Zcrawler: Extracción, clasificación y publicación de información pública desde su perspectiva geográfica

Lic. Damián Barry[1], APU Luís Ignacio Aita[2], APU Juan Manuel Cortez[2]

[1]Depto. de Informática, Facultad de Ingeniería, UNPSJB. Puerto Madryn, Chubut, Argentina +54 280-4472885 – Int. 117. damian_barry@unpata.edu.ar

[2]Sur Software: Soluciones Informáticas. Patagonia 687, Puerto Madryn, Chubut, Argentina +54 280-4454883. damian_barry@unpata.edu.ar

"Un puente es un Hombre cruzando un puente"

Julio Cortázar

"El Caos es un Orden por Descifrar"

José Saramago

Resumen. El presente trabajo, es el resultado de un proyecto de vinculación y transferencia tecnológica entre la Universidad Nacional de la Patagonia San Juan Bosco y empresas nacionales en la modalidad de Innovación Abierta.

Sobre la necesidad de construir una herramienta para gestionar información pública de internet desde la perspectiva de su origen o contenido geográfico, donde se ha desarrollado una herramienta que aplicando técnicas de búsqueda y recuperación de información (information retrieval) y de clasificación de información y construcción de perfiles, permitan entender las preferencias del público que consume información teniendo en cuenta la ubicación geográfica.

La solución se potencia mediante la utilización de técnicas de análisis de redes sociales como una herramienta para la gestión de información desde la perspectiva del comportamiento social.

Palabras claves: Gestión de la Información, Sistema de Innovación Regional, Sociedad de la Información, Lucene, Solr, CMS.

1. Introducción

El presente trabajo tiene como finalidad mostrar una herramienta concreta que permite mejorar la gestión de grandes volúmenes de información pública provenientes de redes sociales y portales de información.

A partir del creciente uso de Internet y de la popularidad en el uso de las Redes Sociales las cuales, según Edgar Morín[Morin99], generan “ruido” en el ciber-espacio es que se crea la necesidad de poder procesar y transformar la gran cantidad de datos en información útil para las personas o para las organizaciones que desean entender sucesos desde la perspectiva social.

Por otra parte 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 solo 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.

En este sentido Edgar Morin[Morin99] en su libro “El Método” dice que “La idea de cibernética–arte-ciencia del gobierno puede integrarse y transformarse en co-cibernética – arte-ciencia de pilotear conjuntamente, donde la comunicación ya no es útil del mando, sino una forma simbólica compleja de organización.”

“La información generativa y la información circulante pueden transformarse la una en la otra, pero la transformación de una información circulante o de señales de información generativa no es posible más que si se encuentra un aparato capaz de registrarla y tratarla.”

“Así la información solo puede nacer a partir de una interacción entre una organización generativa y una perturbación aleatoria al ruido. Ergo la información no puede desarrollarse más que a partir del ruido. Y desde luego, en el nacimiento de una información, siempre se necesita una actitud organizacional de carácter neguentrópico que se supere a si misma transformando el evento en novedad.”

Claramente los buscadores en Internet (Google, Yahoo, Bing, etc.) han solucionado en parte el problema pero han introducido otro que no es menor: la supuesta información recuperada es sesgada al criterio de clasificación y ordenamiento de cada buscador y no precisamente resuelve la problemática de encontrar lo que se necesita, más aún de generar la información correcta como objetivo a la resolución de problemas concretos.

En el caso particular de las denominadas “Redes Sociales” sólo vemos el resultado de la aplicación parcial de técnicas de Análisis de Redes Sociales, donde estas herramientas tienen más la finalidad de ser un sistema de comunicación entre los usuarios a través de sus preferencias que una herramienta que permita transformar al evento en novedad.

El carácter público de gran parte de la información generada en estas redes sociales y gracias al acceso de la misma, mediante interfaces públicas, permite acceder a la misma con técnicas de Information Retrieval.

