Big Data como aporte de información al Proceso de Vigilancia Tecnológica

Rosana Hadad Salomón
UTN – FRT
rosanahadad@gmail.com
(0381)155631065

Elizabeth del Valle Made
UTN – FRT
elimade.5@gmail.com
(0381)155683549

Francisco Alfredo Paz
UTN – FRT
fredypaz2001@gmail.com
(03873) 15415870

Abstract

Big data es una herramienta que empieza a cobrar auge, y a utilizarse en el análisis de la información. Razón por la cual no puede pasar desapercibida por la Vigilancia Tecnológica. Es importante relacionar estos dos conceptos para potenciar la Vigilancia Tecnológica y entregar informes de mayor riqueza para el usuario que la necesita.
En este trabajo se explican los conceptos de Vigilancia Tecnológica y Big Data para una mayor comprensión del tema. Luego se lleva a cabo un caso de estudio para demostrar la utilidad de Big Data como aporte de información al proceso de Vigilancia Tecnológica, desarrollando una aplicación con diferentes módulos y elaborando reportes para visualizar los resultados de búsqueda.

Palabras Clave
Big Data - Vigilancia Tecnológica – Búsquedas especializadas - datos no estructurados

1. Introducción

La Vigilancia Tecnológica (en adelante VT) es un proceso de búsqueda especializada de información que implica la observación y el monitoreo permanente de la información y que requiere de una constante actualización en cuanto al uso de herramientas, técnicas y procedimientos que le permitan a una organización agilizar este proceso. En la práctica, la VT se vuelve tediosa y lenta sin las herramientas adecuadas, los procedimientos que apoyen la tarea y una metodología que permita llevar adelante los procesos con éxito.

Es allí, en el afán de buscar nuevas herramientas, que el equipo autor del presente trabajo, consideró la importancia de la utilización de las técnicas de Big Data como medio y herramienta para la obtención especializada de información.

2. La Web Profunda

La fuente de la cual la VT trae información actualizada y precisa de los últimos avances tecnológicos es la llamada "Web Profunda" de donde se puede extraer datos fiables que le servirán a la organización para estar al día en los distintos cambios que se realizan en el mercado actual permitiéndole tomar decisiones necesarias de acuerdo a los datos obtenidos a través de la investigación que se realizó.

La Web Profunda no sólo incluye archivos estáticos subidos a la Web, sino también aquellos archivos que están en difusión constante como lo son las redes sociales donde los usuarios están constantemente posteando. De estos posteos hay que distinguir que no todos son datos útiles, y de ahí surge la necesidad de filtrar, de algún modo, los datos que son recogidos allí.

3. Vigilancia Tecnológica

La significancia de la VT se define como "una actividad que garantiza la supervivencia de las organizaciones en un mundo donde se hace necesario estar atentos a todo aquello que se presenta en el entorno".

Esto lleva a la organización a adoptarla como un proceso sostenido en el tiempo y de carácter cíclico al que denominamos "Ciclo de VT". Este proceso resulta transversal a la organización cobrando mayor importancia y con distintas responsabilidades en las áreas que así lo requieran manteniendo presente su entorno influyente.

Conviene recordar que la tecnología aporta a la VT productividad, velocidad y eficiencia al trabajo humano. Estos aportes pueden llegar a ser tan importantes que marquen la viabilidad o no de su realización. Lo cual no debe llevar a pensar en modo alguno que las soluciones que hoy ofrecen la tecnología y su capacidad de acceso y trabajo con un mundo de información de gran magnitud sin precedentes puedan sustituir el trabajo humano.

