I+D

N° 2
SEPTIEMBRE 2019

Revista de investigación y desarrollo
de la Universidad Gastón Dachary

ISSN 2591-3778

UNIVERSIDAD
Gastón Dachary

I+D

Publicación en soporte digital, cuyo objeto es dar a conocer artículos de divulgación científica resultados de investigaciones desarrolladas en el ámbito de la Universidad Gastón Dachary.
Comité Editorial. Editor Responsable: Editorial UGD.
Universidad Gastón Dachary. Salta 1912. Posadas, Misiones.
ISSN 2591-3778

Editorial UGD
Universidad Gastón Dachary
Salta 1912, Posadas, Misiones, Arg.
Tel./Fax: +54 (0376) 4438677
Contacto: editorial@dachary.edu.ar

RED DE EDITORIALES DE UNIVERSIDADES PRIVADAS

ÍNDICE

EDITORIAL

2

Se concreta con esta publicación de la Universidad Gastón Dachary, la segunda edición de la Revista I+D con la firme convicción de continuidad a sabiendas que la divulgación de saberes es un fin esencial de la acción universitaria.

Los artículos presentados en esta segunda edición de I+D, surgen de los resultados obtenidos al término de la ejecución de los proyectos presentados y aprobados pertenecientes a la séptima Convocatoria Interna de Proyectos de Investigación organizadas por la Secretaría de Investigación y Desarrollo y abarca las áreas de Derecho, Economía, Producción Agropecuaria e Informática.

Se pretendió desde el Equipo Editorial de esta revista, en acuerdo con los autores de los artículos presentados, mantener un estilo lo suficientemente afable destinado a una comunidad de lectores más amplia y diversa.

Cotidianamente en estos tiempos expresamos que estamos viviendo en la sociedad del conocimiento y es en esa sociedad donde las instituciones educativas de todos los niveles y singularmente las universitarias, les cabe generar saberes. Los saberes producidos allí y con anclaje directo en la comunidad regional –mediante transferencias científica y tecnológicas socializadas- puede favorecer la conceptualización y la solución de los problemas para que la gente viva mejor.

Es una vocación prosocial la que anima la razón de ser de esta segunda edición de la Revista I+D, llevar la posibilidad de lectura a una cada vez más amplia sociedad de referencia. Con esta nueva revista pretendemos, por otra parte, seguir sumando al saber, motivados como equipo editorial a transferir la producción científica de nuestros equipos de investigadores.

WEBSHAWN: SIMULANDO REDES DE SENSORES INALÁMBRICOS DESDE LA WEB

AUTORES: Godoy, D. Sosa, E. Bareiro, S. Belloni, E. Favret, F. Benitez, J.

RESUMEN

En este trabajo se presentan las actividades y logros alcanzados en el marco del proyecto de investigación IP A07003 "Simulación en las Tics: Diseño de Simuladores de Procesos de Desarrollo de Software Agiles y Redes de Sensores Inalámbricos para la Industria y la Academia" desarrollado en el Centro de Investigación de Tecnologías de la Información y Comunicaciones adjudicado en la 7ma convocatoria de proyecto de Investigación y desarrollo de la Universidad Gastón Dachary. El proyecto abarca dos grandes áreas de estudio de la simulación para procesos de desarrollo de software ágiles y de redes de sensores inalámbricos como base fundamental para la concreción de la visión de Internet de las Cosas y las ciudades inteligentes. Como uno de sus resultados más relevantes se presenta WebShawn, un prototipo de interfaz Web para un simulador de WSN llamado Shawn. Al mismo se le realizaron adaptaciones en su funcionamiento para realizar corridas de simulación y visualización de resultados a través de la Web. Se describen los componentes del simulador, la metodología utilizada para el desarrollo del prototipo y finalmente los aspectos fundamentales de la implementación. Conjuntamente se presentan pruebas de funcionamiento. Como conclusión se puede sostener que se han logrado los objetivos y que se proponen acciones para continuar trabajando en las mismas líneas.

Palabras Claves: Wireless Sensor Networks, Simulación, Shawn.

INTRODUCCIÓN

El proceso de instalación y configuración de un sistema de simulación de propósito específico, puede ocasionar un consumo importante de tiempo. Ya sea porque requiera ser instalado desde un sistema operativo en particular, o porque es preciso adquirir el hardware y el conocimiento para la instalación correcta del simulador en sí.

Además, luego de realizar la puesta en funcionamiento del simulador, se debe entender cómo proceder con la configuración correcta, para obtener resultados de simulación lo más cercano posible a sistemas reales. Esto implica en algunos simuladores actuales conocer no solo de la problemática a simular, sino también de cómo ingresar los parámetros, dónde ubicar los archivos que son necesarios como fuentes del proyecto y finalmente cómo generar la visualización de los resultados de simulación.

En este sentido, al momento de poner en funcionamiento un sistema de simulación, se buscan herramientas portables o multiplataforma y que tengan una interfaz que abstraiga al usuario de la complejidad necesaria en la configuración del sistema de simulación. Este tipo de herramientas evitan que los tiempos de un proyecto de simulación se trasladen en mayor medida a la instalación y puesta en marcha del simulador.

Una de las principales características de los sistemas Web es su portabilidad, así también como su facilidad de acceso en forma remota. Con el uso del modelo cliente-servidor, y al utilizar un navegador Web como interfaz, se abstrae al usuario de toda la complejidad interna de un sistema. La complejidad en este caso queda completamente transparente del lado del servidor.

Con la utilización de Shawn (Fischer, Pfisterer, & Fekete, 2007. ISBN 1-4244-1231-5) como sistema de simulación específico para nodos ISense (Coalesense, 2013), se ha comprobado que es una herramienta de software poderosa para la simulación de Redes de Sensores Inalámbricos. Pero también se ha observado una gran complejidad para realizar la instalación, configuración y puesta a punto de los módulos del simulador (Shawn Community, 2013). Estas características y limitaciones hacen que Shawn sea un sistema de simulación especialmente adaptable a la Web.

Si bien ya existen trabajos similares de simulación basados en la Web, se considera de gran utilidad realizar una adaptación al ámbito de la simulación de redes, particularmente a las Redes de Sensores Inalámbricos (Wireless Sensor Networks - WSN); donde actualmente se disponen de escasas alternativas que conjuguen estas dos tecnologías. Entre las alternativas se pueden mencionar a "A Web-Based Integrated Environment for Simulation and Analysis with NS-2". En el que se desarrolla un completo entorno para la simulación y pos-proceso de las salidas generadas por el Network Simulation version 2 (NS2 Commnity, 2011). La aplicación Web denominada ns2web permite la ejecución remota de simulaciones de redes inalámbrica, incluyendo WSNs y cableadas. También ofrece un conjunto de herramientas para analizar los archivos de rastreo que genera como salida el simulador. Una implementación del simulador se encuentra on-line para público acceso en (NS2 WEB Community, 2013). Otro ejemplo es "Web-based simulation management: A Web-based interface for storing and excuting simulation model", el cual consiste en el desarrollo para el lenguaje de simulación SIMAN (Pegden, 1983). El lenguaje SIMAN permite la simulación de sistemas discretos y dinámicos. SIMAN es un simulador de propósito general, lo cual no se condice con las propiedades específicas para realizar simulación de WSN. Para este caso Shawn ofrece más características que se adaptan mejor.

EL SIMULADOR SHAWN

Shawn es un framework cuya idea central es reemplazar los efectos a bajo nivel de una WSN, con modelos abstractos e intercambiables de manera que se pueda utilizar en la simulación de grandes redes en un tiempo

razonable (Fischer, Pfisterer, & Fekete, 2007. ISBN 1-4244-1231-5). Conceptualmente Shawn se compone de tres partes principales: el Entorno de Simulación, el Secuenciador y los Modelos. El Entorno de Simulación contiene los elementos simulados y sus propiedades, mientras que el Secuenciador y los Modelos influyen en el comportamiento del Entorno o Ambiente de Simulación (Kroeller, Pfisterer, Buschmann, & Fekete, 2005). La Figura 1 muestra los principales componentes de Shawn.

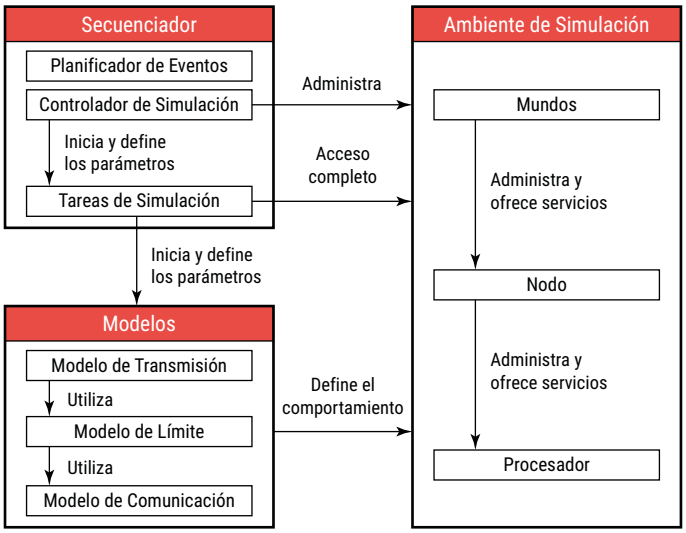


Figura 1. *Arquitectura general de los componentes principales de Shawn (Kroeller, Pfisterer, Buschmann, & Fekete, 2005)*

La principal característica que hace de Shawn sea más rápido (Shawn, 2013) que otros simuladores de WSN es su composición interna o diseño. Mientras que otros simuladores se centran en un fenómeno en particular, Shawn se centra en el efecto causado por el fenómeno.

El modelado que se realiza con Shawn de un sistema apunta a la velocidad de simulación a gran escala. En vez de realizar el cálculo de posibles congestiones en una red, analizando cada paquete en forma individual, simplemente se modela la pérdida de paquetes provocados por el alto tráfico.

Este tipo de modelado, se adapta específicamente al paradigma de WSN, donde el foco de la investigación de un problema de software está en entender la estructura fundamental de la red. Una tarea que a menudo está un nivel por encima de los detalles técnicos referente a los nodos individuales y los efectos de bajo nivel.

Un beneficio directo de este tipo de modelado es la escalabilidad. Escenarios visionarios anticipan Redes de Sensores Inalámbricos con un gran número de nodos individuales. Es por ello que un aspecto crítico en el diseño de Shawn es el de poder realizar exitosamente simulaciones de una red con más de 100.000 nodos en un equipo PC estándar (Kroeller, Pfisterer, Buschmann, & Fekete, 2005).

WEBML

El prototipo del simulador fue desarrollado utilizando Web Modeling Language. WebML es una notación o lenguaje de alto nivel para el modelado, diseño e implementación de aplicaciones Web que hacen uso intensivo de datos (Ceri, Fraternali, & Bongio, 2000).

Como se muestra en la Figura 2 el enfoque de WebML para el desarrollo de aplicaciones Web está compuesto por diferentes fases. Éste enfoque está inspirado en el modelo en espiral de (Boehm, 1988) y en línea con los modernos métodos de desarrollo de software. El proceso de desarrollo con WebML se aplica en forma iterativa e incremental, de manera que las fases se repiten y se refinan hasta que los resultados cumplen con los requisitos de determinada aplicación (Brambilla, Comai, Fraternali, & Matera, Londres, 2008). Este tipo de metodología con ciclos iterativos se adapta especialmente a los desarrollos Web dado la dinámica de cambios que pueden suceder durante el desarrollo y la rapidez con que deben ser desplegados o puestos en funcionamiento los sitios en Internet.

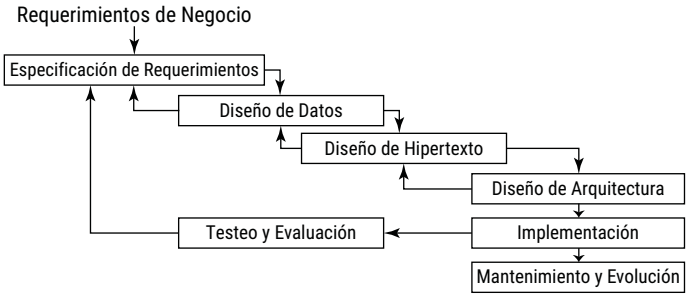


Figura 2. *Fases de WebML (Ceri, Fraternali, Bongio, Brambilla, Comai, & Matera, 2003)*

A continuación, se describe brevemente cada fase:

A- Especificación de Requerimientos: Se centra en la recopilación de información acerca del dominio de la aplicación, las funciones esperadas y sobre la especificación de las funciones por medio de descripciones fáciles de entender.

B- Diseño de Datos: Esta fase implementa una de las disciplinas más tradicionales y consolidadas en la tecnología de la información. Por lo cual WebML utiliza el modelo de datos entidad-relación o el diagrama de clases de UML.

C- Diseño de Hipertexto: El modelo de hipertexto permite la definición de la interfaz front-end de la aplicación, la cual se muestra en el navegador del usuario. Esto permite la especificación de las páginas y su organización interna en términos de componentes para la visualización del contenido. También se pueden definir los enlaces y localización de la información.

D- Diseño de Arquitectura: Consiste en la especificación del hardware, redes y componentes de software que mejor se adapte a los requerimientos de la aplicación.

E- Implementación: Es la fase donde se producen los módulos de software necesarios para transformar el diseño de datos y de hipertexto en una aplicación que se ejecute en la arquitectura seleccionada.

F- Pruebas y Evaluación: Es la actividad de verificación de la conformidad de la aplicación implementada con los requisitos funcionales y no funcionales.

G- Mantenimiento y Evolución: Abarcan todas las modificaciones efectuadas después que la aplicación ha sido desplegada en un entorno de producción.

H- Herramientas CASE para modelado con WebML de WebRatio: WebML posee un conjunto de herramientas para modelado asistido por computadora desarrollado por WebRatio (Web-Ratio, 2014).

CONSTRUCCIÓN DEL SIMULADOR

La arquitectura de software para realizar el desarrollo del simulador se apoya en el esquema de Simulación y Visualización Remota (Whitman, Huff, & Palaniswamy, 1998) y (Byrnea, Heaveya, & Byrne, 2010). Donde el motor de simulación y visualización se encuentran completamente implementados en el servidor. El acceso al sistema de simulación es siempre a través del navegador Web. Este último se utiliza como interfaz liviana para ingresar los parámetros de simulación. Los parámetros ingresados son enviados al servidor Web a través de una intranet o bien Internet. El servidor obtiene los parámetros enviados, que luego son dirigidos al motor de simulación Shawn. Una vez que el motor termina completa satisfactoriamente la simulación, los resultados son devueltos al usuario a través de la interfaz Web.

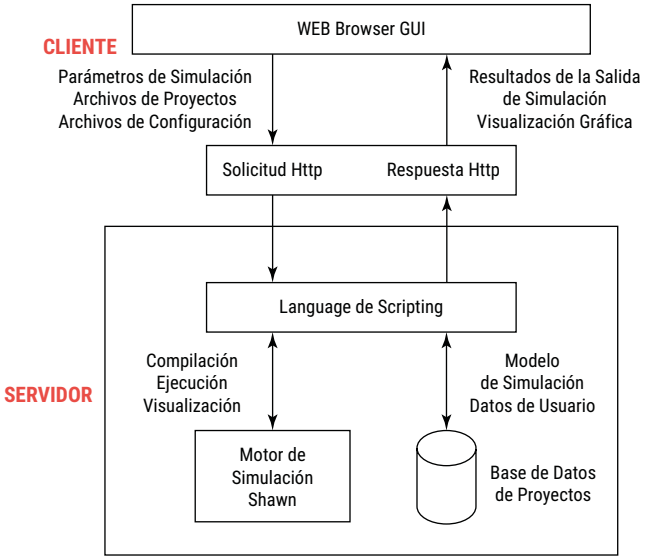


Figura 3. *Esquema general del simulador Basado en la Web.*

Es importante destacar que, además de ingresar los parámetros que controlan la ejecución de la simulación, también se puede editar a través de la interfaz Web todo el código fuente que dará funcionalidad a los sensores.