2. Contexto

2.1 Cambio de Paradigma: Globalización y Sociedad de la Información.

Estamos participando de un cambio de paradigma, de una transformación de los criterios básicos con los que comprendemos la realidad: Durante muchos siglos se conocieron discursos (religiosos, culturales, políticos y científicos) que la pretendieron unificar; siempre se propuso una explicación que redujera lo que sucede a un solo principio: lo que Derrida llamó “logocentrismo” y, en palabras de Gilles Deleuze, se podría denominar “monológica”.

La lógica virtual, la lógica informática, la presencia de Internet, ya se está instalando como aquel criterio en el que se disuelve toda idea de centro.

Por lo tanto a partir de la obtención de esta información y aplicando técnicas de clasificación según criterios de expertos es que podemos transformar el ruido en verdad.

Este acceso democrático a la información púbica generada por personas y organizaciones, nos permite realizar transformaciones concretas para la obtención de información útil que permita retro-alimentar y detectar patrones para su clasificación y seguir aprendiendo en un proceso continuo negentrópico.

En particular si analizamos la problemática dentro del marco del desarrollo socio-produtivo respecto de la información como conocimiento, refuerza este concepto lo expuesto por Yoguel, Robert, Erbes y Borello en su trabajo “Capacidades cognitivas, tecnológicas y mercados” donde se expresa que: “... En este nuevo esquema el conocimiento presenta una característica distintiva respecto a otros factores de producción, su consumo aumenta la producción y no se agota al utilizarse [Cimoli y Correa, 2005, Yoguel 2000]. El consumo y circulación del conocimiento, que asimismo constituyen una fase importante de su producción, depende positivamente de la complejidad y

articulación de las redes y del grado de competencias endógenas de los agentes involucrados y de las complementariedades que se generan en la interacción entre los agentes que las integran (Ocampo, 2005) ..."

"... A su vez, la tecnología es concebida como un complejo proceso de generación, circulación y apropiación de conocimientos no solo codificados sino también tácitos (Nonaka y Takeuchi 1995), que va mucho mas allá de la incorporación de máquinas al sistema productivo. Este conocimiento, que adopta una doble naturaleza se valoriza cuando se transforma (Rullani, 2000) a partir de procesos de aprendizaje formales e informales (Lundvall, 1992; Johnson y Lundvall, 1994, Lam, 1998, Ducatel, 1998, etc.) que desarrollan los agentes económicos en su práctica e interacción productiva. Comienza a existir un creciente acuerdo acerca de que la innovación no constituye un fenómeno individual de firmas u organizaciones (Freeman et al 1991) y crecientemente se enfatiza en el carácter multi-direccional del proceso de aprendizaje, cada vez mas contextual y transdisciplinario con una fuerte interacción y complementariedad de los componentes tácitos y codificados (Nonaka y Takeuchi, 1995). En ese proceso, comienzan a cobrar creciente importancia los mecanismos de intermediación y traducción que se manifiestan en la emergencia de las llamadas instituciones puente (Casalet, 2000a, 2000b). .."

2.2. El hombre con centro generador de conocimiento

¿Cuáles son las características que podemos avistar del nuevo mundo? Por de pronto el debilitamiento de la idea de verdad. De diversos modos, según la disciplina que se trate, la suposición de que existen verdades inamovibles, cede paso a la admisión de la eficiencia. A su vez la eficiencia se admite siempre dentro de un paradigma, que no es prueba de verdad si no que funciona dentro de lo que está preparado para resolver. Otra de las cosas que ha variado es la idea de deber. La vieja ética que suponía verdades absolutas ha cedido paso a una más abierta, más difusa, que sólo reconoce valores dentro de los campos en los que funciona.

La unidad menor de este sistema y objeto mayor es el hombre, organización a su vez biológica-cultural que integra los sistemas y sub-sistemas sociales, económicos, políticos, culturales etc. y que es también permanentemente interactuado por fenómenos directos, indirectos y emergentes o sumergentes de los mismos. Por ejemplo su salud como su calidad de vida, estarán íntimamente ligados a su sistema biológico, cultural, social, político, económico, ecológico, medio ambiente etc. y las interacciones de cada uno de estos, más sus propios comportamientos como interactuante, receptor, generador y modificador de los mismos.

