Extracción y clasificación automatizada para la Unidad de Vigilancia Tecnológica e Inteligencia Competitiva de la Patagonia

Claudio Delrieux, Damián Barry, Romina Stickar, Luís Ignacio Aita, Juan Manuel Cortez

Laboratorio de Investigación en Informática
Departamento de Informática, Facultad de Ingeniería,
Universidad Nacional de la Patagonia San Juan Bosco.
Puerto Madryn, Chubut, Argentina. +54 280-4472885 – Int. 116
cad@uns.edu.ar, damian_barry@unpata.edu.ar, romistickar@gmail.com, ignacioaita@gmail.com, juanmanuelcortez@gmail.com

Resumen

El presente trabajo presenta la constitución, desarrollo y actividades realizadas por la Unidad de Vigilancia Tecnológica e Inteligencia Competitiva de la Patagonia. En particular el desarrollo realizado de una herramienta que permite automatizar y homogeneizar las ecuaciones de búsqueda que utilizarán los expertos de la Unidad de Vigilancia.

Palabras clave: big data analytics, extracción, clasificación, Vigilancia Tecnológica, Innobación Abierta.

Contexto

La Unidad de Vigilancia Tecnológica e Inteligencia Competitiva de la Patagonia (VTeIC Patagonia) está conformada por los consorcistas que integran el Parque Tecnológico Puerto Madryn (PTPM) y tiene la finalidad de incorporar a la ciudad de Puerto Madryn una plaza más de la red de Antenas de Vigilancia Tecnológica e Inteligencia Competitiva que estratégicamente están planteadas dentro del Plan Argentina Innovadora 2020, formando parte del Programa Nacional de Vigilancia Tecnológica e Inteligencia Competitiva (VINTEC) del Ministerio de Ciencia, Tecnología e Innovación Productiva de la República Argentina (MinCyT) [1][2][10].

Las actividades presentes se enmarcan en el proyecto de investigación "Clasificación de Información en BigData mediante la utilización de Técnicas de Inteligencia Artificial y Análisis de Redes Sociales" que el Laboratorio de Investigación en Informática (LINVI) acredita en la Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB).

Introducción

Los territorios deben enfrentar nuevos desafíos para el diseño de estrategias de desarrollo dentro de un contexto de mayor complejidad, incertidumbre y velocidad de cambios, adquirir mayores competencias, adaptarse a las exigencias del mercado y avanzar hacia el desarrollo del territorio[10].

Así, la utilización de las potencialidades endógenas se presenta como la estrategia para lograrlo. Teniendo en cuenta que la difusión de las innovaciones y el conocimiento entre las empresas y la organización de los

sistemas productivos en formas más flexibles, son dos pilares fundamentales para el proceso que mejoran las economías internas de las firmas y favorecen el posicionamiento competitivo de las ciudades y territorios[10].

Puerto Madryn presenta espacios institucionales activos en los que participan representantes de diferentes sectores que avanzan en experiencias concretas de articulación y cuyo aprendizaje aportan las bases para la creación de una oficina de VTeIC. A la vez, esta unidad permitirá enriquecer tales espacios ínter-institucionales

En la actualidad existe la necesidad de administrar grandes volúmenes de información no estructurada. Este incremento, debido a la enorme producción de información digital, ya sea desde la perspectiva de información producida en internet como así también la enorme cantidad de información producida por las empresas y organismos que en su gran mayoría, por no ser administradas correctamente, se termina dejando sin uso en algún repositorio de información. Normalmente esta información termina siendo eliminada sin evaluar correctamente su utilidad por parte de una comunidad.

La producción y obtención de información ha pasado a ser uno de los grandes activos de las organizaciones, ya sean públicas, mixtas o privadas. En este sentido el desarrollo y estudio de la generación, administración, explotación, interpretación y clasificación de información se ha convertido en un desafío tecnológico y científico a nivel mundial. Para poder abordarlo, no sólo se requiere del soporte de científicos y tecnólogos en el área de la informática sino además de la integración con investigadores y expertos de distintas áreas vinculadas con las actividades que se desean analizar y comprender, donde a través de la conformación de equipos multidisciplinarios generen verdadero valor a la información circundante.[1][2]

Un Sistema de Vigilancia e Inteligencia consta de 7 (siete) fases preliminares: planificación, identificación de necesidades, búsqueda de información y herramientas, monitoreo y validación, tratamiento y análisis, difusión y protección y evaluación y seguimiento conformando lo que se llama el Ciclo de VTeIC.

- Fase 0: Estructura Organizativa, planificación de actividades de VT, proyección de productos y servicios de VTeIE, recursos físicos y humanos, etc.
- Fase 1: Identificación de necesidades e interpretación del sector - árbol tecnológico, fuentes de información, definición de palabras claves y términos técnicos, recopilación documental, distribución geográfica.
- Fase 2: Búsqueda de información, herramientas y generación de ecuaciones de búsqueda.
- Fase 3: Monitoreo y validación de la información.
- Fase 4: Tratamiento y análisis de información.
- Fase 5: Difusión y protección de la información.
- Fase 6: Evaluación, seguimiento y actualización del proceso de VeIE y presentación de Informe Final

Herramienta de Vigilancia automatizada

Como hemos expresado entendemos que no es posible pensar implementar un tratamiento manual para la gestión de conocimiento realizado por expertos para la implementación de las fases 2, 3 y 4

planteadas en la sección anterior, es por esto que la Unidad de VTeIC de la Patagonia ha diseñado y desarrollado una herramientaintegral que permita realizar la extracción de información tanto científica, de patentes como comercial y de mercado de forma automatizada y homogénea para su tratamiento.

