Nieke, M., Mauro, J., Seidl, C., & Yu, I. C. (2016). User profiles for context-aware reconfiguration in software product lines. Paper presented at the *International Symposium on Leveraging Applications of Formal Methods*.
- Patil, D., & Mason, H. (2015). *Data Driven*: " O'Reilly Media, Inc."
- Patnaik, D., & Becker, R. (1999). Needfinding: the why and how of uncovering people's needs. *Design Management Review*, 10(2), 37-43.
- Pleuss, A., Wollny, S., & Botterweck, G. (2013). Model-driven development and evolution of customized user interfaces. Paper presented at the *Proceedings of the 5th ACM SIGCHI symposium on Engineering interactive computing systems*.
- Pohl, K., Böckle, G. n., & Linden, F. J. v. d. (2005). *Software Product Line Engineering: Foundations, Principles and Techniques*: Springer-Verlag New York, Inc.
- Pressman, R. S. (2005). *Software engineering: a practitioner's approach*: Palgrave Macmillan.
- Sarikaya, A., Correll, M., Bartram, L., Tory, M., Fisher, D.: What Do We Talk About When We Talk About Dashboards? *IEEE Transactions on Visualization Computer Graphics* (2018)
- Schwaber, K. (1997). *Scrum development process Business object design and implementation* (pp. 117-134): Springer.
- Schwaber, K., & Beedle, M. (2002). *Agile software development with Scrum* (Vol. 1): Prentice Hall Upper Saddle River.

- Seidl, C., Schaefer, I., & Aßmann, U. (2014). Capturing variability in space and time with hyper feature models. Paper presented at the Proceedings of the Eighth International Workshop on Variability Modelling of Software-Intensive Systems.
- Sharda, R., Delen, D., & Turban, E. (2013). Business intelligence: a managerial perspective on analytics: Prentice Hall Press.
- Tufte, E., & Graves-Morris, P. (2014). The visual display of quantitative information.; 1983.
- Wang, S., Keivanloo, I., & Zou, Y. (2014). How do developers react to restful api evolution? Paper presented at the International Conference on Service-Oriented Computing.