Por otra parte algunos de estos individuos son integrantes de sub-sistemas de las distintas organizaciones. Su visión del problema y su resolución será diferente porque de acuerdo a su ubicación y su situación las determinaciones que apliquen serán distintas.

Conectividad (Internet), intercambio de información, redes de información de todo tipo, interacción de las partes e inter-operabilidad, inteligencia colectiva, gestión del conocimiento, transformación de valores y saberes tácitos en codificados, simulación social y la construcción de nuevos espacios semánticos y sociales, serán el desafío tecnológico de los sistemas de información del futuro que garantizarán una mejor calidad de vida al hombre asegurando una mejor utilización de los recursos económicos, profesionales, logísticos, culturales y sociales disponibles dentro de cualquier tipo de organización política y garantizando el equilibrio entre todas las partes tanto públicas como privadas, equilibrando sus contradicciones y potenciando sus cualidades. Obligando al uso racional de la información y a reducir la brecha sobre el conocimiento que existe actualmente.

2.3 Construcción de conocimientos a través de información pública.

Como se ha mencionado la gran proliferación de información publica generada en internet, ha provocado un fenómeno de identidad donde pareciera que la sociedad tiende a la uniformidad de criterios y comportamientos. Actualmente las comunidades ligadas a la información en Internet se están re-ordenando, generando un nuevo orden conceptual sobre las identidades y particularidades en sus comportamientos.

Esto lo demuestra la creciente influencia de las redes sociales, especialmente en los jóvenes que han cambiado sus hábitos de encuentro social. El incremento en la necesidad de generar estas particularidades formando grupos de interés con características geográficas, ideológicas y culturales comunes permiten identificar la necesidad de clasificar la información según los segmentos mencionados con la finalidad de segregarla en ordenes de interés.

La segmentación de la información en áreas de interés ligados a los perfiles de cada individuo, permitirán a los distintos usuarios de la red encontrarse mejor con la información buscada.

Para ello es necesario contar con:

- Extracción de información publica de estructura heterogénea. Utilización de técnicas de Crawling e Informatrion Retrieval
- Segmentación de la información mediante etiquetas: clasificación, segmentación, generación automática de atributos.

- Ecualizador de interés: perfil del usuario basado en pertenencia geográfica, identidad cultural y hábitos.
- Motor de búsqueda que jerarquice la información analizando la clasificación de la misma y los hábitos del usuario.

El desarrollo experimental realizado se centra en los dos primeros puntos, extracción y clasificación pública.

3. El proyecto Zonales

3.1 Historia y contexto del proyecto

El proyecto “Zonales” se realizó en el marco de un acuerdo de transferencia tecnológica entre la Facultad de Ingeniería de la Universidad Nacional de la Patagonia San Juan Bosco y la firma Mediabit S.A.

La firma Mediabit es una empresa de servicios tecnológicos aplicados a internet, donde su principal línea de trabajo es el desarrollo de portales web, marketing digital y comunity management.

Resultado de su basta experiencia en el desarrollo de productos para empresas en marketing digital es que surge dentro de la empresa la necesidad de generar un producto que les permitiera insertarse como productores de información digital a nivel nacional.

La primer etapa del proyecto tuvo dos metas para la firma. Por un lado la organización del área tecnológica y transferencia de tecnología respecto de gestión de proyectos tecnológico mediante la modalidad de gestión ágil. Y por otra el desarrollo de una plataforma web que permita la gestión distribuida geográficamente de agencias productoras de noticias, dando un sentido de descentralización a la producción de contenidos. Como resultado de esta primer etapa se concretaron dos productos con muy buen resultado para la firma: La plataforma de gestión de contenidos distribuidos geográficamente y lo que se denominó el “Ecualizador de Interés”, herramienta que permite a los usuarios recuperar información según sus preferencias. Los resultados del ecualizador de interés fueron publicados en Intercom2010 en Puno, Perú [Barry, Páez, Cortez 2010]. Este trabajo demostró la gran potencialidad del esquema de búsqueda y ordenamiento (boosting) de información utilizada en la herramienta Solr-Lucene.

