

Comunidades web de inteligencia

F. de la Rosa T. y R. M. Gasca

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
{ffrosat, gasca}@lsi.us.es

Resumen. En este trabajo se presenta el concepto de comunidades web de inteligencia. Estas comunidades utilizan herramientas colaborativas web 2.0 para implementar sistemas de inteligencia competitiva que aprovechen la inteligencia colectiva. Se definen tres comunidades web de inteligencia basadas en distintas herramientas web 2.0: marcado social, enlaces externos y sistemas que usan rastreadores y extractores. Para cada tipo de comunidad se ha realizado un experimento que permite una aproximación a los resultados que pueden esperarse de estos sistemas.

1. Introducción

La evolución de Internet ha sido espectacular, sobre esta infraestructura se han construido distintos tipos de webs. En sus inicios desarrolló la web 1.0 que se caracterizaba por componerse de páginas estáticas o de solo lectura. Para muchos el punto de ruptura hacia la web 2.0 (Tim O'Reilly 2005), comenzó con la fiebre del .com. Esta web se caracteriza por la aparición de las páginas wiki (Ward Cunningham 1994) que concede a los usuarios un papel activo al poder leer y escribir contenidos. Se comienza a hablar de la inteligencia colectiva y de la autoorganización como factor clave para el desarrollo de forma colaborativa de contenidos enciclopédicos como la wikipedia. Las folcsonomías o marcadores sociales [1] han sido utilizadas para clasificar noticias ó enlaces en sistemas de votaciones como Menéame o del.ici.us. Aparecen nuevas formas de comunicación como: blogs, microformatos, RSS y podcats. Y finalmente destacar sistemas como facebook, tuenti ó twitter, que han facilitado el networking en las redes sociales. Todo esto ha provocado una gran transformación social.

Según los expertos el futuro está en la web3.0 o web semántica [2], donde se ve la red como una base de datos gigante donde los usuarios además de poder leer y escribir podrán asociar un significado a los contenidos de los sitios web. Hasta ahora se ha definido lenguajes como OWL o RDF que permiten definir ontologías como FOAF. También se han creado lenguajes para realizar consultas como SPARQL y motores de inferencia como JENA. Pero para que la web semántica sea una realidad aún hacen falta sortear algunas dificultades, como: asociar o anotar contenidos web con un significado, realizar consultas eficientes sobre grandes volúmenes de datos, mapear conceptos parecidos entre distintas descripciones ontológicas, poblar ontologías, etc.

En este contexto a caballo entre la web 2.0 y la web 3.0 están apareciendo nuevos conceptos relacionados con la inteligencia como *intelligence web* o *inteligencia web* [3, 4] una disciplina que integra el marketing con la minería web o el *web mining*. El objetivo de esta disciplina es estudiar como los usuarios utilizan los sitios web, permitiendo a las empresas buscar patrones o tendencias que ayuden a tomar mejores decisiones y a comprender mejor las características de sus productos. Existen dos grandes áreas que investigan sobre los sistemas de inteligencia: la *inteligencia empresarial* o *business inteligente* y la *inteligencia competitiva* o *competitive intelligence* [5]. La primera se caracteriza por trabajar con las fuentes internas de la empresa y con datos cuantitativos y la segunda por trabajar con fuentes externas y datos cuantitativos. Dependiendo de las fuentes de información utilizadas se pueden organizar sistemas de inteligencias en cuatro ejes principales: competitivo, comercial, tecnológico o de entorno. Por tanto dependiendo del tipo de fuente y datos analizados la *inteligencia web* o *intelligence web* se puede clasificar como subárea de cualquiera de estas dos grandes áreas y su eje principal serían los estudios de mercado.

Los sistemas de inteligencia pueden considerar diseños muy variados y no existe un acuerdo definitivo sobre las características que los definen que variarán en función de la disciplina desde la que se estudian. En el marco de este trabajo los sistemas de inteligencia presentan dos características fundamentales: 1) generan una información especialmente diseñada para ayudar a tomar decisiones *a nivel estratégico* y 2) permitir una difusión adecuada de la información. Por ejemplo, un sistema de recomendaciones como el que ofrece Amazon a sus clientes sería adecuado para tomar decisiones a nivel operativo pero no ofrecería inteligencia, y sin embargo una nube de etiquetas podría ser útil para tomar decisiones a nivel estratégico ya que ofrece una *perspectiva holística* de las tendencias de las etiquetas. Aun ofreciendo una información muy distinta ambos sistemas pueden ser considerados como de interés para la *inteligencia web*. Por tanto, *para que podamos hablar de un sistema de inteligencia, que ayude a la toma de decisiones a nivel estratégico, es necesario que la información que produzca ofrezca una visión holística del escenario que se analiza*. Necesariamente en este proceso de resumen o análisis se producen pérdidas de información, será por tanto necesario que no se pierdan matices importantes que ayuden a tomar decisiones.

