Modelo de Análisis de Información Desestructurada Utilizando Técnicas de Recopilación y Minería Web

Karina B. Eckert[a], Fabián E. Favret[b], Roberto Suénaga[c], Tokuji Kairiyama[d]

Universidad Gastón Dachary

Avda. López y Planes 6519, Posadas, Misiones-Argentina. Tel: +54 (0376) – 447699

{karinaeck[a], fabianfavret[b]}@gmail.com, {rsuenaga[c], kairiyama[d]}@ugd.edu.ar

Resumen

Debido al gran avance en las Tecnologías de la Información y las Comunicaciones (TICs) se ha facilitado y simplificado de manera significativa el acceso, el procesamiento y el almacenamiento de los datos las organizaciones en general. Estas tecnologías han cambiado el paradigma del análisis de la información, ya que hoy en día el problema no es la escasez, sino la excesiva cantidad de datos disponibles que, claramente, no todos son de utilidad para el proceso de Toma de Decisiones (TD). Otro inconveniente, además del gran volumen de información que se necesita analizar, es que los datos disponibles están desestructurados y distribuidos, dificultando aún más la tarea de detectar aquellos que pueden ser útiles. Este proyecto tiene como objetivo desarrollar una herramienta para búsqueda de información para el proceso de TD.

Palabras clave: Toma de decisiones, minería web, análisis información desestructurada

Contexto

La presente investigación se encuentra en ejecución, acreditado en la Secretaría de Investigación y Desarrollo de la UGD por resolución N° 07/A/17.

Introducción

Sistemas de Apoyo para el Proceso de TD

Un sistema eficiente para dar apoyo al proceso de TD debe contener las funcionalidades necesarias para el análisis y recopilación de requerimientos de usuario, recolección de la información en base al análisis eficiente de la misma y la presentación adecuada de los resultados.

Un gran número de proyectos de desarrollo de aplicaciones fracasan por no realizar una adecuada definición, especificación, y administración de los requerimientos.

Problemas como la falta de definiciones del usuario, requerimientos incompletos y el mal manejo de los cambios en los requerimientos son factores determinantes [1]. Existen varias técnicas propuestas para recopilar requerimientos como entrevistas y cuestionarios, prototipos o simulaciones del posible sistema o incluso casos de uso [2]. Claramente todas estas técnicas requieren la interacción con el usuario para determinar de manera correcta las necesidades de información. Una vez que se han establecidos los requerimientos del usuario, el sistema debe comenzar a

buscar la información que pueda ser de utilidad para el proceso de TD.

Claramente, para este tipo de actividad, la fuente más importante de información es la Web. La gran cantidad de información desestructurada y distribuida en Internet es una complicación al momento de decidir qué datos son relevantes y cuáles no. Para ello existen múltiples técnicas y métodos que derivan del área del aprendizaje automático y que se utilizan para hacer Minería Web (MW) [3][4].

La MW tiene como objetivo descubrir información útil e intentar extraer conocimiento de la estructura de hipervínculos, de los contenidos de las páginas y de los datos de uso. La MW utiliza muchas técnicas de Minería de Datos (MD) [5], no es puramente una aplicación de MD tradicional debido a la heterogeneidad y la naturaleza semi-estructurada o no estructurada de los datos de la Web. Sobre la base de los tipos principales de datos utilizados en el proceso de minería, las tareas de MW se pueden clasificar en tres tipos: la minería de la estructura de la Web, la minería de contenidos Web y la minería del uso de la Web [3].

El proceso de MW es similar al proceso de MD, la diferencia radica en el proceso de obtención de los datos. En la MD, los datos se colectan y luego se guardan en un almacén de datos. En la MW, la recopilación de datos puede ser una tarea demasiado compleja, sobre todo para determinar la estructura y luego analizar el contenido, tarea que implica el rastreo de un gran número de fuentes de información.

Existen múltiples algoritmos y técnicas que se utilizan para determinar la relevancia de la información y ellos se basan en extracción de conocimiento de los datos. En este sentido, la obtención de reglas que describen tanto las preferencias como el comportamiento de los usuarios se realiza mediante el análisis de uso de los recursos web, por ejemplo mediante reglas de asociación [6][7] y minería de patrones secuenciales [8][9][10].

Este tipo de técnicas intentan encontrar regularidades en los datos sin tener una referencia explícita externa que direccione la búsqueda. Es decir, se trabaja directamente con los datos resultantes de la interacción del usuario con los recursos web.