De este modo el usuario puede ir guardando sus diferentes modelos de simulación en el servidor para luego realizar pruebas con diferentes soluciones. En la Figura 3 se muestra un esquema general de la solución.

La interfaz de usuario que se ejecuta en el navegador Web del lado del cliente se desarrolla utilizando HTML5, JavaScript, CSS y estará disponible para dispositivos tipo PC. Con el fin de almacenar información relativa a los proyectos de simulación de los usuarios, datos de la aplicación, modelos de simulación y estructuras de datos, se empleará el sistema de gestión de bases de datos PostgreSQL.

El esquema de navegación del simulador utiliza solapas o pestañas que son progresivas en el uso de la herramienta. En la primera pestaña se puede gestionar todo lo relacionado al árbol de archivos de cada proyecto (Figura 4).

Figura 4. *Interfaz Web de gestión de archivos de proyectos de simulación*

La siguiente pestaña en la navegación permite ingresar los parámetros de control. En esta página se puede fácilmente controlar alguno de los más importantes en la simulación. Como se puede ver en la Figura 5, al seleccionar los diferentes archivos de configuración, estos son cargados en la entrada correspondiente. Desde allí se pueden modificar y guardar los nuevos valores.

Figura 5. *Interfaz Web de control de simulación*

En la última pestaña se puede compilar los archivos fuentes del proyecto, ver las salidas generadas de este proceso y finalmente ejecutar la simulación. Cuando se ejecuta la misma, la visualización generada se puede ver en el navegador Web en diferentes formatos como pdf, png o ps. Una vez que obtuvieron los resultados, se puede descargar en formato comprimido absolutamente todos los archivos del proyecto.

Figura 6. *Interfaz Web de salida de simulación*

PRUEBAS

Para realizar las pruebas se desarrolló una aplicación denominada SimpleApp. Como su nombre indica, es una aplicación bastante simple. Su objetivo es simplemente para probar el comportamiento del simulador y la interfaz Web. El caso de prueba se describe a continuación: En la primera ronda de simulación, cada nodo envía un único mensaje. El nodo que recibe el mensaje imprime su propio identificador dentro de la red, así como el del remitente. Lo que se puede observar con esta aplicación es principalmente la conectividad de los nodos. Para simplificar en este caso solamente se ingresaron como parámetros a través de la interfaz Web seis parámetros; la cantidad de nodos (count), el rango de cobertura del sensor (range), las medidas de ancho y el alto del medio rectangular (width y height), la semilla (seed) de la simulación, y cantidad de iteraciones (max iterations).

En la Tabla 1 se puede visualizar los parámetros ingresados en la primera prueba.

Tabla 1. *Parámetros de control ingresados en la primera prueba de simulación*

PARÁMETRO	***VALOR***
count	*22*
range	*1*
width	*7*
height	*7*
seed	*1331177*
max iterations	*5*

Luego de correr la simulación, mediante las visualizaciones generadas (Figura 7) que algunos nodos no tienen conexión y quedan sin comunicación con los otros nodos de la red de sensores.

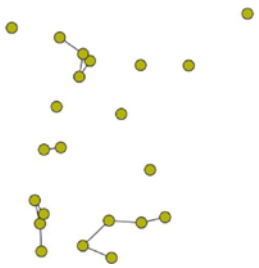


Figura 7. *Visualización generada por ShawnWEB de interconexión con algunos nodos sin conexión*

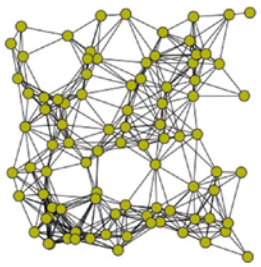


Figura 8. *Visualización generada por ShawnWEB de interconexión entre todos los nodos*

Modificando la cantidad de nodos y el rango de cobertura, como se puede visualizar en la Tabla 2 se puede observar que todos los nodos quedan totalmente interconectados (Figura 8). Mediante estas pruebas se pudo constatar las ventajas que otorga el prototipo desarrollado. Permite realizar en forma ágil modificaciones tanto en los parámetros de control como archivos fuentes de simulación. Además, luego de hacer las modificaciones, se puede fácilmente obtener visualizaciones para comprobar los efectos generados sobre la red.

Tabla 2*. Parámetros de control ingresados en la segunda prueba*

PARÁMETRO	***VALOR***
count	*90*
range	*1.5*
width	*7*
height	*7*
seed	*1331177*
max iterations	*5*

CONCLUSIONES Y TRABAJOS FUTUROS

En este proyecto de investigación se combinan tres ramas del conocimiento en Tecnologías de la Información y la Comunicación. La simulación, el desarrollo de proyectos de software con metodologías agiles y las redes de sensores inalámbricos enfocados en la industria y la academia.

En cuanto a la simulación y las redes de sensores inalámbricos se puede considerar que los objetivos han sido cumplidos, ya que se han logrado publicaciones en congresos y revistas internacionales, destacándose el trabajo "WebShawn, Simulating Wireless Sensors Networks from the Web" (Godoy, Sosa, Díaz Redondo, & Bareiro, 2017).

El desarrollo final del simulador servirá en un futuro para que investigadores y estudiantes puedan realizar simulaciones de WSN enfocándose solamente en la aplicación y el diseño de la red, sin necesidad de tener que lidiar con cuestiones de instalación y de funcionamiento de Shawn. Actualmente se utiliza en simulaciones para diversos proyectos de investigación relacionados con ciudades inteligentes como: a) Redes Inalámbricas de Sensores: Una experiencia en la Industria del Té, cuyo objetivo es realizar una implementación en una empresa productora de Té negro de la provincia de Misiones interesada en mantener mejor control en las variables del proceso utilizando una tecnología que es clave en Internet del Futuro; b) Análisis Y Comparación De Modelos De Propagación Para Optimizar La Localización Geográfica Del Ganado, cuyo objetivo es comparar modelos de propagación de señal de radiofrecuencia con el fin de optimizar la localización geográfica del ganado en un escenario de terrenos con topografía irregular y diferentes estratos de vegetación.

En el marco de este proyecto de investigación se han desarrollado más de diez trabajos finales de carrera de Ingeniería en Informática y Licenciatura en Sistemas de Información y continúan en ejecución otros cinco.

Entre las actividades de transferencia/extensión más destacadas se pueden mencionar la participación con un Tutorial denominado Internet de las Cosas y Ciudades Inteligentes en Simposio Argentino Sistemas Embebidos SASE 2017/2018 y la realización de Curso de Desarrollo de Aplicaciones Web en la UGD.

Se destaca en este sentido la participación como miembros plenos representando a la UGD como grupo en la Red Temática CYTED Red 518RT0558 denominada Ciudades Inteligentes Totalmente Integrales, Eficientes y Sostenibles (CITIES).

Se pretende, evaluar Frameworks de Internet de las Cosas y culminar con la construcción de una IDE para desplegar programas en Python en la CIAA.

En vistas de poder contar con una plataforma de pruebas de aplicación de las redes de sensores para Internet de las Cosas y Smart Cities, se pretende diseñar e implementar un sistema de estacionamiento inteligente para la sede del Campus Urbano Posadas de la UGD. Parte de esta plataforma permanecerá estable para recolectar datos que serán utilizados en futuros proyectos de I+D.

BIBLIOGRAFÍA

• Boehm, B. (1988). *A Spiral Model of Software Development and Enhancement*. IEEE Computer.

• Brambilla, M., Comai, S., Fraternali, P., & Matera, M. (Londres, 2008). *Designing web applications with WebML and WebRatio*. Web engineering: Springer.

• Byrnea, J., Heaveya, C., & Byrne, P. (2010). *A review of Web-based simulation and supporting tools*. Simulation Modelling Practice and Theory Vol 18: Elsevier Publisher.

• Ceri, S., Fraternali, P., & Bongio, A. (2000). *Web Modeling Language (WebML): a modeling language for designing Web sites*. Computer Networks.

• Ceri, S., Fraternali, P., Bongio, A., Brambilla, M., Comai, S., & Matera, M. (2003. ISBN 1-5586-0843-5). *Designing data-intensive Web applications*. Morgan Kaufmann.

• Coalesense. (2013). Obtenido de iSense Wireless Sensor Network Software: http://www.coalesenses.com/index.php?page=isense-software

• Fischer, S., Pfisterer, D., & Fekete, S. P. (2007. ISBN 1-4244-1231-5). *Shawn: The fast, highly customizable sensor network simulator*. Braunschweig University of Technology and University of Lubeck, Alemania.

• Godoy, D. A., Sosa, E. O., Díaz Redondo, R. P., & Bareiro, H. S. (2017). WebShawn, Simulating Wireless Sensors Networks from the Web. *IEEE WiMob – International Workshop on Cooperative Wireless Networks*. Roma: IEEE.

• Kroeller, A., Pfisterer, D., Buschmann, C., & Fekete, S. (2005). *Shawn: A new approach to simulating wireless sensor networks*. Alemania.

• NS2 Commnity. (2011). Obtenido de NS2 Wiki: http://nsnam.isi.edu/nsnam

• NS2 WEB Community. (2013). Obtenido de NS2 WEB : http://vlssit.iitkgp.ernet.in/ns2web/ns2web

• Pegden, C. D. (1983). *Introduction to SIMAN*. Proceedings of the 15th conference on Winter simulation-Volume 1. IEEE Press.

• Shawn. (6 de 2013). Obtenido de Shawn: The fast, highly customizable sensor network simulator: http://goo.gl/kCe9I3

• Shawn Community. (2013). Obtenido de Wiki de Shawn: https://github.com/itm/shawn/wiki

• Web-Ratio. (2014). Obtenido de Web-Ratio: La Nueva Ecuación de los Negocios-TI: http://www.webratio.com/

• Whitman, L., Huff, B., & Palaniswamy, S. (1998). *Commercial simulation over the web*. 30th conference on Winter simulation: IEEE Computer Society Press.

ANÁLISIS ECONÓMICO DE LOS SISTEMAS DE LABRANZA EN YERBALES DE OBERÁ

AUTORES: Bárbaro, S. Santa Clara, S. Kairiyama, T. Marín, R. Robulino, S. Widla, F.

RESUMEN

La débil capitalización de las chacras genera una gestión económica deficiente afectando el rendimiento de los cultivos tradicionales como la yerba mate (Ilex paraguarienses Saint Hill) (YM). Así como los sistemas de labranza afectan la sustentabilidad del suelo, por su efecto sobre la materia orgánica, también lo hacen sobre la rentabilidad de la producción. Cada sistema de control de malezas conlleva la utilización de diferentes insumos, equipos y mano de obra que harán más o menos beneficioso económicamente el control de malezas. El objetivo de esta investigación fue generar información sobre el costo resultante que genera cada uno de los sistemas de labranza empleado. Para evaluar los costos de los sistemas de control de malezas se instaló un ensayo de campo para evaluar la frecuencia y los insumos utilizados para casa sistema como así también el tiempo (jornales) que se emplea con cada uno. De manera complementaria para corroborar que la información obtenida en un ensayo controlado sea representativa de lo que realizan los productores se realizaron encuestas/entrevistas a productores yerbateros sobre los insumos, jornales y equipos necesarios en cada sistema de control.

Los resultados obtenidos en la investigación nos permiten concluir que el menor costo económico para el control de malezas se logra con la desmalezadora y la rastra traccionados por el tractor. El de mayor costo económico es con moto guadaña. Finalmente, el control de malezas con desmalezadora es lo adecuado teniendo en cuenta el costo económico y el menor daño en la sustentabilidad del suelo.

Palabras clave: sustentabilidad, yerba mate, malezas, costos.

INTRODUCCIÓN

Actualmente, en el sistema de producción de YM el control de malezas se realiza de la siguiente manera: a) con remoción de suelo, utilizando la rastra de discos b) realizando el control químico con herbicida y c) sin remoción de suelo, utilizando una desmalezadora de eje vertical enganchada a tres puntos del tractor o moto guadaña.

Desde un punto de vista económico, una maleza es una planta, cuya presencia genera la reducción de la rentabilidad del sistema agrícola. Cualquier planta no cultivable que aparece espontáneamente, son usualmente consideradas como malezas en los sistemas agrícolas altamente desarrollados. Éstas compiten con los cultivos por los nutrientes del suelo, el agua y la luz, además interfieren con la cosecha del cultivo e incrementan los costos de tales operaciones (FAO, 1996). Para que estas malezas no afecten la productividad de los cultivos se realizan diferentes prácticas de control buscando siempre la rentabilidad de la producción, controlando óptimamente los costos que generan dichas prácticas.

Para el control de estas malezas se incurren en gastos que afectan tanto a la sustentabilidad ambiental como a la rentabilidad económica de la producción. Según González y Paglietini (2001), en el costo de elaborar un producto, costo de producción, se deben computar todas las asignaciones que es necesario efectuar para garantizar la continuidad de la producción. En consecuencia, un costo no involucra únicamente los gastos en efectivo que deben efectuarse para lograr el producto total, sino que también incluye otro tipo de imputaciones y retribuciones que necesariamente deben considerarse a efectos de que los factores de la producción permanezcan en la empresa.

En lo referido a costos de producción de YM el mayor rubro lo representa la cosecha, la participación de los insumos es baja por la poca fertilización y la amortización (de la plantación) es alta pero el productor difícilmente lo tome en cuenta en consecuencia en la práctica, la producción de YM es muy intensiva en mano de obra lo que hace poco sustentable en el largo plazo (Lysiak, 2012). Por tanto el desafío en las plantaciones de YM es buscar el manejo más amigable con el medio ambiente y que además nos permita un rango de rentabilidad aceptable en la producción.

ANTECEDENTES

La superficie cultivada con Yerba Mate asciende a las 207.604 hectáreas involucrando a 17.841 productores, de los cuales el 75% producen YM en menos de 10 has (INYM, 2012). El rendimiento promedio general es de 5.400 kg de hoja verde por hectárea, siendo aún inferior los rendimientos del 75% de productores antes mencionado que es de 3178 kg/ha (INYM, 2012). Estos rendimientos están muy por debajo del potencial de una plantación de YM que debería promediar los 15000 kg/ha.

Según Acuña en Bongiovanni et al, 2012, los márgenes brutos de plantaciones de Yerba Mate fueron calculados para diferentes zonas agroeconómicas (Gunter et al, 2008) clasificando al sistema productivo en diferentes niveles tecnológicos por manejo de suelo, manejo de planta y fertilización. En dicho trabajo concluyen que para la zona de Oberá los márgenes brutos son reducidos para todos los niveles tecnológicos planteados.

En trabajos sobre análisis de costos realizados por Lysiak en Bongiovanni, 2008, resalta que el principal componente del costo de la YM representa la cosecha, sin embargo, la producción primaria de YM está compuesta por gran cantidad de pequeños productores quienes

no contabilizan el costo de la mano de obra familiar para los manejos del cultivo, entre ellos el control de malezas, pudiendo tener un mayor impacto en la grilla de costos dependiendo del sistema de labranza para controlas estas malas hierbas.

OBJETIVO

Calcular los costos de cada sistema de control de maleza y relacionarlos con el costo total de producción del cultivo y la sustentabilidad de la producción.