Luego de esta primer etapa la empresa detectó que el modelo de negocios pensado resultada de difícil implementación pues requería de la instalación o contratación de agencias de noticias distribuidas a lo largo y ancho del país. Esta problemática nos presentó un nuevo desafío: cómo

obtener información local (obtención de información distribuida geográficamente) de forma automática y clasificarla para utilizar la potencialidad del ecualizador de interés.

Para esta segunda etapa denominada “Georreferenciación de contenidos” es donde surge la necesidad de desarrollar una herramienta de extracción de contenido público de internet (crawling), que permita recuperar información de redes sociales y portales y su clasificación, fundamentalmente desde su perspectiva geográfica, ya que el concepto “Zonales” así lo requería.

3.2 Construcción de un motor de extracción de información

Para la recuperación de contenido público de internet se desarrollo un motor de extracción (Crawler) denominado dentro del proyecto como Zcrawler, el mismo se compone de un lenguaje de extracción que le permite a los usuarios mediante una especificación comenzar a extraer información según múltiples criterios, permitiendo la clasificación de esta información según a quien se está evaluando.

3.2.3 Gramática del lenguaje de extracción

En la siguiente figura podemos observar la gramática EBNF definida dentro del proyecto:

```
<Zcrawling>     ::=     ["**" <descripcion> "**"]
                        "extraer para la localidad" <localidad>
                        "mediante la fuente" <fuente>
                        ["asignando los tags" <tags>]
                        ["a partir" <criterios>]
                        ["incluye comentarios" ["de los usuarios" <usuarios>]
                        ["y filtrando por" <filtros> ]
                        ["incluye los tags de la fuente"]
                        "."

<descripcion>   ::=     ? Cualquier cadena de caracteres ?
<localidad>     ::=     ? para toda localidad definida en el arbol de localidades de zonales encerrado entre comillas simple ?
<tags>          ::=     <tag> {","<tag>}
<tag>           ::=     ? para todo tag (distinto de la zona) definido en la estructura válida de tags de zonales entre comillas simple ?
<fuente>        ::=     "facebook" | "twitter" | "feed ubicado en" <uri_fuente> [ubicacion]
<uri_fuente>    ::=     ? uri de rss o atom bien formada ?
<criterios>     ::=     <criterio> {"y" <criterio>} ["pero no" <criterio> {"y" <criterio>}]
<criterio>      ::=     "de los usuarios" <usuarios> |
                        "de amigos de los usuarios" <usuarios> |
                        ["si o si"] "de las palabras" <palabras>
<usuarios>      ::=      <usuario> [ubicacion] {"," <usuario> [ubicacion]}
<usuario>       ::=     ? string (palabra) que identifica a un usuario en la fuente (facebook, twitter, linkedin, etc) encerrado entre comillas simple?
<ubicacion>     ::=     "[" latitud "," longitud "]"
<palabras>      ::=     <palabra> {"," <palabra>}
<palabra>       ::=     ? cualquier palabra ?
<filtros>       ::=     <filtro> {"y" <filtro>}
<filtro>        ::=     "al menos <min num shuld match> "de las palabras deben estar" |
                        "con una dispersión entre palabras no mayor a" <numero_entero>" |
                        "lista negra de usuarios" |
                        "lista negra de palabras" |
                        "con al menos" <numero_entero> "actions"
<min num shuld match> ::= ? según la especificación y formato de solr ?
<uri_fuente>    ::=     ? formato uri apuntando al url de definición, encerrada entre comilla simples ?
```

imagen 1: Gramática de extracción

La estructura básica de extracción permite definir una zona geográfica vinculada con la información a extraer, esta decisión se enmarca en la necesidad de georreferenciación de contenidos. Luego se especifica la fuente a extraer, indicando al motor de extracción el componente específico de

extracción a utilizar. Además de la georreferenciación se permiten utilizar otros clasificadores mediante la utilización de tags. Finalmente el motor de extracción permite configurar la frecuencia de extracción.