También existen algoritmos como los árboles de decisión [11], naive bayes [12][13] y las máquinas de vectores de soporte [14] que son técnicas basadas en robustos principios matemáticos y que han sido muy utilizados para la clasificación de datos. Estos algoritmos, así como algunas redes neuronales [15], requieren la utilización de etiquetas asociadas a los datos. Otro grupo de algoritmos muy utilizados en el análisis de información son los denominados algoritmos de clustering o agrupamiento [16][17]. Este tipo de técnicas tienen como objeto agrupar datos por similitud utilizando alguna función para medir la distancia entre ellos. La idea subyacente es que los datos similares deben estar cerca ya que comparten características similares y es por ello que se intenta medir la mencionada distancia que los separa.

Muchos de estos algoritmos son utilizados para recuperar información de la Web, como en la utilización de Crawlers (programas que recorren la estructura de lo hiperlinks), analizadores de hiperlinks y contenido web y analizadores de estructura de la Web. La implementación de estas técnicas otorga valor significativo al proceso de búsqueda, análisis y selección de la información útil para el proceso de TD.

Al obtener información de la Web, es necesaria que sea presentada de manera

adecuada al usuario. Claramente, cuando más precisa y clara sea la forma de mostrar la información, se podrá sacar mayor provecho de la información. Uno de los inconvenientes al presentar la información de relevancia es la integración que presenta la misma [3]. Debido a diversas fuentes de información en la Web, los diferentes sitios suelen utilizar diferentes palabras o términos para expresar la misma información o similar. Con el fin de hacer uso de los datos extraídos de varios sitios, se necesita integrar mediante técnicas semánticas la información de estos sitios. De esta manera se intenta que coincidan los datos que son semánticamente lo mismo pero expresan de manera diferente en distintos sitios [18].

Otro aspecto a considerar en la presentación de la información al usuario es la visualización de datos, que tiene por objeto comunicar la información de forma clara y eficaz a través de la representación gráfica [5]. Se pueden utilizar las técnicas de visualización para exponer conocimiento que de otra manera no sería fácilmente observable mediante el examen de los datos en bruto.

Si bien existen muchas técnicas de visualización de los datos [19] como la proyección geométrica, la visualización jerárquica, la basada en íconos o en relaciones de datos, entre otras, todas tienen como objetivo destacar los aspectos relevantes de la información. De esta manera, las herramientas de visualización son parte esencial en los sistemas para dar apoyo al proceso de TD.

Sector productivo Té

La falta de información útil para el proceso de TD afecta a todas las organizaciones, pero las más vulnerables son las pequeñas que están inmersas en economías regionales, caracterizadas por ser productoras de bienes primarios y no tener recursos destinados para adaptarse y/o reconvertirse para hacer frente a los cambios del entorno.

El Clúster de Té de Misiones o Conglomerado Productivo Tealero de Misiones se constituye en el año 2006 en el marco del programa FONTAR (Fondo Tecnológico Argentino) donde mediante el Plan de Mejora de la Competitividad (PMC) se analizó las necesidades del sector, se detectaron los factores críticos de éxitos y se propuso diferentes líneas de acción con sus propuestas superadoras.

En el año 2011 se realizó una actualización del PMC donde se analizó el enfoque estratégico contrastando la mirada de los actores locales del té de Misiones con la percepción de los compradores nacionales e internacionales, que en base al diagnóstico surgieron nuevas estrategias para cada eslabón.

Entre las necesidades detectadas dentro del sector del té, se destaca tener acceso a información sobre: nuevos mercados, alternativas sobre agregación de valor al té y conocimiento del mercado internacional. La información permitirá tomar decisiones acerca del análisis para el reposicionamiento del té en el mercado local e internacional, la estrategia comercial de toda la cadena productiva y definir una caracterización del producto en la región que permita reposicionar respecto al mercado regional e internacional.

Líneas de Investigación, Desarrollo e Innovación

Este proyecto abarca cinco etapas:

1. Obtención de información respecto a las técnicas de recopilación de necesidades, búsqueda automática, exploración y MW, y herramientas de TD.. Esta etapa se divide en dos actividades principales: (i) relevamiento

de información específica sobre las técnicas de MW, búsqueda automática y sistemas de soporte para las decisiones y (ii) análisis sistematizado de la información relevada a fin de determinar el estado del arte de los algoritmos y técnicas sobre nuevas formas de búsqueda de información.

2. Recopilación de información detallada del sector productivo del té, incluyendo un análisis de la situación actual y del contexto (programas de mejora, alternativas de financiamiento, proyectos vigentes, etc.). Se estudian técnicas y algoritmos de identificación y análisis de información contextualizada a requerimientos específicos (demandas sectoriales).