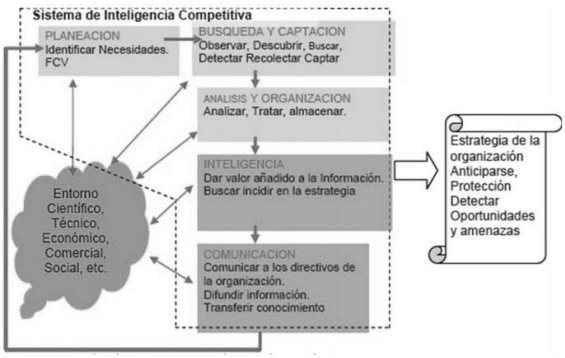

Fig. 1. Ciclo de la Vigilancia Tecnológica

La VT tradicional utiliza datos estructurados para sus procesos. Los datos estructurados tienen la ventaja de ser fácilmente introducidos, almacenados, consultados y analizados. [1]

La actividad previa al inicio de la búsqueda es el armado de lo que llamamos "la frase de búsqueda". Esta frase se arma a partir de identificar las necesidades de quien solicita un proceso de VT (el usuario final) y es armada por un equipo especializado. Esta frase luego es usada en las diferentes herramientas para la búsqueda de información, auditoría de la información y demás.

4. Big Data

La administración de datos no estructurados es uno de los mayores problemas aún no resueltos en la industria de la tecnología de la información. La principal razón radica en que las herramientas y técnicas existentes han sido desarrolladas para el tratamiento de información estructurada pero fallan al momento de procesar información no estructurada. [2]

Los datos no estructurados son todas esas cosas que no pueden ser clasificadas de manera fácil y caber en una caja ordenada: fotos e imágenes gráficas, videos, streaming de datos de instrumentos, páginas web, archivos PDF, presentaciones de PowerPoint, correos electrónicos, entradas de blogs, wikis y documentos de procesadores de textos. El dato semi-estructurado es una combinación entre los dos. Se trata de un tipo de dato estructurado, pero carece de la estructura del modelo de datos. Con el dato semi-estructurado, etiquetas u otros tipos de marcadores se utiliza para identificar ciertos elementos dentro de los datos, pero los datos no tienen una estructura rígida.

"Big data" es un término aplicado a conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable. Los tamaños del "big data" se hallan constantemente en aumento. En 2012 se dimensionaba su tamaño en una docena de terabytes hasta varios petabytes de datos en un único data set. En 2001, en un informe de investigación que se fundamentaba en congresos y presentaciones relacionadas, el analista Doug Laney del META Group (ahora Gartner) definía el crecimiento constante de datos como una oportunidad y un reto para investigar en el volumen, la velocidad y la variedad. [3]

5. Caso de estudio: Uso de la Big Data para la Vigilancia Tecnológica

Las organizaciones tratan de averiguar qué tipo de información deben analizar en vez de enfocarse en el problema. Existen muchos tipos de datos y tratar de dar una clasificación ayudaría a su mejor comprensión.

1.- Web and Social Media: Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, blogs, etc.

2.- Machine-to-Machine (M2M): M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

3.- Big Transaction Data: Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

4.- Biometrics: Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.

5.- Human Generated: Las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc. [4]

En este caso de estudio se tendrá un alcance de sólo dos de las redes sociales, Facebook y Twitter, por ser consideradas las más populares en estos momentos.

Ampliando el concepto de Social Media se tiene que, según Boyd y Ellison (2007), una red social se define como un servicio que permite a los individuos (1) construir un perfil público o semipúblico dentro de

un sistema delimitado, (2) articular una lista de otros usuarios con los que comparten una conexión, y (3) ver y recorrer su lista de las conexiones y de las realizadas por otros dentro del sistema.