METODOLOGÍA

Para abordar el objetivo se utilizaron dos metodologías de trabajo a saber:

a) Se instaló un ensayo a campo en una parcela de YM donde se evaluaron cuatro sistemas de control de malezas (tratamientos), con el propósito, en primer lugar, de medir la frecuencia de control de cada sistema en segundo término, obtener información sobre los costos de cada uno (Fotos 1 y 2). Los tratamientos empleados fueron: rastra, desmalezadora, herbicida y moto guadaña. Para este ensayo se utilizó una parcela de 3.000 metros cuadrados en un lote de YM abarcando 15 líneos de 80 metros de largo con un marco de plantación de 3.5 mts entre líneos y 2 metros entre plantas quedando una superficie entre líneos de por 2,5 metros de ancho, zona donde se aplicaron los tratamientos.

El diseño estadístico utilizado fue de bloques completos aleatorizados. Se instalaron tres bloques o repeticiones por tratamiento. El motivo del bloqueo es por la pendiente.

Foto 1. *Localización del ensayo a campo en la zona rural de Campo Viera. Fuente: Google Earth Pro.*

Foto 2. *Instalación de ensayo en la parcela de yerba mate para medir la frecuencia de control de malezas con tres métodos. Fuente: Elaboración propia.*

b) Como segunda herramienta de análisis para llegar al objetivo se confeccionaron encuestas a productores mediante un cuestionario con preguntas preestablecidas para confirmar la representatividad del ensayo controlado. Además, se obtuvieron respuestas de los éstos que ayudaron a interpretar los resultados hallados en el ensayo y así tratar de comprender las acciones y decisiones adoptadas en el manejo de sus yerbales.

El cuestionario realizado fue el siguiente:

Nombre del productor:

- Fecha:
- Lugar: Campo Viera
- ¿Cuántas prácticas de control de malezas en yerba mate conoce?
- ¿Qué prácticas de control aplica en su explotación yerbatera?
- ¿Cuál es la razón por la que elige realiza esta/s práctica/s?
- En el caso del herbicida ¿Qué dosis utiliza?
- ¿Realizó alguna vez otro tipo de prácticas?
- ¿Qué maquinaria utiliza para realizar las prácticas?
- ¿Posee algún otro implemento para el control de malezas?
- ¿Cuál es el estado de longevidad de dichos implementos y en qué estado están?
- ¿Qué diferencias nota en los rendimientos de un año a otro?
- ¿A qué motivo considera que se debe esa baja en rendimientos?

RESULTADO Y DISCUSIÓN

El ensayo realizado en campo para relevar información sobre la frecuencia y costos de cada tratamiento se mantuvo por un período de 12 meses, tomados entre los meses de noviembre de 2017 y diciembre de 2018. Para los tratamientos rastra y desmalezadora (fotos 3 y 4), se utilizó un tractor Bertolini 445 de 45 hp; por tratamiento (tres líneos 600 metros cuadrados) el consumo fue de 0,54 litros de gas-oíl, por lo que se pudo determinar que el consumo en una hectárea fue de 9 litros de gas-oíl.

Foto 3, 4, 5 y 6.
Rastra de discos, tractor con desmalezadora, moto guadaña y mochila pulverizadora.

Para la aplicación del herbicida se utilizó una mochila pulverizadora marca "guaraní" (foto 5), con un pico de abanico y una dosis de 0,0125 litros de herbicida por cada litro de agua. El consumo en la superficie a muestrear fue de 10 litros de mezcla y 0,125 litros de producto, por lo que se determinó que el consumo de herbicida por hectárea fue de 2,08 litros en 166.66 litros de mezcla. Para el tratamiento con moto guadaña (foto 6) se utilizó una moto guadaña marca Villa modelo mx520 de 51.7 cc y 1.9 hp, el consumo en 600 metros cuadrados es de 0.75 litros lo que da como resultado 12.5 litros de nafta y 0.25 litros de aceite de mezcla (2T).

Los cálculos de los jornales o mano de obra fueron expresados en horas/hombre y los cálculos de amortización y mantenimientos están incluidos en el costo por hectárea y fueron calculados con los coeficientes de Fank (1977).

En las siguientes tablas 1, 2, 3 y 4 se pueden observar los datos relevados para los cálculos de costo expresados en dólares (1 US$= 38 pesos argentino).

DESCRIPCIÓN	***VALOR (US$)***
Valor tractorw	*$ 15.789,47*
Valor desmalezadora	*$ 1.000,00*
Valor rastra	*$ 1.578,95*
Valor pulverizadora	*$ 165,79*
Valor moto guadaña	*$ 315,79*
Valor hora hombre	*$ 1,97*
Valor gas oíl	*$ 1,05*
Valor nafta	*$ 1,11*
Valor glifosato	*$ 5,26*
Valor aceite 2T	*$ 8,95*

***Tabla 1**. Precio de las maquinarias e insumos. Fuente: Franco Widla.*

VIDA ÚTIL DE MAQUINARIA	***en Años***
Tractor	*15*
Rastra 14 discos	*20*
Desmalezadora	*20*
Mochila Pulverizadora (20 Lts)	*5*
Moto guadaña	*5*

***Tabla 2**. Vida útil de maquinarias. Fuente: Gunter (2002).*

HORAS HOMBRE X HECTAREA	
Rastra	*4*
Desmalezadora	*4*
Moto guadaña	*12*
Pulverización	*8*

***Tabla 3**. Horas de mano de obra por hectárea. Fuente: Franco Widla.*

CONSUMOS x LITROS x HECTAREA	
Gas oíl Tractor-Rastra	*9*
Gas oíl Tractor-Desmalezadora	*9*
Nafta - Moto guadaña	*12,50*
Aceite - Moto guadaña	*0,25*
Glifosato - Herbicida	*2,08*

***Tabla 4**. Consumo de insumos expresados en litros. Fuente: Franco Widla*

Con los datos de las anteriores tablas se realizaron los cálculos para obtener el costo por hectárea de cada tratamiento, también se contabilizo el número de veces necesario a entrar a controlar las malezas, la cantidad de veces fue de tres para cada tratamiento no encontrando diferencias entre los sistemas de control empleados. Los resultados se observan en la tabla 5.

TRATAMIENTO	COSTO UNITARIO x Ha	FRECUENCIA ANUAL	COSTO ANUAL x Ha
Rastra	22,28	3	66,84
Desmalezadora	22,05	3	66,14
Herbicida	31,36	3	94,07
Moto guadaña	42,53	3	127,59

***Tabla 5**. Costo en US$ de los sistemas de control de malezas evaluados. Fuente: Elaboración propia*

Los controles con rastra y desmalezadora resultan ser los menos costosos pero de una elevada inversión inicial (inversión en maquinaria), seguido por el control con herbicida, siendo el más costoso el control de malezas con moto guadaña. Estos últimos requieren una baja inversión inicial. De acuerdo a Lysiak (2012) técnico economista del INTA, la estructura de costos de la producción típica de yerba mate se reparte de la siguiente manera: Horas Hombre 12%, Gastos de maquinaria 3%, Insumos 2%, Cosecha y flete 50%, Amortizaciones 28% y otros 4%. Sin duda el gran costo lo lleva la cosecha (se consideraría un gasto fijo) y las horas hombre (dependen del tipo de control de malezas empleado). Si tenemos en cuenta el 50% restante, toda aquella práctica que se pueda hacer mecanizada con tractor, reduce sustantivamente los costos de producción. Según Acuña (2012), para la zona de Oberá se definieron tres niveles tecnológicos bajo, medio y alto. El ensayo a campo se realizó en un lote que concuerda en densidad de plantación y rendimientos con un nivel tecnológico medio, por lo tanto haremos los cálculos de la tabla 6 en base a una producción de 6000 kg de hoja verde (HV) expresado en US$ (1 US$= 38 pesos argentinos).

TRATAMIENTO	6000 kg HV x 0,22 US$	% afectado del ingreso bruto	Menos 50% cosecha	Menos costo de control de maleza	% afectado menos costo de cosecha
Rastra	1320	5,06	660	593,16	10,12
Desmalezadora	1320	5,01	660	593,86	10,02
Moto guadaña	1320	9,66	660	532,41	19,33
Herbicida	1320	7,12	660	565,95	14,25

***Tabla 6.** Porcentaje de afectación del costo del control de malezas y el ingreso bruto por ha en US$. Fuente: Elaboración propia*

Si consideramos casi como un empate técnico entre los costos del control de malezas con rastra y el desmalezado con tractor, pero incluimos en la discusión los beneficios sobre la conservación del suelo aportados por el desmalezado y además, una mayor fertilidad indicada por la materia orgánica (Barbaro et al, 2018), sin dudas que la opción del control de malezas con desmalezadora es la mejor entre las evaluadas. Además con dicho manejo de malezas se esperarían aumentos de los rendimientos como sucede con algunos productores (respuesta 10 de la encuesta) lo que bajaría el % de afectación del control de malezas en el costo de producción.

El control de malezas con moto guadaña también generaría una mejora en la calidad del suelo pero a un costo muy elevado. Este sistema de control de malezas sería indicado para plantaciones de alta densidad (alta tecnología) donde es imposible entrar conW tractor, teniendo en cuenta que en dichas plantaciones los rendimientos necesariamente deberían ser elevados para bajar los costos que demanda esta práctica. El control con herbicida se sitúa en un punto intermedio en costos económicos.

Del análisis surgido a partir de los testimonios obtenidos de las encuestas realizadas se pudo sintetizar las respuestas más frecuentes:

1. En cuanto a las prácticas de control de malezas en YM conocidas:

El productor conoce las siguientes formas: con rastra, desmalezadora, herbicida, macheteo con moto guadaña o machete y carpida con asada.

2. En relación a las prácticas de control aplicada en la explotación yerbatera del productor:

Las prácticas más utilizadas son: rastra y desmalezado con tractor, pulverización con mochila de 20 lts y algunos pocos con tanque pulverizador acoplado al tractor y también utilizan la moto guadaña.

3. Respecto de las razones por las que [el productor] elige realizar esta/s práctica/s:

Para aplastar la maleza, la desmalezadora con tractor o muto guadaña lo utilizan cuando la superficie es poca y el herbicida lo utilizan porque según el productor, tiene un bajo costo y porque demanda poco tiempo de trabajo.

4. Con respecto a la dosis de herbicida utilizada:

La normalmente utilizada de glifosato es de entre 200 a 300 cm3 por mochila de 20 lts.

5. Sobre si realizó alguna vez otro tipo de práctica:

Los productores no utilizan solo una manera de controlar las malezas, van rotando entre las nombradas en la respuesta 1.

6. En cuanto a la maquinaria utilizada para realizar las prácticas…

Las utilizadas son tractor, rastra de 12 a 14 discos, desmalezadora de enganche al tractor, mochila pulverizadora y moto guadaña.

7. En referencia a si posee algún otro implemento para el control de malezas:

Además de las herramientas tradicionales como el machete y la azada, en los últimos años están adquiriendo pulverizadoras de 200 a 400 litros acopladas al tractor.

8. En lo referente a la antigüedad de dichos implementos:

Los tractores tienen en promedio 40 años de antigüedad, la rastra 32 años, la desmalezadora 16, la mochila pulverizadora 3,5 años, y la moto guadaña 4 años.

9. En relación a si nota diferencias en los rendimientos de un año a otro:

La mitad de los productores encuestados aumentaron sus rendimientos y la otra mitad no notó diferencias de rendimientos.

10. En cuanto a qué motivos considera que se deben los cambios en los rendimientos:

Algunos productores atribuyen la mejora en los rendimientos el hecho de no utilizar más la rastra y a la fertilización de sus yerbales.

El productor realiza las prácticas evaluadas en la investigación, esto indica la representatividad del ensayo controlado a campo. No hay un único manejo de malezas utilizado, pero sí está claro que la utilización de la rastra es cada vez menor ya que se han dado cuenta del daño que produce en el suelo y la baja de rendimientos de sus yerbales. Las maquinarias utilizadas son en promedio muy antiguas. El productor, en vez de renovar su tractor o maquinarias opta por comprar mochila pulverizadora y moto guadaña que requieren una menor inversión y afrontar los mayores costos en mano de obra. La evaluación económica de la explotación agrícola es una poderosa herramienta en la que se sustentan muchas decisiones del productor. No obstante, su utilización no es generalizada, o se tiene un conocimiento empírico de las tareas realizadas y de los costos (Gunter 2002).

CONCLUSIONES FINALES

Por todo lo expuesto, podemos decir que:

• Se podría esperar una mayor frecuencia de control de malezas con el uso de la desmalezadora y moto guadaña, ya que no matan totalmente a las malezas y de acuerdo al tipo de clima que tenemos en Misiones su brotación sería rápida. Sin embargo, en el ensayo de campo que duró 12 meses no se han notados estas diferencias respecto a los demás sistemas de control evaluados.

• El menor costo económico para el control de malezas se logra con la desmalezadora y la rastra traccionados por el tractor.

• El de mayor costo económico es con moto guadaña.

• El control de malezas con desmalezadora sería lo adecuado teniendo en cuenta el costo económico y el menor daño en la sustentabilidad del suelo.

Agradecimientos:

A la Universidad Gastón Dachary por financiar el proyecto de investigación.

Al INTA por brindar información complementaria al proyecto.

Al Licenciado en Economía Emiliano Lysiak por sus aportes técnicos en la elaboración del presente trabajo.

BIBLIOGRAFÍA

• Acuña, D.O. (2012). Margen bruto de plantaciones de yerba mate. En: Bongiovanni, R.; Morandi, J.; Troilo, L. (Editores). (2012). *Competitividad y calidad de los cultivos industriales*. 1ra edic. Manfredi Córdoba, Ediciones INTA. 212 p.

• Barbaro S. E; Kairiyama T; Santa Clara S; Marín R; Sosa A; Iwasita B. (2018). *Forest replacement to Yerba Mate plantation and its effect on organic matter INTERNATIONAL UNION OF FOREST RESEARCH ORGANIZATIONS*. Conference Posadas, Misiones, Argentina.

• Bongiovanni, R. (Editor). (2008). *Economía de los cultivos industriales: Algodón, caña de azúcar, maní, tabaco, té y yerba mate*. 1ra edición, Manfredi Cordoba. INTA 2008. 108 p.

• Bongiovanni, R.; Morandi, J.; Troilo, L.; (editores). (2012). C*ompetitividad y calidad de los cultivos industriales: Caña de azúcar, mandioca, maní, tabaco, té y yerba mate*. 1ra edición, Manfredi, Cordoba. INTA 212p.

• FAO. (1996). *Manejo de malezas para países en desarrollo*. Estudio FAO, producción y protección vegetal, 120. Rome, Italy.

• FAO. 2002. Captura de carbono en el suelo para un mejor manejo de la tierra. Cap 3. Manejo de las tierras forestales, de pastoreo y cultivadas para aumentar la captura de carbono en los suelos. N° 96. Viale delle Terme di Caracalla, 00100 Rome, Italy.

• Frank, R.G. (1977). *Costos y administración de la maquinaria agrícola*. Buenos Aires. Edit. Hemisferio. Sur.385 p.

• Gunter, D.F; Colcombet, L; Kornoski, C; Kurt, V.D; Pereyra, L y Tkachuk, J.J. (2002). *Coeficientes técnicos para cálculos de costos de yerba mate*. Cerro Azul: 32 p. Miscelánea Nº 49.

• González, M.C y Pagliettini, L.L.(1983). *Costos de producción, unidad económica y tasaciones rurales*. Buenos Aires, 2da edición. 129p. Tesis UBA.

• INYM. (2012). Instituto Nacional de la Yerba Mate. Superficie cultivada por departamento. Recuperado de http://www.inym.org.ar/inym/imagenes/Estadisticas/sup%20cultivada%20depa.pdf.

• Lysiak, E. (2012). Los cuatro principales eslabones de la cadena de la yerba mate. En: Bongiovanni, R.; Morandi, J.; Troilo, L. (Editores).2012. *Competitividad y calidad de los cultivos industriales*. 1ra edic. Manfredi Córdoba, Ediciones INTA. 212 p.