3. Estudio, análisis y desarrollo de las herramientas que serán integradas en el modelo de recolección de información. La etapa se enfoca en la relación, influencia y resultados del proceso de integración que contendrá la herramienta a desarrollar. Específicamente esta etapa tiene dos actividades prioritarias: (i) un estudio integral de las herramientas a desarrollar e implementar y (ii) el desarrollo e implementación de las técnicas relacionadas con la captura de requerimientos, la búsqueda automática y la presentación de la información al usuario.

4. Formulación de los modelos de negocio a partir de la utilización de la herramienta en el ámbito productivo del té de Misiones. En esta etapa se pretende que el empleo de la herramienta desarrollada aporte características de valor en los nuevos modelos de negocio que puedan ser utilizados en el sector productivo del té.

5. Evaluación de los resultados obtenidos a partir de la implementación de la herramienta para la obtención de información útil en el ámbito del sector productivo del té. En esta etapa se trabajará con el sector productivo del té para determinar el impacto potencial que generen los modelos de negocio desarrollados.

Objetivos

Objetivo general

Desarrollar y utilizar un modelo para el proceso de toma decisiones basado en la integración de técnicas de recopilación, exploración y análisis de información.

Objetivos Específicos

- Estudiar y analizar técnicas de manejo de avanzado de datos (clustering, reconocimiento de patrones, modelos descriptivos y predictivos).
- Definir y analizar los modelos de negocio actuales del sector productivo del té de Misiones.
- Analizar métodos de búsqueda y exploración de información en la Web.
- Integrar métodos de recopilación, exploración y análisis de información en una herramienta informática de entorno web.
- Utilizar la herramienta para obtener información para la toma de decisiones del sector productivo del té, que permitan generar escenarios potencialmente convenientes para mejorar la competitividad.

Formación de Recursos Humanos

El equipo del trabajo de investigación está compuesto por un Dr. en Ingeniería de Sistemas y Computación, un magister en administración de empresas especialización en marketing, dos maestrando, uno en redes de datos y el otro en tecnologías de la información.

Director del Proyecto:

Dr. Marcelo Karanik
Co-Director:
Ing. Roberto Suénaga
Docentes-Investigadores:
Ing. Karina Eckert
Ing. Fabián Favret
Mg. Tokuji Kairiyama

Referencias

[1] I. Sommerville, *Software Engineering*. United Kingdom: Addison-Wesley, 2005.

[2] M. A. Chaves, "La ingeniería de requerimientos y su importancia en el desarrollo de proyectos de software," *InterSedes*, vol. 6, no. 10, 2005.

[3] Bing Liu, *Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data*, First Edition. Chicago: University of Illinois, 2007.

[4] P. K. P. Ranout and A. S. P. Sharma, *Web Mining-Concept, Classification and Major Research Issues: A Review*, vol. 4, 2 vols. 2016.

[5] H. J. J. Pei and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[6] A. R. T Imieliński and A. Swami, "Mining association rules between sets of items in large databases," *Acm sigmod record*, vol. 22, 1993.

[7] Z. Z. R. Kohavi and L. Mason, "Real world performance of association rule algorithms," *ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. ACM*, 2001.

[8] R. Srikant and A. Rakesh, "Mining sequential patterns," *Data Eng. Proc. Elev. Int. Conf. IEEE*, 1995.

[9] J. Ayres, "Sequential pattern mining using a bitmap representation," *Proc. Eighth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. ACM*, 2002.

[10] J. R. A. McCallum, "Using reinforcement learning to spider the web efficiently," *ICML*, vol. 99, 1999.

[11] J. Ross Quinlan, *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. 1992.

[12] P. D. M. Pazzani, "On the optimality of the simple Bayesian classifier under zeroone loss," *Mach. Learn.*, vol. 29, p. 3, 1997.

[13] K. R. B. Becker and D. Sommerfield, "Improving simple bayes," *Proc Eur. Conf. Mach. Learn. ECML'97*, 1997.

[14] B. E. B. I. M. Guyon and V. N. Vapnik, "A training algorithm for optimal margin classifiers," *Proc. Fifth Annu. Workshop Comput. Learn. Theory ACM*, 1992.

[15] S. S. Haykin, *Neural Networks and Learning Machines: A Comprehensive Foundation*, 3rd ed. USA: Pearson, 2009.

[16] A. K. J. R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, 1988.

[17] X. R. D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, 2005.

[18] P. R. P. Heiko, "Semantic Web in data mining and knowledge discovery: A comprehensive survey," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 36, pp. 1–22, 2016.

[19] D. King, "Introduction to the Mining, Analysis and Visualization of Web Content and Usage Minitrack," *49th Hawaii Int. Conf. Syst. Sci. HICSS IEEE*, 2016.