La herramienta de extracción contempla un conjunto de utilitarios que permiten a quien realiza las configuraciones de extracción realizar análisis previos de la información a a extraer y simular resultados de extracción para evaluar si es lo que se desea.

A continuación podemos observar algunos ejemplos de extracción:

- extraer para la localidad 'puerto madryn' mediante la fuente facebook a partir del usuario 'LU17.com' y 'Madryn TV' y de amigos del usuario 'Cultura Puerto Madryn' incluye comentarios de los usuarios: 'Demián Barry', 'Juan Manuel Cortez'.
- extraer para la localidad 'Argentina' mediante la fuente twitter asignando los tags 'política','actualidad' a partir del usuario '@CFKArgentina' y del usuario '@mauriciomacri' y del usuario '@ricalfonsin' incluye comentarios y filtrando por con al menos 50 actions.
- extraer para la localidad 'Córdoba' mediante la fuente twitter asignando los tags 'Deporte','futbol' a partir de las palabras córdona, deporte, futbol, talleres y filtrando al menos 75% de las palabras deben estar y con al menos 25 actions incluye los tags de la fuente.
- extraer para la localidad 'rosario' mediante la fuente rss ubicada en 'www.lacapital.com.ar/rss/ultimomomento.xml' a partir de todo incluye los tags de la fuente.

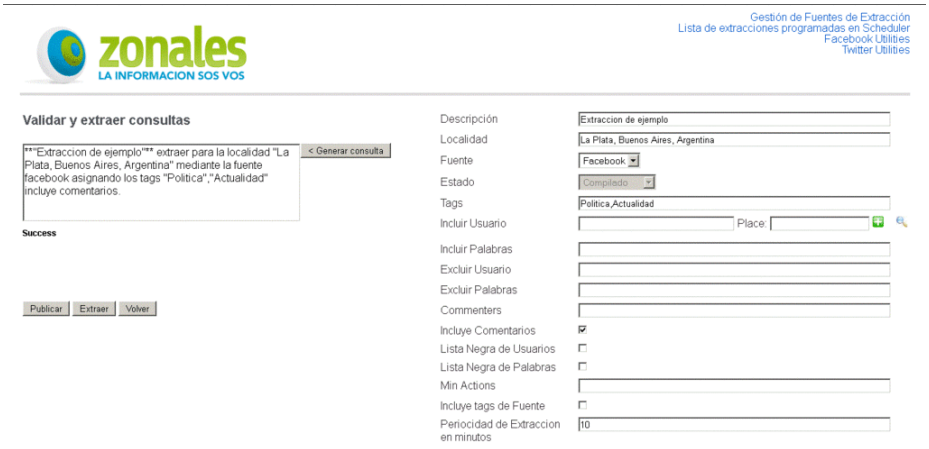

Imagen 2: Carga y validación de extracciones

En la imagen 2 podemos observar el mecanismo de carga de una extracción donde la aplicación permite generar una extracción ya sea mediante su gramática o mediante un asistente.

A continuación en la imagen 3 se pueden observar tanto el resultado de la generación de la gramática de extracción como potenciales resultados de las mismas para que el gestor de extracciones pueda evaluar si es lo deseado.

Imagen 3: Validación de una gramática de extracción

Actualmente se han desarrollado los servicios de extracción para facebook, twitter, RSS estándar y además se han adaptado parser específicos para blogs y diarios en línea que no cuentan con una implementación estándar de RSS. En particular se implementaron conectores para los diarios Jornada y Chubut, ambos de la provincia de Chubut.

Adicionalmente para los grandes medios nacionales se adaptaron extractores ya que sus definiciones de RSS no accedían ni a los comentarios ni a las referencias (links) externas ni a las imágenes.

A partir de poder recuperar imágenes de las noticias y publicaciones extraídas se mejoró la herramienta permitiendo realizar un tratamiento común a todo el material multimedia asociado a una publicación realizando una gestión uniforme tanto para facebook, twitter y las páginas sindicadas.