TECNOLOGÍAS DE LA WEB SEMÁNTICA APLICADAS AL TRATAMIENTO DE DOCUMENTOS JURÍDICOS ELECTRÓNICOS

AUTORES: Ruidías H. J., Lezcano J. M., Eckert K, Caliusco M.L.

RESUMEN

En la actualidad los operadores jurídicos disponen de un acceso sin precedentes al corpus jurídico merced al auge de Internet y las tecnologías asociadas. Esto requiere de una forma de tratamiento documental que ponga atención no únicamente en el acceso y la manipulación de cantidades ingentes de documentos, sino también a la semántica de los mismos. Las tecnologías de la Web Semántica ofrecen una forma explícita de representación para organizar el contenido mediante ontologías, tesauros y otros mecanismos de vocabulario controlado, pero también permiten aplicar procesos de descubrimiento de relaciones a partir de mecanismos de razonamiento automático. En este modelo de relaciones semánticas explícitas, los mecanismos de búsquedas pueden resultar favorecidos al permitir búsquedas sobre un vocabulario menos dependiente de la simple ocurrencia terminológica, y sí más orientado a las estructuras de las categorías empleadas. En este artículo se presenta la problemática señalada y los resultados de una herramienta de búsqueda sobre fallos judiciales.

Palabras Claves: Digestos Jurídicos, Tesauro, Web Semántica, Búsqueda y Recuperación

INTRODUCCIÓN

El fenómeno de la inflación legislativa y la producción de documentos jurídicos hacen necesario desarrollar y promover plataformas y medios de comunicación e información que proporcionen oportunidades de acceder, compartir e intercambiar los recursos científicos, culturales, sociales y económicos que están disponibles en base de datos vinculadas con las ciencias jurídicas (Paul & Baron, 2007).

Atento a que dicho fenómeno se da también en un contexto de crecimiento exponencial de la información producida globalmente por organizaciones, empresas e individuos, alentada sin duda por la dinámica propia de Internet, es precisamente que también sobre la misma coyuntura científico-tecnológica se están desarrollando el mayor número de avances tecnológicos que abordan dicha problemática desde dos aspectos fundamentales: uno relacionado al tratamiento de volúmenes masivos de datos y el otro relacionado a la reducción de la opacidad semántica de dichos datos.

Es por ello que en forma concomitante al desarrollo de la Web, tal como fuera planteada por Sir Tim Berners Lee, también ha ido tomando vigor la idea de la Web Semántica (Berners-Lee, Hendler, Lassila, & others, 2001), donde la información no solamente resultara comprensible por seres humanos, sino también por máquinas.

En el terreno jurídico, el crecimiento de Internet ha propiciado que las normativas legales, digestos, fallos judiciales, sentencias y leyes estén a disposición de un mayor número de interesados a través de portales de acceso público (gobiernos provinciales o nacionales) o privado, los cuales cuentan con una forma de organización determinada (Ej. clasificación por "voces") y búsqueda por términos. A menudo esto último supone la coincidencia exacta de la ocurrencia de los términos en los textos que se devuelven en el resultado, o bien en algunos casos se ofrecen soluciones que permiten una tolerancia a errores en los términos de búsqueda aplicando algunas técnicas de búsqueda aproximada (Ej. utilizando funciones de semejanza). En todo caso, aun considerando la existencia de soluciones de propósito general respecto a la coincidencia exacta, se tendrá una búsqueda deficiente acorde a los requerimientos de un usuario particular de la documentación jurídica, ya sea este un justiciable, letrado, jurista o magistrado. La razón fundamental de ello, es que la documentación jurídica supone aspectos intrínsecos del objeto judicial que posee relaciones de categorías propias del derecho y de la legislación existente que muchas veces no aparecen en forma explícita y que requieren por lo tanto sean identificadas en un proceso de "curación" del documento legal. Un documento legal cuenta así con una riqueza semántica que a menudo se pierde al ser almacenada únicamente como texto plano.

El hecho de que la búsqueda de información jurídica sea deficiente, no resulta únicamente en un inconveniente de índole técnico, sino que atenta contra los mismos cimientos de la democracia al propiciar la falta de transparencia y celeridad necesaria en la resolución de casos judiciales que esconden un entramado de conflictos sociales (Brenna, 2014; Ciuro Caldani, 2007).

EL ORDENAMIENTO JURÍDICO

Los digestos jurídicos han tenido como resultado del análisis normativo, epistemológico y documental, el conocimiento de la validez normativa, que permitió estudiar y evaluar las distintas patologías del cuerpo normativo en general, tales como "las abrogaciones y derogaciones implícitas, las contradicciones normativas, las equivalencias normativas, las pérdidas de vigencia, las interdependencias normativas y las inconsistencias normativas, con un alcance lógico y lingüístico. Para obtener

textos ordenados de su articulado y una propuesta de unificación de los mismos"[1]. Dichos análisis (construcción teórica) a menudo desconocido por operadores del derecho, puede no obstante mejorar sustancialmente el uso y el conocimiento del digesto jurídico, relacionándolo con la práctica profesional; ya que al pensar conjunta y cooperativamente los aspectos teóricos del mismo con el de la práctica de la abogacía, sin dudas se logra el enriquecimiento de la tarea profesional desde la realidad jurídica práctica.

Precisamente y teniendo en cuenta lo mencionado con anterioridad se debe tener en cuenta tres ejes fundamentales de la práctica profesional que son: SABER-HACER-SER. A partir de ello es que toma relevancia la pregunta que alude a qué saberes debería tener el abogado para lograr abarcar las tareas específicas que el mercado laboral demanda ya sea en el estado, empresas, estudios jurídicos, etc.

Seguramente la respuesta sobre qué habilidades debía poseer y por sobre todo qué saber van a tener un punto en común, que es el de la normativa legal vigente, ya que al conocer la estructura, caracteres y especificaciones del Digesto jurídico, el profesional podrá acceder de una forma rápida y segura a dicho saber[2].

TECNOLOGÍAS APLICADAS AL DIGESTO JURÍDICO

En el derecho, la interpretación es un eje fundamental sobre el que se articula el accionar aplicativo de la justicia, y el deterioro legislativo debido a la sobresaturación de leyes y normas, puede suponer un reto a la hora de conocer la vigencia real sobre cierta normativa. Todo esto pone sobre relieve la necesidad de establecer un proceso de ordenamiento jurídico (Brenna, 2014) que culmina en la creación de un Digesto Jurídico. Tal proceso supone básicamente el escrutinio permanente sobre el crecimiento legislativo y el consecuente ordenamiento del mismo. Existen tres categorías fundamentales de intervención que se consideran y que definen la complejidad del proceso (consolidación), que incluye 1) un aspecto sistemático, es decir la forma en que se organizará la normativa, 2) otro lingüístico (morfológico, sintáctico y semántico) y 3) uno de interrelación normativa. Estas tres categorías resultan de vital importancia, no únicamente para el ordenamiento jurídico, sino para la digitalización y su tratamiento a través de sistemas de gestión documental, que permitan almacenar y acceder a dicha normativa.

La creación de un Digesto Jurídico por otra parte supone la conformación a una doctrina, un marco jurídico y normativo, y también el entrecruzamiento entre las ramas clásicas del derecho y otras transversales tales como "equidad de género", "biodiversidad", "derecho a la salud" entre otras (Ciuro Caldani, 2007).

En el marco de un sistema de gestión documental enriquecido con tecnologías semánticas, un Digesto Jurídico se define a partir de vinculaciones semánticas (anotado semántico) entre el texto de la norma y una ontología de referencia. Para ello lo apropiado supone la reutilización de ontologías de alto nivel (Kaneiwa, Iwazume, & Fukuda, 2007; Mascardi, Locoro, & Rosso, 2010) para objetos que no son propios del derecho, tales como aquellos que refieren a eventos temporales y ubicaciones espaciales, y para aquellos que pertenecen a categorías propias del Derecho, emplear ontologías de dominio tales como FOLAW o LRI-CORE (Benjamins, 2005; Breukers & Hoekstra, 2004)

Un antecedente destacado que se puede mencionar es el e-COURT de la Unión Europea, un sistema para manejo de información legal, textual y multimedia, que permite el almacenamiento y recuperación de la misma, y que es tolerante a ambigüedades de índole lingüística gracias a las tecnologías semánticas que emplea (Breuker, Elhag, Petkov, & Winkels, 2002). Otro antecedente es el Digesto Jurídico basado en ontologías creado para la Universidad Nacional del Litoral (Enrique & López, 2016).

TECNOLOGÍAS SEMÁNTICAS Y APLICACIONES PARA EL ORDENAMIENTO JURÍDICO

Los aspectos intrínsecos de las tecnologías semánticas que pueden fomentar por un lado, la codificación y el ordenamiento jurídico antes señalados, supone a priori un trabajo pormenorizado de legisladores y juristas en establecer categorías, relaciones y normativas asociadas, con el fin de lograr una mejora en los procesos de almacenamiento y organización del corpus de documentos jurídicos, pero por si presenta la necesidad de que dicha información pueda ser sometida a procesos de recuperación de información de relevancia en los procesos de impartición de justicia por parte de los distintos actores jurídicos. No obstante, conviene señalar otro aspecto, que se desprende de los aspectos intrínsecos antes mencionados, y es que las ontologías obligan a repensar las relaciones establecidas en el mundo.

Así también, podríamos alegar que la ontología jurídica responde al reclamo de la denominaciones o concepto que utilizamos, y es aquella que trata de asignarle, en el derecho un significado peculiar, aunque abstracto. Por ello parece necesario, para aclarar los conceptos fundamentales del derecho, examinar someramente aquellas consideraciones ontológicas susceptibles de aportar alguna precisión al conocimiento en las ciencias jurídicas desde la perspectiva de su realidad, para luego fijar en qué, sentido y con, qué alcance tales precisiones puede ser parte de la ciencia jurídica de los juristas.

ONTOLOGÍAS

Las ontologías han sido propuestas como artefactos de representación que especifican un vocabulario relativo a cierto dominio en el contexto de los sistemas de información (Guarino, 1998). Dicho vocabulario define entidades, clases, propiedades, axiomas y reglas, además de las relaciones entre dichos componentes, con el objetivo de reducir la ambigüedad, los conflictos terminológicos y las discrepancias semánticas que se presentan en un área de conocimiento determinada. Esto conduce a una definición donde una ontología es entendida como "una especificación formal y explícita de una conceptualización compartida" (Gruber, 1993). Se destaca en la misma dos ideas claves, la de conceptualización compartida y la de especificación formal. La primera refiere a un modelo abstracto, con sus elementos y la vinculación entre estos, explícitamente definida y además cumple con el requisito de ser consensuada y aceptada por una comunidad. El segundo en tanto aborda la cuestión del lenguaje y el vocabulario con que se representa, de tal

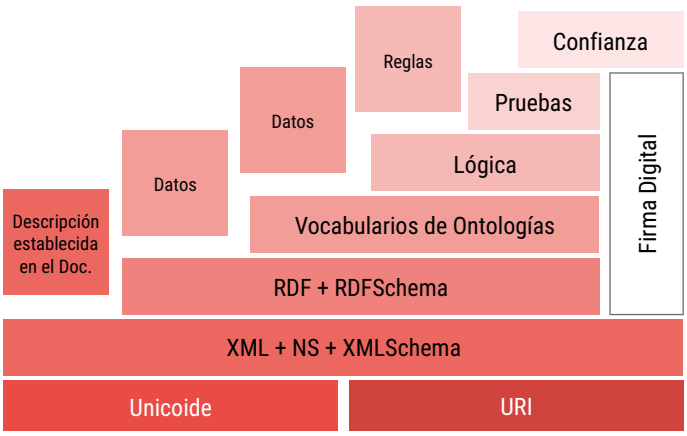
1 Ver tercer informe de avance del D.J.M en línea en http://www.diputadosmisiones.gov.ar/uploads/digesto juridico_3erinforme_cer.pdf

2 Las consultas del D.J.M están online en http://www.diputadosmisiones.gov.ar/digesto juridico_juridico/content.php?id_category=179

manera que sea interpretable de igual manera tanto por personas como por computadoras, siendo por lo tanto una cualidad irrenunciable el que sea procesable por máquina.

El énfasis en el aspecto "consensuado" que propician las tecnologías de la Web Semántica va desde la estandarización y especificaciones, la arquitectura de capas de la W3C presentada en la figura 1, hasta las definiciones de ontologías de alto nivel (o fundacionales) o específicas para distintos dominios, ya sean médicos, de organizaciones y personas o de índole legal.

En lo que respecta a las tecnologías presentadas en la figura 1, a modo de resumen se puede mencionar el conjunto de tecnologías de base sobre las que se construye la Web actual, que incluye una forma univoca de identificar recursos (URI), una codificación estándar (UNICODE), y un lenguaje básico para estructurar documentos (XML+XMLSchema) y sobre el cuál se respaldan los distintos lenguajes de marcado, tales como xHTML y HTML. Sobre tales tecnologías se cimientan las tecnologías específicas de la Web Semántica, que incluyen descriptores de recursos y relaciones entre estos (RDF y RDFSchema) y un lenguaje específico de ontologías (OWL), que aporta mayor expresividad al permitir describir propiedades y clases con restricciones de cardinalidad, exclusión (relaciones disjoint), y una mayor riqueza a la hora de definir propiedades (relaciones). Naturalmente tales características imponen algunas restricciones en función de la expresividad deseada, lo cual permite diferenciar entre tres tipos básicos de familias de ontologías, OWL Lite, OWL DL y OWL FULL.



Las ontologías han demostrado su utilidad desde sus orígenes al ofrecer un marco conceptual útil para reducir la ambigüedad inherente al uso del lenguaje natural. Puesto que en muchos dominios del conocimiento humano se establecen definiciones en términos de construcciones verbales susceptibles a varias interpretaciones, principalmente aquellas que se vinculadas a las ciencias sociales.

TAXONOMÍAS Y TESAUROS

Las definiciones de ontologías antes presentadas pueden suponer la ocurrencia de algunos errores de conceptos, principalmente si puede notarse que comparte descripciones con otros mecanismos de vocabularios controlados, que en algunos casos puede resultar suficientes dadas las características del problema a tratar.

Es por ello que, entre los vocabularios de términos, la estructura más elemental puede ser el de una simple taxonomía, la que, a modo de un sistema de directorio, permite organizar el contenido en relaciones jerárquicas simples.

Está claro que todos los mecanismos de vocabularios controlados que se desarrollan en este capítulo responden a la misma necesidad de reducir la ambigüedad, y sentar las bases para la elaboración de sistemas de nomenclatura (ya sea universal o no), que no solo permita la organización interna de contenidos, sino también la comunicación y transferencia de información de un sistema a otro.

Es por ello que una taxonomía resulta de mucha utilidad para dicha empresa, ofreciendo un esquema simple pero funcional, de términos organizados en forma jerárquica.

Respecto a los tesauros difiere básicamente en que se limita a un solo tipo de relación de padre e hijo entre términos, omitiendo por ejemplo relaciones de sinonimia, o de otra naturaleza. Tampoco se pueden definir atributos (como sí se puede hacer con las ontologías). Es por ello que un Tesauro supone un nivel de mayor complejidad de tesauro controlado, normalmente adhiriendo a algún estar o convención respecto a las notaciones empleadas.

Un tesauro puede definirse como vocabulario controlado organizado formalmente con objeto de hacer explícitas las relaciones semánticas y genéricas que se aplican a una parte especifica del conocimiento, a menudo empleados para organizar, clasificar, codificar y decodificar información con finalidad de que se recuperada por un sistema informático.