En la teoría de los grafos una red es un conjunto de relaciones en la cual las líneas que conectan los diferentes puntos tienen un valor concreto, sea éste numérico o no. Esta posibilidad de cuantificar un vínculo es una de las cualidades que mayor interés puede tener para la sociología. El concepto de red social introducido por Barnes[5] en su estudio sobre los parroquianos de las islas noruegas concuerda aproximadamente con la definición y propiedades que ha enunciado la teoría de los grafos. Aunque esta teoría matemática no es restrictiva a redes finitas, sin embargo en sociología, por cuestiones pragmáticas, normalmente es necesario trabajar con un conjunto identificable de actores (personas, grupos, etc.) y las relaciones que existen entre ellos.[6]

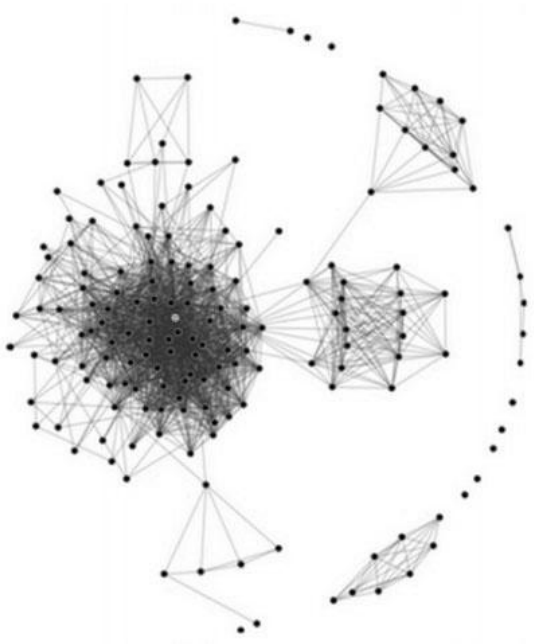

Fig 2. Ejemplo de un diagrama de una red social conocido como grafo social. El nodo con la más alta intermediación centralidad está marcado en amarillo.

Con la solución propuesta en el presente trabajo se pretende desarrollar un producto que ayude al equipo de VT de la Universidad Tecnológica Nacional - Facultad Regional Tucumán a manejar un nuevo grupo de información que hasta ahora no es tratado. Se propone de una aplicación capaz de funcionar en equipos de escritorio o notebooks con acceso a internet que por medio de la frase de búsqueda que es redactada en una de las fases del ciclo de VT, realiza la búsqueda de información en la web profunda, más específicamente, los datos no estructurados. Luego se realiza la extracción de los mismos y se los almacena en un repositorio para luego ser analizados y procesados. Una vez procesados los datos se elaborarán reportes para el uso posterior en las fases siguientes del proceso de Vigilancia Tecnológica.

El objetivo principal de la aplicación desarrollada es demostrar la utilidad de la Big Data en el proceso de VT. La siguiente figura da una vista general de cómo funciona el sistema expuesto:

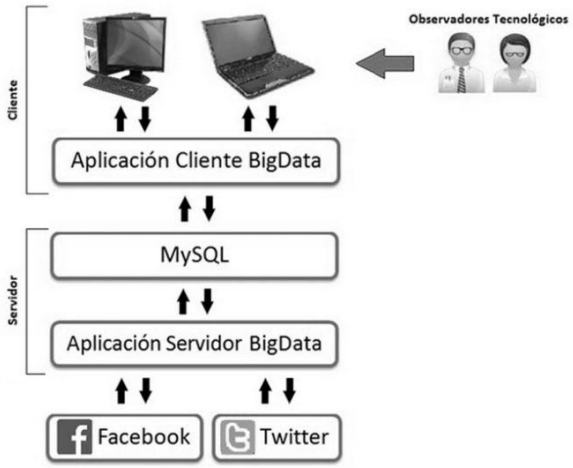

Fig 3. Modelo de funcionamiento de la aplicación

La aplicación se compone de dos partes, el servidor y el cliente. Del lado del servidor se tiene la base de datos y la aplicación que interactúa con las API de Facebook y Twitter extrayendo datos de dichas redes sociales y almacenándolos en la base de datos. Del lado del cliente se tiene una aplicación desarrollada en Java que interactúa con la base de datos y confecciona los reportes para los observadores tecnológicos.