3.3 Clasificación de la información.

El etiquetado automático (autotagging) se incorpora al proyecto “Zonales” como un modo de colaborar con la carga de contenido. El objetivo fundamental es que tanto en la carga manual de contenidos como en su incorporación automática mediante medios de sindicación (Atom, RSS, etc.) o extracción de redes sociales, el editor cuente con opciones de tags sugeridas por el sistema, teniendo en cuanta aspectos de espacios semánticos que fueron previamente determinados por expertos en distintas categorías de noticias.

La versión original sugería tags en función de las palabras de un contenido. Esto no es óptimo porque las palabras no definen la temática de un contenido. Muchas veces son frases o construcciones complejas las que cumplen este cometido. Para esta tarea se han utilizado características avanzadas de la herramienta de gestión de contenidos utilizada Solr.

3.4. Búsqueda de Información y Apache Solr

EL desafío planteado de extracción de múltiples fuentes de información impone otro desafío que el proyecto debía resolver: la escalabilidad, debido al gran procesamiento de información extraída.

Por ello se trabajó intensamente, donde los resultados fueron publicados en el trabajo denominado “Distributed Search on Large NoSQL Databases” [Barry, Aita, Páez, Tinetti, PDPTA2011].

Donde se puede apreciar que:

“La búsqueda secuencial de cualquier tipo de información presenta varios problemas, siendo el principal la falta de escalabilidad. Una solución a este inconveniente es el uso de estructuras de datos que permitan ser rápidamente consultadas.

El indexado transforma los datos desde su forma original en una estructura que facilita la búsqueda y recuperación de los mismos en forma rápida y precisa, por ejemplo un índice invertido [Hatcher2004], un índice de citas, una matriz o un árbol.

El proceso de indexado generalmente requiere un análisis y procesamiento de los documentos a incluir en el índice: lematización, tokenización, análisis fonético, etc. Estos pasos introducen problemas y desafíos importantes al momento del procesamiento [Hatcher2004], que no son alcance del presente trabajo.

Dentro del estudio del presente trabajo, debido a la facilidad de implementación en un ambiente heterogéneo, se llegó a la conclusión que una solución óptima requeriría el uso del concepto de base de datos de partición horizonal o sharding [Henderson2006].

Apache Solr, que utiliza una base de datos no convencional para almacenar el índice de búsqueda (listas invertidas), provee múltiples capacidades para el escalado horizontal, permitiendo dividir la carga de trabajo entre múltiples instancias lo que permite una fragmentación horizontal de los mismos entre múltiples servidores Apache Solr, a los cuales se les denomina shards. Las búsquedas son luego redirigidas a cada shard, y finalmente una respuesta única es construida en base a los resultados

obtenidos de cada uno. Esta técnica es utilizada especialmente cuando se cuenta con un gran volumen de datos sobre el cual realizar consultas.”

3.4 Resultados del proyecto

Si bien no se ha realizado una carga exhaustiva de fuentes para la Argentina, se han definido algunas localidades del gran Buenos Aires con referencia a publicadores destacados de cada zona geográfica, identificando personalidades de cada localidad y los medios digitales de dichas zonas.

En el término de 1 año de extracción de información se han extraído hasta el momento aproximadamente 12 millones de publicaciones.

Desarrollado por la empresa que recibió la transferencia tecnológica, y complementando el proceso de extracción, la misma generó un portal de publicación de noticias orientado a la zonificación de resultados como se puede apreciar en la imagen 4.