Los vocabularios controlados, tales como los tesauros son regulados a partir de especificaciones y estándares que facilitan la interoperabilidad de los mismos a través de diferentes sistemas, tales como el estándar ISO 25964-2:2013 (ISO 25964-2:2013, 2013)

RECUPERACIÓN DE INFORMACIÓN

En la tarea de Recuperación de Información (RI, en inglés information retrieval), la relevancia denota en qué medida un documento de una colección satisface la necesidad de información de un usuario. El acceso a información relevante se ve limitado por el volumen de datos que constantemente se encuentra en aumento.

Según informa la Internacional Data Corporation (IDC), los datos globales aumentarán de 33 ZB[3] en 2018 a 175 ZB para el 2025. Se deduce entonces que pese a que un usuario disponga del acceso a toda esta información (exhaustividad), la misma será inútil ya que dicho usuario deberá realizar el trabajo de búsqueda, lo cual –debido a un volumen tan elevado de información– resulta inviable. Por este motivo, la precisión en los motores de búsqueda y modelos (técnicas, enfoques) de RI son vitales para la obtención de resultados que satisfagan las necesidades del usuario; e indefectiblemente se requiere de la utilización de nuevos modelos y estrategias capaces de abordar el crecimiento exponencial de dicha información.

En el dominio jurídico, la Recuperación de Información Legal (RIL, en inglés legal information retrieval) incide en gran manera en el desempeño profesional de los especialistas en Derecho. La búsqueda y RI en este ámbito se diferencia de otros dado que se considera muy compleja

3 Un zettabyte equivale a 1021 bytes.

y sus procesos dependen en gran medida de la interpretación del conocimiento por un experto humano. Dicha complejidad está dada por la estructura sintáctica y la terminología dependiente del dominio.

En contraste a otras disciplinas las taxonomías rara vez son inherentes a la ley. Los vocabularios legales contienen términos abiertos, siendo intrínsecamente dinámicos y cuyas aplicaciones dependen de las interpretaciones de los especialistas, resultando en términos legales semánticamente ambiguos.

Por lo tanto, las búsquedas booleanas basadas en la aparición y repetición de los términos de la consulta ingresada por el usuario, pueden resultar ineficientes dado que no consideran un análisis global en el cual se ponderen los aspectos semánticos de dicha consulta. Servicios como Westlaw, LexisNexis y Findlaw también implementan este tipo de búsquedas.

Este método para calificar la relevancia se ve limitado al hecho de que la consulta del usuario posea una ocurrencia exacta de los términos y a la repetición de los mismos en los documentos. De este modo, pueden existir palabras no pertenecientes a la consulta –como ser sinónimos, hiperónimos, hipónimos, jerarquías entre términos, ontologías y tesauros– las cuales a pesar de que se presenten en el documento y estén fuertemente relacionados con dicha consulta, no inciden en la relevancia del mismo ya que no se atienden las relaciones entre vocablos (ontologías); es decir, este método de relevancia no contempla la semántica entre el documento y los términos ingresados para la recuperación –y menos aún la distancia semántica entre los mismos–.

Por otro lado, si los términos de una consulta no pertenecen al dominio –como ser términos en relación con aspectos culturales, folksonomías, medios de comunicación, redes sociales, etc.–, existen escasas probabilidades de recuperar documentos pertinentes. Esto también se ve reflejado con la aparición de nuevas figuras legales –homicidio agravado, femicidio, etc.– que no tienen lugar en los documentos anteriores a las mismas de modo que, al no contemplar la semántica del texto, la búsqueda no recupera –o lo hace en forma ineficaz– documentos relevantes.

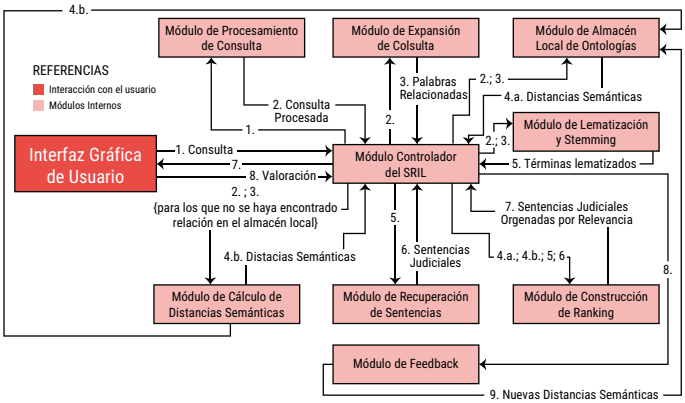
No considerar los factores semánticos y los aspectos propios del dominio legal pueden conducir a búsquedas cuyos resultados presenten documentos no relevantes, clasificados como relevantes –o documentos relevantes clasificados como irrelevantes–, no respondiendo así a las necesidades de los usuarios.

En este trabajo se ha diseñado e implementado un modelo basado en ontologías –tanto legales como genéricas y de propósito general– y distancias semánticas para mejorar la consulta del usuario y obtener sentencias relevantes. Esto se fundamenta en el hecho de que las ontologías aseguran una recuperación eficiente al permitir inferencias basadas en el conocimiento del dominio.

Se utilizaron diversas ontologías con dos objetivos principales. El primer objetivo consiste en establecer una relación entre una consulta ingresada por el usuario –que puede no estar expresada con vocablos relativos al ámbito legal– y términos legales. Como segundo objetivo se procuró expandir la consulta del usuario y, mediante la utilización de distancias semánticas generar un ranking –dado que la exhaustividad resulta inviable– para la presentación de los resultados al usuario.

MODELO PROPUESTO

El modelo propuesto para la recuperación y clasificación de documentos legales se encuentra representado en la Figura 2. Mediante este modelo se busca establecer la relevancia de los documentos, otorgando como resultado final un ranking a partir del cual el usuario puede vislumbrar un orden basado en este enfoque de búsqueda. Los módulos del mismo serán descritos a continuación.



Para llevar a cabo la implementación del modelo se utilizó un entorno de ejecución para JavaScript conocido como Node, debido a la gran compatibilidad con diferentes tecnologías y APIs utilizadas en este trabajo.

Más específicamente para la interacción con el usuario (frontend) se utilizó el framework conocido como Angular y para el desarrollo interno del backend, Express (Agostinho, Dutra, Jardim-Gonçalves, Ghodous, & Steiger-Garção, 2007; Google, 2019; StrongLoop, Inc. y otros colaboradores de expressjs.com, 2019).

MÓDULO DE PROCESAMIENTO DE CONSULTA

Este módulo se encarga de mejorar los términos de la consulta ingresada por el usuario. Al recibir la consulta, se identifican los términos y/o frases que forman parte de la misma, proceso similar a la tokenización. Por cada aparición del carácter especial "+" los vocablos (y frases) son separados para evaluarse individualmente. Sucesivamente, para cada uno de éstos, se realiza la corrección ortográfica (en español) – proceso conocido como Spelling – de los tokens involucrados en la consulta. Además, se eliminan signos y caracteres tales como "!", "@", ";", "#", entre otros.

Este procesamiento incide en otros módulos dado que la modificación de los vocablos repercutirá en las búsquedas de relaciones con otros términos y cálculos de distancias semánticas.

MÓDULO DE EXPANSIÓN DE CONSULTA

La utilidad de la tarea de expansión de consultas realizada en este módulo se ve manifiesta en que las palabras claves de una consulta pueden no encontrarse textualmente en los documentos, pero sí sus propiedades y/o relaciones definidas en las ontologías. Por lo tanto, se procura expandir una consulta de búsqueda a partir de las palabras claves, mediante la obtención de sinónimos y términos relacionados (hiperónimos, hipónimos, merónimos) en las ontologías. En el ámbito de la RI, las ontologías permiten expandir las consultas. Análogamente sucede con la RIL y las ontologías legales (Wagh & Anand, 2017).

En este trabajo se utilizaron distintas ontologías y redes semánticas; algunas de éstas con relaciones propias del ámbito legal y otras de propósito general dado que los usuarios pueden expresar sus consultas con términos que no estén referidos específicamente al ámbito jurídico. Por

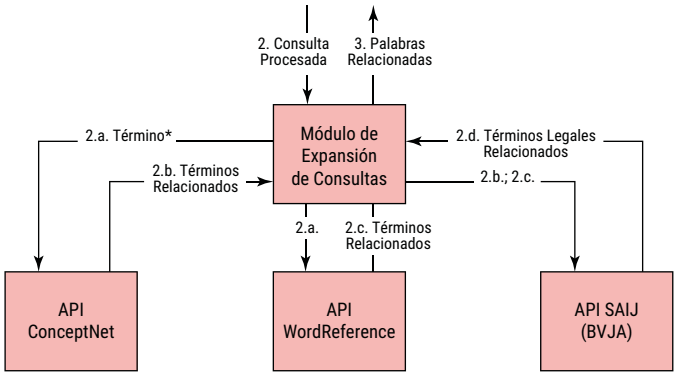
lo tanto, se busca establecer relaciones semánticas entre los términos, que permitan relacionar la consulta del usuario con los vocablos propios de los documentos legales.

Seguidamente se definen las ontologías y redes semánticas utilizadas en este trabajo:

• ConceptNet: donde se obtienen relaciones como "Synonym" (sinónimos), "RelatedTo" (relacionado con), "IsA" (es un), "DerivedFrom" (derivado de), "EtymologicallyRelatedTo" (etimológicamente relacionado con), entre otras.

• WordReference: servicio web empleado para la obtención de sinónimos.

• BVJA (API de SAIJ): El Sistema Argentino de Información Jurídica (SAIJ) ofrece un servicio abierto al Tesauro Saij de Derecho Argentino, de donde se obtienen relaciones como "usado por" (UP), "término general" (TG), "término específico" (TE), "término relacionado" (TR), entre otros.



* Por cada término individual de la consulta se obtienen sus relacionados.

Figura 3: *Interacción del Módulo de Expansión de Consultas con APIs externas.*

El módulo presentado en la Figura 3 se encarga, a partir de cada uno de los tokens o frases individuales, de obtener otros términos y relaciones. Primeramente, se realizan peticiones a ConceptNet y WordReference y luego, por medio de este subconjunto de términos obtenidos (más los tokens), se solicita al BJVA los términos legales relacionados a los anteriores.

Esta tarea aporta flexibilidad dado que las redes semánticas van cambiando conforme avanza el tiempo, por lo cual se obtiene una mayor versatilidad al no depender de términos locales estáticos en el tiempo.

El usuario accede al servicio a través de una interfaz gráfica basada en tecnologías Web, la cual permite el ingreso de los términos de búsqueda tal como se consigna en la Figura 4 para luego procesar la consulta del lado del servidor y devolver los resultados en forma de una lista donde los resultados aparecen ordenados por la relevancia en función de las ontologías y las distancias semánticas, según se puede apreciar en la Figura 5.

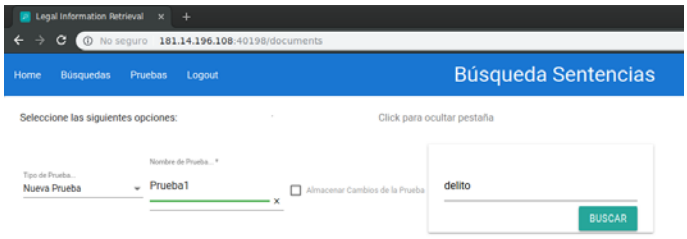


Figura 4: *Ingreso de los términos de consulta (aplicación alojada en el servidor de UGD).*

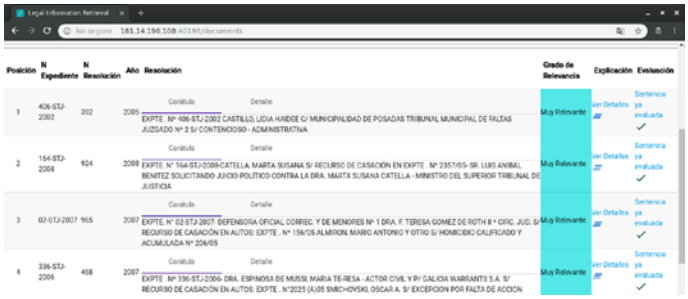


Figura 5: *Pantalla de resultados del prototipo (alojado en el servidor de UGD).*

CONCLUSIONES

El trabajo desarrollado sentará las bases para la transferencia de soluciones tecnológicas al sector judicial, partiendo desde un prototipo funcional, que empleó datos aportados desde el Poder Judicial de la Provincia de Misiones mediante la firma de un convenio de transferencia de conocimientos entre la Universidad Gastón Dachary y dicho organismo estatal. Asimismo, desde el punto de vista académico, desde el trabajo llevado a cabo con tecnologías semánticas, se puede ver ya no solo un campo fértil para la aplicación de tales tecnologías, sino también la posibilidad de acceder a una fuente de escenarios de prueba para la validación de las propuestas desarrolladas tanto en el marco del presente proyecto como de futuros proyectos.

También en el desarrollo de la investigación se ha conformado y consolidado un grupo de investigación de tecnologías semánticas y sus aplicaciones en la UGD, a la vez que se ha fortalecido las relaciones en ciencia y tecnología con la Universidad Tecnológica Nacional, Facultad Regional Santa Fe, y en particular con el grupo de la Dra. María Laura Caliusco perteneciente al Centro de Investigación y Desarrollo de Ingeniería en Sistemas de Información de dicha facultad.

BIBLIOGRAFÍA

• Agostinho, C., Dutra, M., Jardim-Gonçalves, R., Ghodous, P., & Steiger-Garção, A. (2007). EXPRESS to OWL morphism: making possible to enrich ISO10303 Modules. En G. L. BSc MSc & R. C. BEng (Eds.), *Complex Systems Concurrent Engineering* (pp. 391–402). Recuperado de http://link.springer.com/chapter/10.1007/978-1-84628-976-7_44

• Benjamins, R. (2005). *Law and the Semantic Web - Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*. Recuperado de http://www.springer.com/us/book/9783540250630

• Brenna, R. G. (2014, junio 16). Thomson Reuters | EL DIGESTO JURÍDICO ARGENTINO. Primer paso en el camino de la democratización de la información jurídica. Recuperado el 3 de junio de 2019, de http://thomsonreuterslatam.com/2014/06/el-digesto-juridico-argentino-primer-paso-en-el-camino-de-la-democratizacion-de-la-informacion-juridica/

• Breuker, J., Elhag, A., Petkov, E., & Winkels, R. (2002). Ontologies for legal information serving and knowledge management. Legal Knowledge and Information Systems, *Jurix 2002: The Fifteenth Annual Conference*, 1–10.

• Breukers, J. a. P. J., & Hoekstra, R. J. (2004). Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law. *CEUR Workshop Proceedings*. Recuperado de http://dare.uva.nl/record/1/424828

• Ciuro Caldani, M. A. (2007). EL COMPLEJO DEL FUNCIONAMIENTO DE LAS NORMAS. *INVESTIGACIÓN Y DOCENCIA, 40*. Recuperado de http://www.centrodefilosofia.org.ar/IyD/iyd40_4.pdf

• Enrique, F., & López, M. (2016). *Desarrollo de algoritmos de procesamiento para la indización y la búsqueda en lenguaje natural del contenido de un digesto basado en tecnologías semánticas*. Universidad Nacional del Litoral.

• Google. (2019, abril 17). Angular. Recuperado el 17 de abril de 2019, de https://angular.io/

• Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowl. Acquis., 5*, 199–220. https://doi.org/10.1006/knac.1993.1008

• Guarino, N. (1998). *Formal Ontology in Information Systems: Proceedings of the First International Conference (FOIS'98), June 6-8, Trento, Italy*. IOS Press.