Para esta primera versión de la aplicación, se usa una base de datos MySQL pensando con urgencia en migrar a una base de datos no relacional, y sólo se trabaja con Facebook y Twitter.

5.1 Funcionalidades del sistema

Del lado del servidor se usa código PHP y diferentes scripts que interactúan con la GAPH API de Facebook y la REST API de Twitter para extraer los datos que son almacenados en la base de datos.

Los administradores del servidor de la aplicación deciden de quien extraer datos, porque no es lo mismo una persona que pública sobre sus sentimientos personales a una entidad científica que postea sobre

avances en la tecnología. De este modo se pone un primer filtro a los datos que hay en las redes sociales.

Un segundo filtro es evitar la redundancia de datos en la base de datos, y para este caso, periódicamente, se corre un script que revisa la base de datos en busca de post o twits duplicados de una misma persona.

Del lado del cliente está la aplicación JAVA, con la cual se tiene que estar registrado, en el sistema, como usuario Administrador o usuario Normal para el uso de la aplicación. Una vez adentro se hace uso de la palabra clave para hacer las búsquedas entre los twits y los posts. Se pueden filtrar los resultados de la búsqueda seleccionando el año y el mes de búsqueda y eligiendo en que red social buscar.

5.2 Muestra de resultados

Los resultados de las búsquedas son desplegados en diferentes gráficos usando la herramienta Processing 2

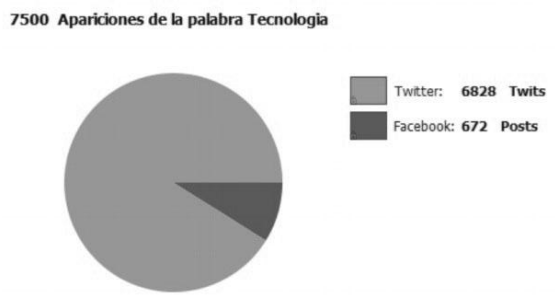

Fig 4. Gráfico de torta con apariciones de la palabra "Tecnología"

Se despliegan gráficos de torta mostrando la cantidad de coincidencias con la palabra clave; gráficos de nubes de palabras, mostrando las palabras con las que más se relaciona la palabra clave como vemos en la Fig.5;

Fig 5. Gráficos de nube de palabras

gráficos de barra, mostrando los paises en los que más aparece la palabra clave como vemos en la Fig.6;

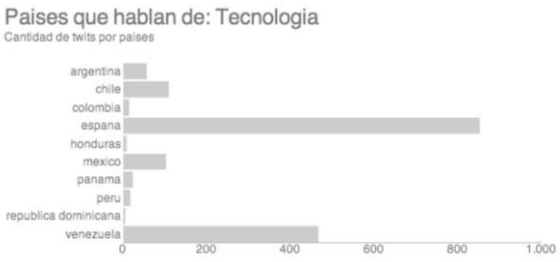

Fig 6. Gráfico de barras con países y cantidad de coincidencias en cada uno de ellos.

gráfico de mapa de calor, donde se visualiza a través de los colores la intensidad con la que se menciona la palabra clave en cada país, siendo el color más oscuro el más intenso y color menos oscuro el menos intenso. Para más detalle se muestran los resultados de los paises de manera individual separados en continentes.

Fig 7. Mapa de calor

Fig 8.Apariciones por cada país

gráficos de linea, donde se muestan las apariciones de la palabra clave através del tiempo. A modo de ejemplo se buscó la palabra clave “Tecnologia” en el mes de Octubre del año 2014;