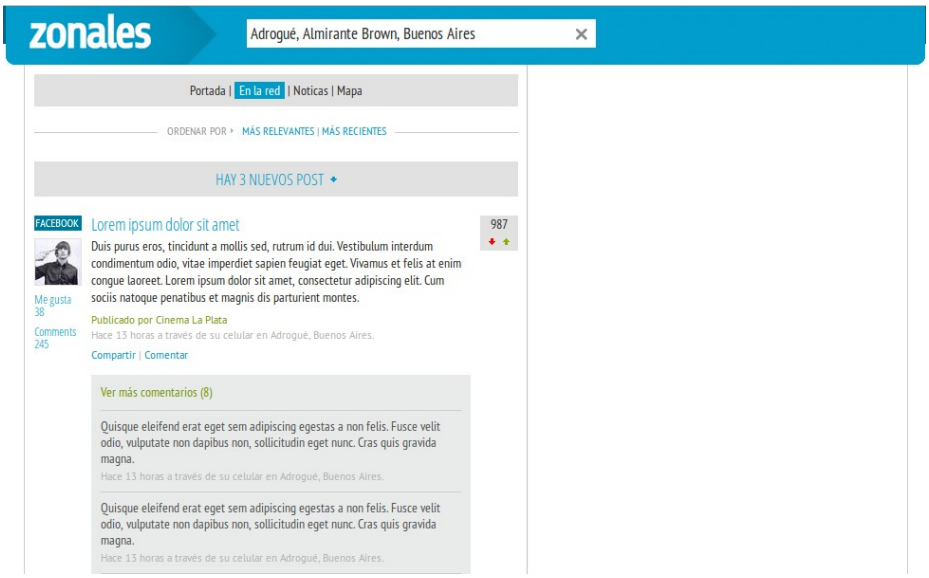

Imagen 4: Portal Zonales

4. Conclusiones y Futuros Trabajos

El proceso de extracción de información y su correcta clasificación es el primer paso analítico para el procesamiento de grandes volúmenes de información.

En este sentido “Zonales” y más particularmente las herramientas de extracción y clasificación cumplen la misión de recolectar información de interés tanto para usuarios que requieren encontrar información según sus preferencias y ubicación geográfica y no por si la noticia tiene más impresiones que otra, premiando en este sentido la centralidad de la información.

Por otra parte la gran cantidad de información recuperada puede ser utilizada para el reconocimiento de patrones sociales que permitan tanto la identificación de “rumores” como la sensibilidad sobre ciertos aspectos y temas de sensibilidad social.

Aplicando técnicas de Análisis de Redes Sociales podemos identificar geográficamente clusters (grupos de personas con intereses comunes) y vinculadores que permiten unir distintos clusters. Todo esto analizado especialmente desde la perspectiva geográfica, potenciando a la herramienta en una herramienta avanzada de vigilancia tecnológica o ser usada como

herramienta de marketing político o clima social. Por supuesto todos estos aspectos requieren de la conformación de equipos multidisciplinarios integrando técnicas avanzadas de análisis de información con el reconocimiento de patrones específicos según ejes temáticos concretos evaluados por expertos en cada área.

5. Bibliografía

1. Ignacio Katz, Retornar al pensamiento lógico. Revista Médicos Número, NEWSLETTER 82 / 24 de Marzo del 2003.
2. Jalfen, Luis J. Globalización y Lógica Virtual. Primera Edición, Ediciones Corregidor, 1998.
3. Morin, Edgar. El Método. Quinta Edición, Ediciones Cátedra, 1999.
4. 5. Borello, J.; Milesi, D.; Novick, M.; Roitter Sonia; Yoguel, G. (2003); Las nuevas tecnologías de información y comunicación en la industria argentina: difusión, uso y percepciones a partir de una encuesta realizada en la región metropolitana de Buenos Aires; en: Nuevas tecnologías de información y comunicación. Los límites en la economía del conocimiento; Boscherini, F.; Novick, M.; Yoguel, G. (comps.); Buenos Aires; Miño y Dávila Editores.
5. Gabriel Yoguel, Verónica Robert, Analía Erbes y José Borello. Capacidades cognitivas, tecnologías y mercados: de las firmas aisladas a las redes de conocimiento. 2005.
6. Apache Solr, http://lucene.apache.org/solr/
7. Damián Barry, Juan Manuel Cortez, Francisco Páez: Construcción de un Ecualizador de Interés Mediante el uso de Lucene-Solr, IEEE Intercon 2010.
8. Erik Hatcher, Otis Gospodnetić. “Lucene in Action”, 2nd. ed, Manning Publications Co. 2004.
9. Cal Henderson: “Building Scalable Web Sites”, O'Reilly Media, 2006
10. Damián Barry, Ignacio Aita, Franciasco Páez: Distributed Search on Large NoSQL Databases, PDPTA2011.