Para realizar este producto se establecieron 3 etapas:

1. Extracción y pre-clasificación de información.
2. Clasificación avanzada de información y construcción de espacios semánticos y taxonómicos mediante técnicas de inteligencia artificial.
3. Explotación de la información clasificada. Automatización de los informes y reportes dela Unidad de Vigilancia Tecnológica e Inteligencia Competitiva.

Líneas de Investigación, Desarrollo e Innovación

Las líneas actuales de investigación se enmarcan dentro de las 3 etapas propuestas en la sección anterior. Para ello se llevan adelante las siguientes líneas de investigación.

- Bases de datos no estructuradas NoSQL. Incluyendo conceptos de Extract, Transform and Load (ETL) para la recuperación y homogeneización de información heterogénea[8][9].
- Técnicas de recuperación de información (information retrieval)[11][12]. Dentro de esta línea se incluyen entre otras temáticas:
    - Lematización.
    - Reconocimiento de entidades nombradas.
    - Extracción de relaciones.
- Técnicas de análisis de redes sociales para la categorización, evaluación y valuación de la red de autores, temáticas y países.
- Técnicas de machine learning y datamining orientado al análisis de textos.
- Técnicas de visualización inteligente de información.

Resultados y Objetivos

Motor de extracción de información

Para la extracción de contenido público de bibliotecas digitales y portales de patentes se desarrollo un motor de extracción de contenido público de internet denominado CrawlingExtractor. El mismo está basado en un trabajo anterior para extraer noticias de orden público georreferenciadas según su contenido denominado “ZCrawler”[4], el mismo se compone de un lenguaje de extracción que le permite a los usuarios mediante una especificación formal comenzar a extraer información siguiendo múltiples criterios y permitiendo la clasificación de esta información según a quien se está evaluando.

Para ello se desarrollaron servicios de ejemplo para las API de Sciencedirect[5] Springer[6], Scopus[7] y para medios digitales se construyeron con RSS estándar.

Adicionalmente se adaptaron todos los extractores ya que las definiciones de API y/o RSS no accedían a información complementaria de los artículos publicados como por ejemplo el perfil de los autores y las citas bibliográficas.

Arquitectura de Extractores

El éxito de la herramienta se basa en la posibilidad de estandarizar en un sólo formato y lenguaje todos los sitios y plataformas de extracción de información, evitando de esta forma la necesidad que

los expertos aprendan muchos lenguajes para elaborar las ecuaciones de búsqueda. Además el esfuerzo de trabajar con formatos heterogéneos de información y estandarizarlos permite luego, independientemente de la fuente unificar la forma de consultar la información.

Para ello es necesario que se programen los distintos formatos de extracción para cada fuente. Si bien el mismo requiere de un esfuerzo por parte de programadores, una vez desarrollado la misma fuente sirve para múltiples consultas.

Para almacenar las extracciones se utiliza una base de datos NoSQL orientada a la construcción de índices invertidos denominada Apache Solr[8], basada en el motor de índices de Apache Lucene[9].

Actualmente la herramienta cuenta con 4 extractores definidos y asociado a cada uno ecuaciones de búsqueda sobre el sector Aluminio en particular en las temáticas relacionadas a los materiales de colada de aluminio y el sector de Pesca y Acuicultura particularmente orientada a la producción de alimento wakame a partir del alga undaria pinnatifida.

Próximos Objetivos

1. Mejorar el esquema de extracción de información en lo referente a lematización, stopwords y reconocimiento de atributos.
2. Extraer información relevante de publicaciones, autores y citas a una base de datos orientada a grafos para realizar análisis de redes sociales (ARS) especialmente en determinar clústers de información relacionada.
3. Incorporar clasificación automática de tópicos de los documentos mediante técnicas de machine learning.

Formación de Recursos Humanos

- Actualmente 2 alumnos de la carrera de grado realizan sus tesinas en el marco del presente proyecto. Cada uno enmarcado en las líneas de investigación comentadas.
- Como parte de las actividades del proyecto de Investigación, actualmente se está desarrollando el Curso de posgrado “Introducción a la ingeniería de Ontologías”

Referencias

1. Edgar Morin. El m ́etodo iii. el conocimiento del conocimiento. Madrid: Cátedra, 2:24, 1988.
2. Edgar Morin and Marcelo Pakman. Introducción al pensamiento complejo. Gedisa Barcelona, 1994.
3. Wayan Vota, Rajendra Singh, Siddhartha Raja, Jude Genilo, Shamsul Islam, Marium Akther, and Rohan Samarajiva. Digital government: building a 21st century platform to better serve the american people. 2012.
4. Barry Damián, Aita Luis Ignacio, and Cortez Juan Manuel. Zcrawler: Extracción, clasificación y publicación de información pública desde su perspectiva geográfica. JAIIO 2014.
5. Elsevier B.V. Sciencedirect.
6. Springer Science+Business Media. Springer.
7. Elsevier B.V. Scopus.

8. Apache Fundation. Apache solr reference guide.
9. Apache Fundation. Apache Lucene Core.
10. Jorge A Sabato. Ensayos en campera. Juárez, 1979.
11. Charu C. Aggarwal • ChengXiang Zhai. Mining Text Data. Springer Science+Business Media. 2012.
12. Katariina Nyberg. Document Classification UsingMachine Learning and Ontologies. Espoo, January 31, 2011.
13. Ian H. Witten, Eibe Frank, Mark A. Hall. Data MiningPractical Machine Learning Tools and Techniques. Third Edition. Elsevier Inc. 2011