• ISO 25964-2:2013. (2013). ISO 25964-2:2013. Recuperado el 2 de mayo de 2019, de ISO website: http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/36/53658.html

• Kaneiwa, K., Iwazume, M., & Fukuda, K. (2007). An Upper Ontology for Event Classifications and Relations. En M. A. Orgun & J. Thornton (Eds.), *AI 2007: Advances in Artificial Intelligence* (pp. 394–403). https://doi.org/10.1007/978-3-540-76928-6_41

• Mascardi, V., Locoro, A., & Rosso, P. (2010). Automatic Ontology Matching via Upper Ontologies: A Systematic Evaluation. *IEEE Transactions on Knowledge and Data Engineering, 22*(5), 609–623. https://doi.org/10.1109/TKDE.2009.154

• StrongLoop, Inc. y otros colaboradores de expressjs.com. (2019, abril 17). Express - Infraestructura de aplicaciones web Node.js. Recuperado el 17 de abril de 2019, de https://expressjs.com/es/

• Trinkunas, J., & Vasilecas, O. (2009). Ontology Transformation: from Requirements to Conceptual Model. *Scientific Papers, University of Latvia, 751*, 52–64.

• Wagh, R., & Anand, D. (2017). Application of citation network analysis for improved similarity index estimation of legal case documents: A study. *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, 1–5. https://doi.org/10.1109/ICCTAC.2017.8249996

APLICACIÓN DE HERRAMIENTAS DE IBM WATSON SOBRE EL DOMINIO DEL CLÚSTER DEL TÉ

DIRECTOR: Karanik, M. J. **INVESTIGADORES:** Suénaga, R. Favret, F. Eckert, K. B. **COLABORADORES:** Kairiyama, T. Candia, G. Alegre, G.

RESUMEN

El área de análisis de grandes volúmenes de información ha tomado una relevancia excepcional dentro de las TICs (Tecnologías de la Información y las Comunicaciones). Esto se debe a que la tecnología, a medida que evoluciona, facilita el desarrollo de algoritmos para tareas de análisis de datos, tales como clasificación, búsqueda de patrones, determinación de tendencias, construcción de modelos descriptivos y predictivos, entre otras. Como consecuencia, las técnicas y algoritmos son cada vez más eficientes y producen resultados más precisos, convirtiéndolos en herramientas apropiadas para la obtención de información útil.

En el trabajo de Eckert K. et al. (WICC 2016) se utilizan técnicas y procesos de búsqueda con resultados sobre información publicada en relación a la producción y comercialización del té a nivel global, a partir del cual se requieren procesos de mejoras para destacar aquellas publicaciones que sean más representativas según los parámetros indicados por el usuario (empresa u organismo vinculado con el Cluster del Té de Misiones).

En base a lo expuesto antes, este trabajo comprende el estudio e implementación de técnicas de análisis de información útil en base a herramientas de la plataforma Watson de la empresa IBM. Específicamente se utilizaron las herramientas Watson Knowledge Studio y Watson Discovery Service para el análisis y puntuación de documentos según su importancia teniendo en cuenta los parámetros que representan los requerimientos del usuario.

Con la puntuación asignada se establecieron ranking de los documentos considerados, los que fueron verificados por los usuarios en representación del Cluster del Té de Misiones, obteniendo resultados dispares que merecerían la comprobación a partir de un conjunto mayor de documentos, probablemente bajo una temática diferente para lograr una contrastación significativa.

INTRODUCCIÓN

En un entorno tan complejo y con tanta incertidumbre como lo es la economía global en general y, en particular la economía en Argentina, se puede observar que, a diferencia de las grandes empresas, las microempresas de las economías regionales descuidan el manejo de la información estratégica para potenciar sus negocios. Este tipo de comportamiento está asociado a un problema complejo de adaptación que les impide a estas organizaciones lograr un crecimiento sostenido y sustentable en el tiempo.

Un caso evidente de esta situación es el Clúster del Té de la Provincia de Misiones que aglomera a microempresas dedicadas a la producción, industrialización y comercialización del Té. Estas microempresas por lo general son familiares y una de las causas de la mencionada desatención se debe a que cuentan con escasos recursos para analizar y resolver problemas estratégicos. Es decir, es prácticamente imposible que este tipo de empresas cuente con recursos que se dediquen a buscar y analizar información para luego pensar cómo aplicarla en la solución de problemas estratégicos. Esto sucede porque todos sus integrantes están abocados a lidiar con problemas operativos cotidianos.

Esta problemática no es nueva y, claramente, la posibilidad de añadir información de calidad al proceso de toma de decisiones, se vuelve una tarea irrealizable si se pretende analizar todas las fuentes disponibles manualmente. Por ello, en sintonía con el avance en las Tecnologías de la Información y de las Comunicaciones (TIC), los Sistemas de Soporte de Decisiones (SSD) han incorporado capacidades para el análisis de información que le permiten determinar los grados de relevancia de la misma en base a los requerimientos definidos por el usuario.

Por lo general, los SSD obtienen los requerimientos de información mediante la interacción con los usuarios. De esta manera se intenta acotar el dominio de problema y qué tipo de información se requiere. Luego, se indican las fuentes y el SSD comienza el proceso de análisis basándose en los criterios definidos anteriormente. Este tipo de herramientas, si bien poseen una autonomía limitada, no requieren de la constante supervisión del usuario para realizar el proceso de análisis y luego mostrar los resultados obtenidos.

Para mantener la ventaja competitiva, las organizaciones necesitan establecer una estrategia. El punto de partida para la formulación de esa estrategia es identificar y valorar los recursos y capacidades disponibles en la organización (Quero L. 2008).

Pero no todos los conocimientos de una organización se convierten en fuente de ventaja competitiva sostenible, lo serán solamente aquellos que contribuyan a la generación de valor económico. Aquí, el conocimiento se entiende también como habilidades, experiencia, información contextualizada, valores, actitudes, know how, que en conjunto se denominan conocimientos esenciales.

ANÁLISIS SECTOR TEALERO

Es sabido que, en un escenario globalizado, el conocimiento que hace posible la innovación, se ha convertido en una fuente de ventaja competitiva sostenida y, en consecuencia, de crecimiento y sostenibilidad para las organizaciones.

En este sentido, la Provincia de Misiones, junto a las provincias de Chaco, Formosa y Corrientes, conforma la Región Noreste de Argentina (NEA), la misma se extiende en una superficie de 29.801 km², representando sólo el 0,8% de la superficie nacional. Un 90% de sus fronteras son internacionales, limitando al oeste con la República del Paraguay,

al norte y este con la República Federativa de Brasil y al sur, con la Provincia de Corrientes. Posee una población total de 1.101.593 habitantes y manifiesta un aumento del 14% intercensal 2001-2010 siendo una de las Provincias más densamente pobladas de la República Argentina. El 73,7 % es población urbana mientras que la población rural representa el 26,3 % (21, 09 % es población rural dispersa y el restante 5,21 % población rural agrupada).

La producción tealera se concentra en la provincia de Misiones y en noreste de Corrientes. Conforman la zona más austral del mundo en la que se cultiva el té. La provincia de Misiones cuenta con más de 38.000 ha implantadas, correspondiendo al 95 % de la superficie de té cultivada en el país (el otro 5% corresponde a la provincia de Corrientes). Aporta el 17,3% al valor de las exportaciones de la provincia de Misiones, y es un garante de ingreso y empleo de la población local. Aproximadamente 6800 productores cultivan té en Misiones. El 93% de ellos tiene menos de 10 ha, el 5% hasta 50 ha y sólo el 0,5%, más de 50 ha. Hay una gran cantidad de productores pequeños (microemprendimientos familiares) que surgen, en general, como herramientas de supervivencia de los trabajadores y sus familias, con pocas hectáreas implantadas, siendo el eslabón más débil de la cadena y el más afectado ante los cambios que involucran al sector (Ministerio de Economía y Finanzas. 2013).

Dentro de este contexto, el Clúster del Té de Misiones surge como una estrategia de las Microempresas (de producción, de industrialización y de comercialización) que conforman la cadena tealera, para aunar esfuerzos que, de manera aislada, no alcanzan para obtener resultados que les permitan ser competitivas. En estas estructuras se pueden detectar relevantes ineficiencias en su cadena de valor (tanto en costo como en tiempo), y encuentran en el clúster un instrumento básico para mejorar las relaciones cliente-proveedor y una plataforma para articular la cooperación entre empresas.

Dentro de las actividades principales se distinguen un primer grupo, denominado extractores de materia prima, un segundo nivel, de transformadores intermedios o de primera transformación, una tercera agrupación denominada creadores de producto o segunda transformación, y un cuarto grupo que son los comercializadores.

HERRAMIENTAS UTILIZADAS

Utilizando las credenciales de IBM Watson, se analizaron cada una de las herramientas que podían ser de utilidad para los fines del proyecto y considerando la aplicabilidad de dichas herramientas al contexto del dominio. La herramienta que cumple con los requisitos de aplicabilidad Watson Discovery Service (WDS). Los conceptos principales de la herramienta fueron estudiados y los requisitos preparados para proceder a la aplicación de las mismas en el contexto del proyecto. Se consideró WDS para categorizar y analizar documentos específicos del dominio, considerando la serie de pasos que propone IBM para la creación de un Proyecto en WDS y posterior realización de consultas en el Lenguaje de Consultas de Discovery (LCD).

NECESIDADES DE INFORMACIÓN DEL DOMINIO DE APLICACIÓN

El dominio de aplicación propuesto por el Clúster del Té de Misiones, se basó en la identificación de propiedades fisicoquímicas del té, su origen, entre otros. Las mismas se detallan en la Tabla 1.

Tema general	*Tea physical-chemical properties*
Características particulares del tema	*Antioxidants, Minerals, Vitamins, Energizing*
Ámbito	*Global*
Objetivo	*Searching for analysis and studies in producing countries / search for research on tea properties*
Fuentes de datos	*Research papers, journal articles, conferences*
Temas a excluir	*Buy, purchase, sale, advertising, offers*

Tabla 1. *Parámetros de Búsqueda*

Estos parámetros refieren a una consulta que fue realizada previamente durante las pruebas del Sistema de Búsqueda Lexicográfico (SBL) (Eckert K. et al. 2016 y Favret F. et al. 2016), que se encarga de obtener los primeros resultados devueltos por motores de búsqueda tradicionales, y luego explora de manera independiente los recursos enlazados a dichos resultados, proporcionando como salida una serie de documentos Web representativos de tales parámetros de búsqueda. Dichos resultados se encuentran listados en un ranking según su relevancia respecto a los parámetros especificados anteriormente, donde cada documento tiene una calificación, dada por el experto en el dominio de aplicación. De esta manera, al considerar durante la aplicación de los recursos de IBM Watson los mismos parámetros que los expuestos anteriormente, se busca obtener una medida comparativa entre ambas tecnologías: dichos recursos y el SBL. En consecuencia con lo mencionado, y teniendo en cuenta los parámetros de interés, se realizó un análisis exploratorio de los documentos calificados con los puntajes más altos en el ranking por el experto en el dominio, buscando establecer el modo de relacionar dicho contenido con los parámetros de interés. Ello puso en evidencia la necesidad de traducir tales parámetros al modo de funcionamiento propio de las herramientas a emplear, lo cual implica la definición de Entidades, y Relaciones entre estas Entidades, propias del dominio.

RELACIÓN	***ENTIDADES***
TIENE	*TE* *PROPIEDAD*
USADO_PARA	*TE* *USO*
PROVIENE_DE	*TE* *ORIGEN*

Tabla 2. *Entidades y Relaciones definidas para el dominio del Clúster del Té*

En este contexto, las Entidades representan la forma en que se categorizan elementos del mundo real, mientras que las Relaciones definen una correspondencia binaria y ordenada entre dichas entidades. Las entidades y relaciones que se definieron son las presentadas en la Tabla 2.

De esta manera, la entidad TE hace referencia a las distintas menciones referidas al té que pueden ocurrir en los documentos. Por ejemplo, "tea", "oolong tea" y "black tea". De manera similar, PROPIEDAD refiere a ocurrencias de menciones relacionadas con propiedades fisicoquímicas, tales como "antioxidants", "energizing" y "polysaccharide".

Las entidades y relaciones específicas del dominio deben ser introducidas a los recursos de IBM Watson mediante un proceso de entrenamiento, que luego es utilizado para extraer menciones a dichas entidades y relaciones a fin de determinar la relevancia de nuevos documentos de texto. Este entrenamiento consiste en, a partir de un corpus de documentos de ejemplo, indicar cada una de las palabras que hacen referencia a una entidad, y cada par de entidades que hacen referencia a una relación. Como resultado del entrenamiento se obtiene un modelo de Machine Learning que luego es explotado para identificar entidades y relaciones de nuevos documentos de texto. Después, según la existencia y cantidad de ocurrencias de tales atributos en los documentos, es que se termina por definir la relevancia de cada uno, conformando finalmente un ranking de documentos según qué tan bien se ajusten éstos a dichas entidades y relaciones, definidas por la consulta de interés.

Los servicios de IBM Watson que permiten realizar estos procedimientos son Watson Knowledge Studio (WKS) y Watson Discovery Service (WDS). WKS se utiliza para el entrenamiento del modelo de machine learning. WDS se utiliza para la extracción de entidades y relaciones, y ordenamiento consecuente de los documentos en un ranking.

PROCESOS REALIZADOS

A continuación se describe el proceso seguido durante la utilización de estos recursos a fin de alcanzar los objetivos del proyecto de investigación. Se pueden identificar tres etapas principales: el entrenamiento de un modelo de machine learning en WKS, el despliegue de dicho modelo sobre WDS, y la ejecución de consultas sobre WDS para la obtención de rankings. A continuación se describen los pasos comprendidos durante estas tres etapas. Los pasos descritos se realizan utilizando la interfaz gráfica provista por IBM para gestionar los servicios mencionados.

ENTRENAMIENTO DEL MODELO DE MACHINE LEARNING (WKS):

1. Definición del Tipo de Sistema (Type System): consiste en introducir los tipos de entidades y relaciones relevantes del dominio.

2. Definición del diccionario: los diccionarios permiten agrupar las palabras y frases que deberían ser tratadas de manera equivalente por el modelo de Machine Learning (ML) en una lista, cuya utilidad se expresa en el punto 6. Estos diccionarios se crean agregando una entrada por cada ocurrencia que, de suceder en el texto, debe ser catalogada como perteneciente a cierta entidad.

3. Carga de documentos: este paso consiste en cargar los documentos de texto de ejemplo, que luego deberán ser etiquetados. En este caso se subió un documento de texto plano, cuyo contenido fue extraído de una publicación científica relacionada a uno de los resultados devueltos por el SBL.

4. Crear Conjuntos de Anotación (Annotation Sets): los conjuntos de anotación son agrupaciones de documentos que luego serán asignados a personas específicas que se encargarán de etiquetar los documentos contenidos en dichos conjuntos. En este caso, como se trabajó con un solo documento de entrenamiento, se creó un solo conjunto de anotación. En "Base set" se elige el conjunto de documentos al que se asociará el de anotación. "Overlap" es un porcentaje que se utiliza para definir la proporción de documentos del conjunto que deberá superponerse entre los distintos anotadores humanos, cuando el proyecto incluye más de un anotador. Dicho parámetro permite hacer un seguimiento respecto de si los anotadores etiquetan de la misma manera estos documentos superpuestos. "Annotator" permite seleccionar el nombre de la/s persona/s que etiquetarán los documentos. "Set name" es el nombre que se le desea dar al conjunto de anotación.

5. Pre-etiquetado: este paso permite acelerar el proceso de etiquetado mediante un pre-etiquetado, en donde el sistema realiza un etiquetado automático asignando a cada ocurrencia de las palabras del diccionario, su entidad correspondiente. La opción "Apply This Pre-annotator" da inicio al pre-etiquetado con los diccionarios definidos.

6. Crear tarea: una vez creado el conjunto de anotación, y asignados los anotadores humanos, es necesario crear una tarea. La creación de una tarea consiste en establecer una fecha límite para la finalización del etiquetado de un conjunto de anotación, lo cual habilita al quien hace el etiquetado a comenzar dicho proceso.