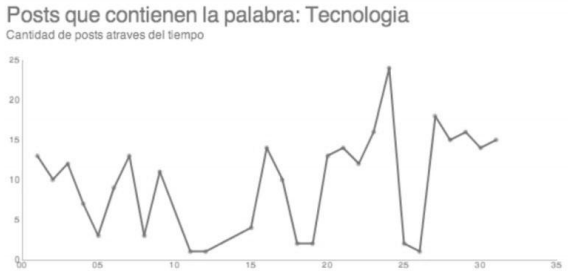

Fig 9. Coincidencias en Facebook a través del mes de Octubre del 2014

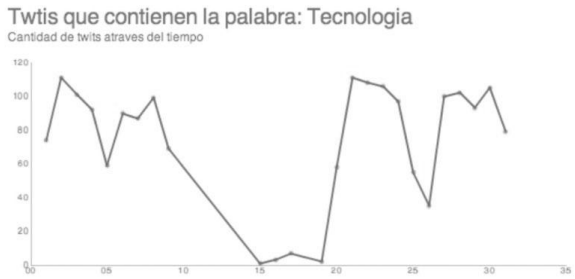

Fig 10. Coincidencias en Twitter a través del mes de Octubre del 2014.

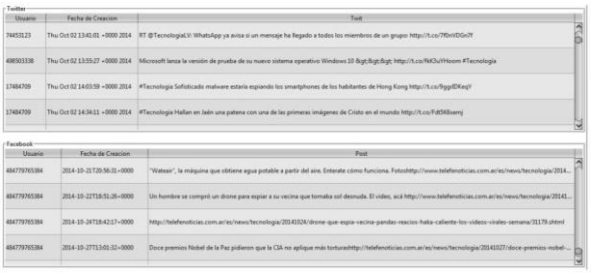

Fig 11. Imagen de tabla de Tweets y Posts.

6. Resultados

Los resultados obtenidos a lo largo de este trabajo fueron alentadores, ya que el prototipo construido cumplió mínimamente los requerimientos propuestos inicialmente. Se pudo agregar un nuevo grupo de información al servicio de la VT, y de este modo, expandiendo el rango de búsqueda. Otra característica importante presente en los resultados fue la velocidad de respuesta relativamente alta del prototipo a las peticiones del usuario y a las buenas cualidades del software. Con el deseo de mejorar los resultados obtenidos, este trabajo servirá como referencia para futuros desarrollos.

7. Conclusiones

Con la idea de buscar nuevas herramientas y expandir las fronteras de búsqueda nos encontramos con un concepto relativamente nuevo y en crecimiento, Big Data.

La mayor dificultad que se encontró a lo largo del desarrollo del prototipo no es el almacenamiento de la información, sino el procesamiento de ella. Además, es mucho más barato almacenar información que procesarla. Por estos motivos es que hay tanta información en la web profunda que no es explotada.

Para el caso de la VT, este grupo de información es importante, porque una vez refinada agrega valor a los reportes finales.

8. Referencias

[1] Pastorino, Maria I; Hadad Salomón, Rosana; Choma, Patricia E & Quiroga Hamoud, Maria C. *Paradigmas de la Vigilancia Tecnológica*. San Miguel de Tucuman: UTN-FRT

[2] J. Fernandez, N. Miranda, R. Guerrero, F. Piccoli. *Datos no Estructurados No Textuales: Desarrollo de Nuevas Tecnologías*.

[3] Area, Eduardo. (14 de Septiembre del 2012) ¿Qué es Big Data? [Mensaje en un blog]. Recuperado de https://eduarea.wordpress.com/2012/09/14/que-es-big-data/

[4] Ricardo Barranco Fragoso (2012). *¿Qué es Big Data?* Recuperado de http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html

[5] Barnes, J.A.*Class and committees in a Norwegian Island Parish, Human Relations, vol. 7*, 1954, pp. 39-58.

[6] Cascales, A.; Real, J. J. & Marcos, B. (2011). *Las redes sociales en internet*. Edutec-e, Revista Electrónica de Tecnología Educativa, 38. Recuperado desde http://edutec.rediris.es/Revelec2/Revelec38/redes_sociales_internet.html.

[7] Mayer-Schönberger, V; Cukier, K.(2013). *Big Data. La Revolución De Los Datos Masivos.* Madrid, España, Turner Publicaciones S.L.