7. Etiquetado: al etiquetar, el primer paso es seleccionar, de entre los documentos del conjunto de anotación, aquel que se va a etiquetar. Luego se comienza el proceso de etiquetado, donde la recomendación es comenzar por etiquetar todas las entidades, y luego etiquetar las relaciones.

a. Etiquetado de Entidades: se selecciona una palabra (o grupo de palabras consecutivas) que refieren a una entidad.

b. Etiquetado de Relaciones: una vez etiquetadas las entidades, para el etiquetado de relaciones se procede seleccionando las menciones correspondientes a las entidades

8. Entrenamiento: una vez finalizado el etiquetado de los documentos del conjunto de anotación, que en este caso fue el archivo relacionado a la publicación científica, se accede a la ventana de "Performance", y en la misma se hace clic a "Train and evaluate", para comenzar el entrenamiento y la evaluación del modelo de machine learning; o sobre "Train", en caso que no se desee evaluar la performance del modelo. Luego se selecciona el conjunto de anotación con el cual realizar el entrenamiento, y los porcentajes para el conjunto de Training, Test y Blind. En este caso, como se contaba con un solo documento dentro del conjunto, se destinó el 100% de documentos al entrenamiento.

Finalizado el entrenamiento, el modelo se encuentra listo para su despliegue sobre WDS.

VINCULACIÓN DE WKS CON WDS:

1. Despliegue de WKS sobre WDS: una vez entrenado el modelo de ML, WKS permite gestionar las versiones de dicho modelo, donde cada versión puede ser desplegada a uno (o varios) servicios, tales como WDS. Para realizarlo, a partir de la solapa "Versions", se selecciona la opción "Deploy" y se identifica el recurso sobre el cual se desea desplegar el modelo. Finalizado el despliegue, el servicio proporciona un identificador del modelo, que será necesario para el paso siguiente.

2. Enriquecimiento de WDS: a partir del identificador obtenido en el paso previo, se accede al recurso de WDS, y se crea una colección. Una colección consiste en un conjunto de documentos. Luego, dicha colección es enriquecida con el modelo de ML, agregando las opciones "Entity Extraction" y "Relation Extraction", e introduciendo el identificador del modelo en el campo destinado a tal fin, según se muestra en la Figura 1.

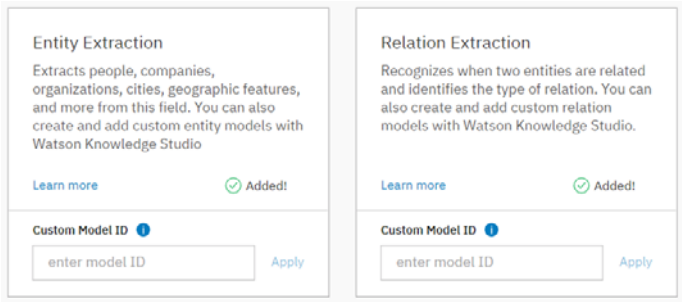


Figura 1. *Enriquecimiento de una colección en WDS*

3. Subir documentos: finalmente, se agregan los documentos a la colección. Los documentos fueron obtenidos en formato HTML utilizando la consola de comandos de cURL para descargarlos a partir de los links del ranking. Una vez que se suben los documentos, Discovery los enriquece automáticamente con análisis de las entidades y relaciones entrenadas, además de otros parámetros, como ser el sentimiento y la emoción que, según interpreta el sistema, despierta el contenido de cada documento.

UTILIZACIÓN DE WDS

1. Operadores de consulta: los operadores son separadores entre distintas partes de la consulta. Entre los operadores que más se utilizaron se encuentran:

a. Delimitador JSON (.): separa los niveles de jerarquía en el esquema JSON.

b. Incluye (:): especifica una coincidencia para el término de la consulta.

c. Coincidencia exacta (::): especifica una coincidencia exacta para el término de la consulta.

d. Caracter de escape (/): para las consultas que tienen la necesidad de consultar términos con literales de series que contienen caracteres que de otra manera serían de control.

e. Consulta de frase (""): se utiliza para analizar todo el contenido de una consulta de frase de forma literal, es decir, sin analizar caracteres especiales (excepto si la frase posee comillas dobles, donde debe usarse un operador de escape).

f. Agrupación anidada ([], ()): se pueden definir para detallar información más específica.

g. Or (|): operador booleano "or".

h. And (,): operador booleano "and".

2. Especificación de consultas: el servicio de WDS ofrece potentes funciones de búsqueda mediante consultas, una vez que el contenido se cargó y el servicio lo ha enriquecido con extracción de entidades y relaciones. Este servicio genera un documento JSON luego de enriquecer los documentos en su recopilación, y es necesario conocerlo para poder crear consultas con Discovery Query Language (DQL). Este documento JSON es jerárquico, de forma que las consultas se deben de escribir siguiendo esa misma jerarquía. Para un JSON con la estructura definida en la Figura 2, la consulta debería ser estructurada según se presenta en la Figura 3.

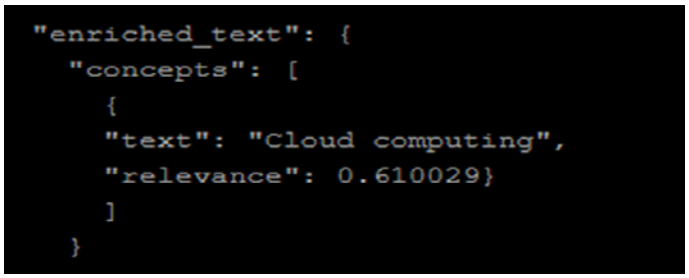


Figura 2. *Ejemplo de respuesta JSON a una consulta realizada en WDS*

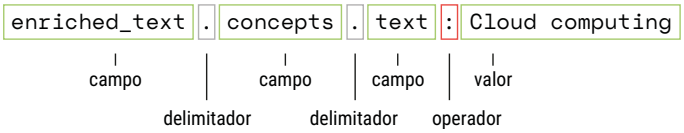


Figura 3. *Estructura genérica de una consulta en DQL*

Una vez realizada la consulta, el resultado es obtenido en forma de un ranking que se puede visualizar dentro del contenido de la página Web, en formato HTML. Dicho ranking muestra, para cada documento, el título correspondiente al mismo según la etiqueta "title" de cada HTML. A partir de dichos atributos, se confecciona una planilla de cálculo donde el ordenamiento de los títulos se corresponde con el otorgado por WDS según su relevancia respecto a la consulta dada.

RESULTADOS Y CONCLUSIONES

Una vez traducidos los parámetros de interés a consultas con análisis de entidades y relaciones, se obtuvo como resultado de cada consulta, un listado de documentos ordenados de manera descendente según la relevancia de los mismos en relación a los parámetros de consulta. Dichos listados pueden considerarse como rankings.

Para establecer una medida de qué tan representativo es un ranking, lo que se consideró fue la media de la distancia entre la posición en el ranking de cada documento, y la que debería ocupar según la puntuación que le asignó el experto. Para realizar este cálculo, se trasladaron cada uno de los rankings a una planilla de cálculo. Luego, cada ranking fue ordenado de manera descendente según las puntuaciones asignadas a los documentos por el experto del dominio. En este paso, se calculó

la media de la distancia entre la posición en el ranking de cada documento, y la que debería ocupar según dicha puntuación asignada. De esta forma, a cada ranking se le asoció una distancia media, representativa de su precisión. Así, cada ranking se acerca más al deseado en tanto y en cuanto dicha distancia sea más cercana a cero.

Para cada una de las consultas mencionadas en la sección anterior, WDS devuelve como resultado un listado de documentos ordenados de manera descendente según la relevancia de los mismos en relación a los parámetros de consulta, el cual puede considerarse como un ranking.

Para el análisis comparativo de los resultados, en primer lugar se trasladan los listados obtenidos a una planilla de cálculo. Luego, y para cada listado, partiendo de las puntuaciones asignadas a los documentos por el experto del dominio, se ordena en forma descendente, y se calcula una media de la distancia entre la posición en el ranking de cada documento, y la que debería ocupar según dicha puntuación asignada. De esta forma, el ranking obtenido por cada consulta tiene asociada una distancia media, representativa de su precisión, que será mejor cuanto menor sea su valor nominal.

Tabla 3. *Análisis comparativo de los rankings obtenidos utilizando Conjunción*

	DQL con modelo ML	***DQL sin modelo ML***
Documentos Obtenidos	*5*	*17*
Distancia Promedio	*0*	*2*
Puntuación Promedio	*5*	*3.94*

El corpus original obtenido por SBL consta de 50 documentos, de los cuales 12 tienen una puntuación máxima (5) y los demás tienen una valoración entre 4 y 0; por lo que la distancia promedio para SBL, según el ranking obtenido es de 9.64. En la Tabla 3 se muestra la cantidad de documentos obtenidos, la distancia media y la puntuación promedio obtenida por (a) la consulta en DQL utilizando parámetros de entrenamiento y (b) la consulta en DQL sin utilizar parámetros de entrenamiento; utilizando el operador AND (conjunción), para restringir los resultados obtenidos.

Con el uso del modelo de ML se obtiene como resultado 5 documentos, en este caso todos ellos tienen la puntuación superior (5), por lo que la distancia promedio es cero (0); lo cual indica que dicho wmodelo retorna documentos de buena calidad, en función a lo evaluado por el experto, pero de los mismos descarta 7 documentos con dicha valoración (además de los 38 documentos con calificación inferior). Por otro lado, la no utilización del entrenamiento (DQL sin modelo ML) devuelve una mayor cantidad de documentos (17), con una distancia promedio de 2, dado que se obtienen 11 de los 12 documentos con valoración 5, en su mayoría ubicados en las primeras posiciones; adicionalmente se incluyen documentos con valoraciones más bajas (entre 1 y 3).

Los resultados expuestos (Tabla 6) se encuentran ligados a las restricciones que conlleva el uso de los operadores lógicos de conjunción (AND), los cuales descartan los resultados en caso de no cumplir con todos los parámetros dados. En busca de una mejoría en la recuperación de documentos con buenas valoraciones, se utilizaron operadores de disyunción (OR), cuyos resultados se exponen en la Tabla 4.

Tabla 4. *Análisis comparativo de los rankings obtenidos con Disyunción*

	DQL con modelo ML	***DQL sin modelo ML***
Documentos Obtenidos	*20*	*22*
Distancia Promedio	*1.90*	*3.77*
Puntuación Promedio	*3.80*	*3.64*

Como se puede observar (Tabla 4), la cantidad de documentos devueltos por DQL con y sin modelo ML aumentan al utilizar disyunción en los parámetros de búsqueda (20 y 22 documentos respectivamente); en este caso, de los 12 documentos con valoración 5, se obtuvieron 11; los demás documentos cuentan con una valoración inferior, entre 4 y 1 respectivamente.

Finalmente, tomando en consideración los 18 documentos con valoraciones superiores (5, 4 y 3), al evaluar la inclusión y exclusión de documentos en DQL, sin considerar el orden dentro del ranking; se puede mencionar que empleando DQL con ML y conjunción, se seleccionan 5 documentos con valoración 5, mientras que para las demás pruebas se recuperan 11 de los 12, esto conlleva a una precisión del 41,67% y del 91,67% respectivamente. En el caso de documentos con puntuación 4, al utilizar disyunción con o sin ML no se incluyen resultados, y al utilizar conjunción en ambos casos se recuperan 1 de los 2 etiquetados con dicho valor. Por último, en el caso de documentos valorados con 3, el orden fue de manera creciente en los resultados: 0, 1, 2 y 3 respectivamente, de un total de 4 documentos esperados.

De los resultados obtenidos se concluye que todos los documentos incluidos en el análisis contenían algún tipo de información vinculada con los requerimientos, pero la influencia del etiquetado y el entrenamiento califican determinados documentos más o menos importantes de los que el usuario los asignó, lo que se podría verificar con el tratamiento de una cantidad mayor de documentos.

BIBLIOGRAFÍA

• Eckert K., Alvarenga V. M., Barboza M., Witzke L. M., and Araldi L. (2016). *Vigilancia tecnológica e inteligencia competitiva basada en técnicas de minería de la web*, presented at the XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).

• Eckert K., Favret F., Barboza M., Witzke L. M. and Alvarenga V. M. (2016). *Modelos de análisis de información para la toma de decisiones estratégicas del sector tealero*, presented at the XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina).

• Favret F., Montiel R., Alvarenga V., Barboza M. and Witzke L. (2016). R*ecuperación de información basada en técnicas de minería Web*, p. 7.

• Ministerio de Economía y Finanzas. (2013). *Plan de Competitividad Conglomerado Productivo Tealero de la provincia de Misiones*.

• Quero Luisa. (2008). *Estrategias Competitivas: Factor Clave de Desarrollo*.

MODELO DE ANÁLISIS DE INFORMACIÓN DESESTRUCTURADA UTILIZANDO TÉCNICAS DE RECOPILACIÓN Y MINERÍA WEB

DIRECTOR: Karanik, M. J. **INVESTIGADORES:** Suénaga, R. Favret, F. Eckert, K. B. **COLABORADORES:** Rojas, M. Pfeifer, H.

RESUMEN

Cuando se hacen búsquedas de información en internet, saber exactamente lo que se busca es primordial para localizar rápidamente lo que se requiere. Pero cuando lo que se necesita es identificar qué documentos hay acerca de un tema en particular (sin identificar su denominación precisa), los buscadores identifican infinidad de sitios y documentos que hace casi imposible revisarlos individualmente hasta dar con aquellos que se acercan a lo que al usuario le interesa.

Esa situación requiere de soluciones especializadas que utilicen técnicas de interpretación de contenido de páginas web y documentos de internet que, a partir de una especificación de lo requerido por el usuario, busque, analice y clasifique la información disponible de acuerdo a lo solicitado.

Este trabajo aborda la situación descripta precedentemente, a partir del cual se propone desarrollar y utilizar un modelo para el proceso de identificación de sitios y documentos, basado en la integración de técnicas de recopilación, exploración y análisis de información en la web.

Primeramente se describen las características de los procesos de identificación de información en internet, como así también las métricas de evaluación que permiten identificar la relevancia de la información de interés.

En el desarrollo del trabajo se propone un modelo correspondiente a un proceso que consiste en la utilización de resultados de los buscadores de internet (Google, Bing, MSXML Excite e Intelligo), a partir del cual se desencadena un proceso de exploración de los enlaces de cada sitio identificado, para luego proceder a asignar puntajes de acercamiento a los requerimientos del usuario (uno de los parámetros significativos es asignado por el análisis semántico), finalmente se ordenan los documentos de acuerdo a los valores asignados (raniking).

El modelo se probó en dos escenarios, uno referido a información sobre herramientas de educación digital y el segundo referido a información sobre técnicas vinculadas con seguridad informática. Las pruebas del modelo proporcionaron ordenamientos que ubicaron mejor a los recursos más relevantes en los rankings construidos, los que fueron validados por usuarios especializados que configuraron los escenarios de búsqueda inicial.

Palabras clave: Minería web; análisis semántico; recuperación de información.

INTRODUCCIÓN

Durante mucho tiempo las dificultades de acceso a las fuentes de información fue el factor preponderante para conseguir datos fiables, pero con el avance de las tecnologías de la información y las comunicaciones (TICs) este inconveniente fue desapareciendo dando lugar a otro con el mismo efecto: la sobresaturación de información. Por ello, existen grandes volúmenes de información que no pueden ser manejados con métodos tradicionales debido al nivel de desestructuración, su disponibilidad en distintos formatos, que se encuentran parcial o totalmente desconectados y están altamente distribuidos.

El área de análisis de grandes volúmenes de información ha tomado una relevancia excepcional dentro de las TICs. Esto se debe a que la tecnología, a medida que evoluciona, facilita el desarrollo de algoritmos para tareas de análisis de datos, tales como clasificación, búsqueda de patrones, determinación de tendencias, construcción de modelos descriptivos y predictivos, entre otras. Como consecuencia, las técnicas y algoritmos son cada vez más eficientes y producen resultados más precisos, convirtiéndolos en herramientas apropiadas para la obtención de información útil.

En base a lo expuesto antes, este proyecto abarca el estudio e implementación de técnicas de búsqueda y análisis de información útil. Específicamente se busca modelar un sistema integral de información que incluya investigación sobre sistemas de recopilación de necesidades, búsqueda automática, exploración y minería web y herramientas de toma de decisiones.

RECUPERACIÓN DE INFORMACIÓN Y RECUPERACIÓN DE DATOS

Existen diferencias marcadas entre la recuperación de información (RI) y la recuperación de datos (RD). Estas diferencias están relacionadas a los tipos de objetos con los que trata cada una, la representación de estos objetos, la especificación de las consultas y los resultados que se obtienen. En primer lugar, la RD trata con valores y claves de búsqueda con una estructura definida, mientras que la RI debe lidiar, en ambos casos, con las dificultades del procesamiento del lenguaje natural. Por otro lado, se puede ver a la RD como una aproximación a la RI, en las situaciones en que se busca determinar los recursos de una colección que contienen las palabras de la clave de búsqueda ingresada por el usuario. Sin embargo, desde el punto de vista de la RI, lo más probable es que los resultados obtenidos sean irrelevantes para la necesidad de información que el usuario posee, ya que la ocurrencia de palabras de la

clave de búsqueda en un recurso, no es suficiente como para afirmar si éste es relevante o no (Baeza-Yates y Ribeiro-Neto, 1999).

RELEVANCIA Y MÉTRICAS DE EVALUACIÓN

Determinar si un recurso es relevante es una actividad compleja, debido a que se trata de un juicio subjetivo, lo que significa, que dos usuarios pueden determinar niveles de relevancia distintos para un mismo recurso. Por lo tanto, se hace evidente que la relevancia consiste en realidad en la consideración de varios factores, entre los que se encuentran las características del recurso, las características de la necesidad de información y la subjetividad del usuario que elabora la clave de búsqueda (Tolosa et al. 2008).

Con respecto a esto, en Blair D. (1990) se afirma que es más fácil llevar a cabo la determinación de la relevancia de un recurso determinado, que explicar cómo fue obtenida o qué criterios se utilizaron.

Con el objetivo de evaluar, que tan acertada es la respuesta proporcionada al usuario y, por lo tanto, que tan efectiva es la determinación de relevancia del sistema de recuperación de información (SRI), es necesario utilizar un conjunto de métricas que permitan llevar a cabo dicha labor. Existen varias métricas que permiten evaluar la calidad de los resultados obtenidos, de las cuales las más utilizadas son la precisión y exhaustividad, planteadas en Cleverdon C. (1966).

La precisión se define como la proporción de los recursos recuperados que son relevantes y permite evaluar la habilidad del sistema para generar un ranking en el que las primeras posiciones estén ocupadas por recursos relevantes (Landauer y Laham. 1998).

La exhaustividad se define como la proporción de los recursos relevantes que han sido recuperados y permite evaluar la habilidad del sistema para encontrar todos los recursos relevantes de una colección (Baeza-Yates y Ribeiro-Neto, 1999).

MODELOS PARA RECUPERACIÓN DE INFORMACIÓN

Existen varios enfoques de modelos de RI, que representan distintos puntos de vista. Gran parte de estos, se centran en el cálculo de la probabilidad de que el contenido de los recursos se relacione al contenido de la clave de búsqueda. Algunos modelos de RI más representativos, son: el modelo Booleano, el Vectorial, el Probabilístico y el enfoque Semántico.

MODELO PROPUESTO

En Eckert et al. (2016) y Favret et al. (2016) se describe un modelo de RI que tiene dos objetivos bien definidos: el primero consiste en capturar de manera precisa la necesidad de información que posee el usuario y presentar los resultados obtenidos. El segundo, consiste en la implementación de métodos y técnicas que permitan la identificación, recuperación y análisis continuo de recursos existentes en la Web. En principio, los dos objetivos mencionados, son implementados mediante el *Módulo de Recopilación de Requerimientos* (MRR) y el *Módulo de Minería Web* (MMW), respectivamente.

Con el fin de implementar un sistema que lleve a cabo la determinación de la relevancia de recursos, mediante Técnicas Semánticas (TS) ya implementas en [6] [7] [8], se define la arquitectura del sistema, que se presenta en la Figura 1.

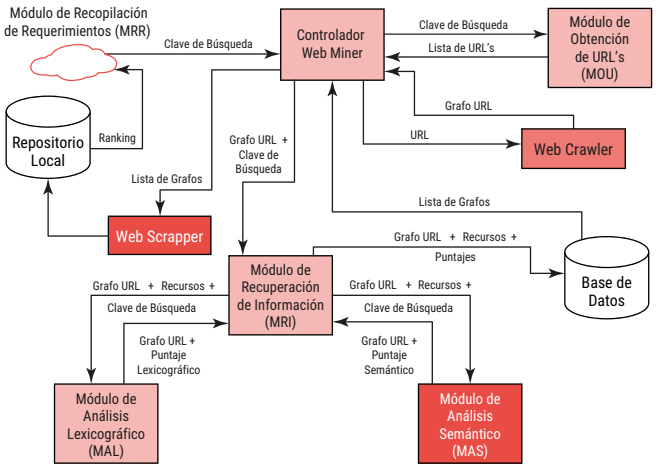


Figura 1. *Arquitectura léxico-semántica*

El proceso de análisis de recursos y generación de rankings comienza cuando el MRR envía la clave de búsqueda ingresada por el usuario al Controlador Web Miner, que se encarga de coordinar el funcionamiento de todo el sistema. Esta clave, se pasa al Módulo de Obtención de URL's (MOU), donde se genera una lista única de URL's conformada por los primeros diez resultados de los siguientes buscadores: Google, Bing, MSXML Excite e Intelligo. Confeccionada dicha lista, se la retorna al Controlador Web Miner, donde se descartan las repetidas e inicia el proceso de análisis.

Cada una de las URL resultantes, son representadas mediante nodos, con estado de procesamiento "No Explorado". La representación mediante nodos permite mantener una estructura de datos por cada URL, compuesta por los puntajes procedentes de los análisis, el contenido de dicha URL (recurso Web) y el estado de procesamiento, que puede ser "Explorado" o "No Explorado".

La siguiente operación consiste en enviar uno de los nodos marcado como "No Explorado" al módulo Web Crawler, donde se descubren los enlaces directamente relacionados a su URL y se construye un grafo acíclico dirigido que represente estas relaciones. En este grafo, el nodo recibido es la raíz y los enlaces descubiertos se representan como sus nodos hijo. Como resultado de esto, se cambia el estado de procesamiento de la raíz a "Explorado" y se retorna el grafo completo al Controlador Web Miner.

Seguidamente, se envía el grafo generado y la clave de búsqueda, al Módulo de Recuperación de Información (MRI), donde por cada nodo se obtiene el contenido del mismo (HTML o PDF) y se ejecuta el análisis de relevancia, que determina su grado de correlación con respecto a la clave de búsqueda. Para el análisis de relevancia se utilizan dos técnicas: las TL implementadas mediante el Módulo de Análisis Lexicográfico (MAL) y las TS implementadas mediante el Módulo de Análisis Semántico (MAS).

Como resultado de la ejecución de estos módulos, por nodo, se obtienen tres puntajes correspondientes a las técnicas CRank, Okapi y VSM, y un cuarto puntaje que representa a la relevancia obtenida mediante las técnicas semánticas.

Al finalizar estos análisis, se almacena en la base de datos los nodos del grafo analizado, es decir, los puntajes, el contenido de la URL y el estado de procesamiento de cada nodo.

Posteriormente, en el Controlador Web Miner, se comienza con la presentación de los resultados. Para ello, inicialmente, se genera una lista compuesta por los grafos obtenidos hasta el momento.

Esta lista se envía al módulo Web Scrapper, donde se calcula el puntaje final de cada nodo mediante el Modelo de Integración Léxico – Semántico (MILS) y se confecciona el ranking de recursos a ser presentado al usuario.

Este ranking consta de cincuenta posiciones, con el fin de limitar la cantidad de información presentada al usuario. Para generarlo, en principio, se agrupan los recursos de acuerdo su dominio.

Ya con los grupos conformados, se determinan las cincuenta posiciones del ranking, considerando los puntajes arrojados por el MILS de los recursos principales de cada dominio.

Por cada posición del ranking, se genera un archivo JSON que contiene el puntaje de relevancia correspondiente al recurso principal, la posición que ocupa en el ranking y las URL's de los recursos relacionados al mismo. Dichos archivos son almacenados en un repositorio local, permitiendo que sean recuperados por el MRR, para presentarlos al usuario.

El proceso continúa con la expansión de los demás nodos marcados como "No Explorado". Finalizada la primera iteración, se procede a verificar si existen nodos obtenidos a partir del descubrimiento de enlaces relacionados. De ser así, se da inicio a una nueva iteración, lo que implica el descubrimiento de nuevos nodos y la realización del análisis de relevancia por cada uno de ellos, resultando en una reestructuración del ranking a ser presentado al usuario. En caso contrario, se da por finalizado el proceso de análisis de recursos y generación de rankings. Además, cabe destacar que el usuario puede finalizar dicho proceso cuando desee, lo que supone otro punto de corte para la ejecución del sistema.

MÓDULO DE ANÁLISIS SEMÁNTICO

El objetivo de este módulo es implementar métricas de relación y similitud semántica, que contribuyan a determinar la relevancia de los recursos analizados. El esquema general de este módulo se presenta en la Figura 2.

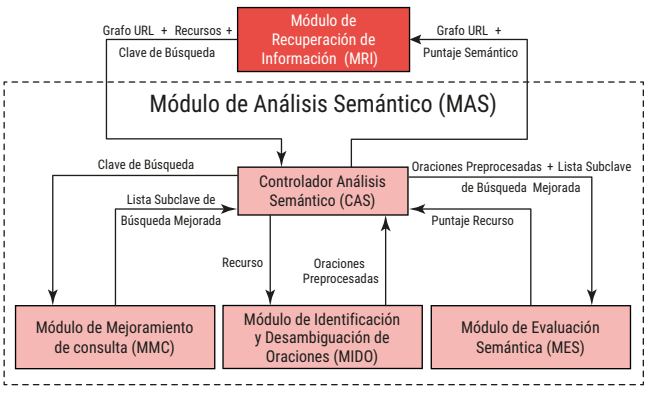


Figura 2. *Esquema del MAS*

Como se puede apreciar, el MAS, está compuesto a su vez por un conjunto de módulos que interactúan entre sí, donde el Controlador Análisis Semántico (CAS) es el encargado de coordinar toda su operatoria.

El proceso de análisis semántico comienza al recibir del MRI el Grafo URL a analizar y la clave de búsqueda, donde cada nodo del grafo está compuesto por su correspondiente recurso.

En primera instancia, el CAS envía la clave de búsqueda al Módulo de Mejoramiento de Consulta (MMC), donde se eliminan errores ortográficos, los stopwords[1] y se identifica el sentido de los términos que la componen. Además, se la segmenta en subclaves, lo que permite que se tenga en cuenta la importancia de las distintas partes de la clave de búsqueda. Como resultado, se retorna la lista de subclave de búsqueda mejorada al CAS.

Luego se envía un recurso correspondiente a un nodo, al Módulo de Identificación y Desambiguación de Oraciones (MIDO), donde se segmenta a su contenido en las oraciones que lo conforman. Por cada oración, se desambigua el sentido de las palabras que la componen. Como resultado, se obtiene una lista de oraciones preprocesadas, que se retorna al CAS.

A continuación, se envía esta lista de oraciones preprocesadas junto a la lista subclave de búsqueda mejorada al Módulo de Evaluación Semántica (MES), donde se aplica la métrica de relación y similitud semántica para determinar el puntaje de relevancia correspondiente al recurso analizado. Finalmente se retorna el puntaje recurso al CAS.

Hecho esto, se comprueba si existen nodos por analizar en el grafo, de ser así, se vuelve a realizar el mismo procedimiento. En caso contrario, se finaliza el análisis semántico, retornando al MRI el Grafo URL junto a los puntajes semánticos correspondientes.

CONCLUSIONES

Para realizar las pruebas y analizar los resultados del modelo, se consideraron dos escenarios. El primero corresponde al ámbito de la educación digital, mediante la utilización de la técnica "Digital Storytelling". El segundo escenario correspondió al área de la seguridad informática, más precisamente en relación a técnicas de ataques por envenenamiento de cookie (Cookie Poisoning).

Durante el desarrollo del trabajo se persiguió el objetivo de analizar las técnicas propuestas, lo que permitió evaluar distintos aspectos relacionados a la recuperación, la generación de rankings, y los criterios de las técnicas semánticas. Las pruebas de recuperación muestran la efectividad de las técnicas semánticas, obteniendo recursos relevantes desde internet, descartando los no relevantes.

Se logró verificar la efectividad en la determinación de la relevancia de documentos específicos sobre el conjunto infinito de recursos existentes en la web, lo cual incide en la cantidad de recursos relevantes presentados al usuario.

Las pruebas de ordenamiento, por otro lado, se centraron en la capacidad de cada técnica de ubicar mejor a los recursos relevantes en los rankings que construyen. En una situación ideal, los recursos más relevantes debieran ubicarse en las primeras posiciones del ranking. Sin embargo, esto no siempre se presentó así en la práctica, debido a la complejidad inherente a la determinación de la relevancia.

Se llevaron a cabo las pruebas sobre el MILS, que permitieron observar los resultados del ordenamiento de recursos a partir de la combinación

1 Stopword: Palabras vacías, o comúnmente utilizadas y que no contribuyen al significado.

de los criterios considerados por las técnicas semánticas y comprobar como la consideración de distintos enfoques, puede contribuir a la correcta estimación de la relevancia.

En definitiva, las pruebas realizadas proveyeron de elementos de análisis considerando distintas perspectivas que permitieron observar y destacar distintos aspectos relacionados al comportamiento del modelo.

BIBLIOGRAFÍA

• Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York : Harlow, England: ACM Press ; Addison-Wesley.

• Blair, D. C. (1990). *Language and representation in information retrieval. Amsterdam;* New York: New York, N.Y., U.S.A: Elsevier Science Publishers; Distributors for the U.S. and Canada, Elsevier Science Pub. Co.

• Cleverdon, C., Mills, J. and Keen, M. (1966) ASLIB Cranfield Research Project: factors determining the performance of indexing systems.

• Eckert, K., Alvarenga, V. M., Barboza, M., Witzke, L. M., and Araldi, L. (2016). *Vigilancia tecnológica e inteligencia competitiva basada en técnicas de minería de la web*, presented at the XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).

• Eckert, K., Favret, F., Barboza, M., Witzke, L. M. and Alvarenga, V. M. (2016). *Modelos de análisis de información para la toma de decisiones estratégicas del sector tealero*, presented at the XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina)

• Favret, F., Montiel, R., Alvarenga, V., Barboza, M., and Witzke L. (2016). *Recuperación de información basada en técnicas de minería Web*, Pag. 7.

• Landauer, T. K., Foltz, P. W. and Laham, D. (1998). *An introduction to latent semantic analysis*, Discourse Process., vol. 25, no. 2–3. Pag. 259–284.

• Tolosa, G. H. and Bordignon, F. R. A. (2008). *Introducción a la Recuperación de Información*. Tolosa y Bordignon.

UNIVERSIDAD
Gastón Dachary