La gran cantidad de información disponible en Internet ofrece una oportunidad única para el desarrollo de los sistemas de inteligencia. Pero a su vez origina tres grandes retos: 1) cómo guiar al sistema para buscar y recopilar la información que se necesita, 2) cómo extraer dicha información y 3) cómo trabajar con fuentes escritas en distintos idiomas. Para conseguir esto se hace necesario disponer de un corpus de entrenamiento, el problema está en que construir un corpus de entrenamiento no es una tarea fácil y además cada corpus está especializado en una cierta área de interés. Estas necesidades pueden hacer evolucionar a los sistemas de inteligencia hacia lo que llamamos *comunidades web de inteligencias* o *web communities of intelligence*, que serían sistemas que aprovechan las tecnologías que ofrece la web 2.0 para encauzar la inteligencia colectiva hacia la resolución de los retos planteados. Las principales tareas que tendría que afrontar estas comunidades se puede resumir en: clasificar las fuentes de información, clasificar textos, anotar textos con entidades y relaciones, construir *gazetteers* o *lexicones* e implementación de extractores y navegadores [3, 4, 6]. Para que tengan éxito este tipo de comunidades, los sistemas

desarrollados deberán ofrecer herramientas capaces de atender no solo uno, sino los múltiples objetivos de análisis que pudiesen plantear los miembros que participan en la comunidad, ofreciendo resultados en un corto plazo de tiempo.

En el contexto de las comunidades web de inteligencia o web communities of intelligence presentamos tres experimentos.

- La primera sección explica en marco teórico sobre el que se han desarrollado los distintos experimentos.
- La segunda sección está dedicada a las comunidades web de inteligencia basadas en marcadores sociales.
- La tercera, describen las comunidades web de inteligencia basadas en enlaces externos.
- Y en la cuarta sección se presentan las comunidades basadas en rastreadores y extractores.
- Finalizará el trabajo con las conclusiones y los trabajos futuros.

2. Marco teórico de las redes de términos

A lo largo de este trabajo se presentan varias redes léxicas, para la confección de estas redes se han utilizado conjuntamente dos medidas el número de apariciones conjuntas y la co-ocurrencia [7, 8, 9] que pasamos a definir. Sea $T=\{T_1, \dots, T_n\}$ un conjunto de n términos y $D=\{D_1, \dots, D_m\}$ un conjunto con m de documentos o corpus. Podemos definir la matriz de apariciones, A , como:

$$A_{ij} = \begin{cases} 1 & \text{si } T_i \in D_j \\ 0 & \text{si } T_i \notin D_j \end{cases} \quad (1)$$

Donde la expresión $T_i \in D_j$ indica que el documento j contiene el término i . A partir de la matriz A se define las siguientes medidas:

$$0 \leq I_i = \sum_{k=1}^m A_{ik} \leq m \quad (2)$$

$$0 \leq I_{ij} = \sum_{k=1}^m A_{ik} * A_{jk} \leq m \quad (3)$$

$$0 \leq C_{ij} = \frac{I_{ij}^2}{I_i * I_j} \leq 1 \quad (4)$$

Donde I_i es el impacto de una palabra en el corpus, I_{ij} es el número de apariciones conjuntas y C_{ij} es la co-ocurrencia de dos palabras o términos en el corpus. Para el filtrado de la red se eligen dos umbrales de cortes uno para el número de apariciones conjuntas y otro para la co-ocurrencia. De esta forma dando mayor importancia al

número de apariciones conjuntas frente a la co-ocurrencia se observarían los temas principales del corpus y dando más importancia a la co-ocurrencia frente al número de apariciones conjuntas se resaltan las temáticas emergentes.

En la visualización de la red, el color de los nodos indica el impacto del término (marrón indica un impacto mayor), el grosor de las aristas representan el impacto conjunto y el color de la arista la co-ocurrencia (marrón y rojo indica una co-ocurrencia mayor). Para la disposición de los nodos en el plano se utiliza el algoritmo de Fruchterman y Reingold [10].

3. Comunidades basadas en marcadores sociales o folcsonomías

Menéame y Delicious son gestores sociales de noticias (GSN) [11] estos sistemas permiten a los usuarios agregar noticias al sistema y simultáneamente permiten utilizar marcadores sociales o folcsonomías para clasificarlas. En general se habla de noticias pero se puede agregar cualquier tipo de recurso que disponga de un enlace o url, como pueden ser blogs, páginas html, pdf, etc. También existen sistemas como es el caso de Menéame que permite votar las noticias en función del interés que originan.

Los GSN pueden ser utilizados para crear comunidades de inteligencia especializadas en alguna área de interés como la Inteligencia Artificial o la energía solar fotovoltaica. Los usuarios de estas comunidades alimentarían el sistema con documentos relacionados con el área y simultáneamente los clasificarían con marcadores sociales. De esta forma el módulo de análisis del sistema de inteligencia tendría a su disposición un repositorio de enlaces a documentos relevantes clasificados mediante etiquetas o marcadores sociales. Este repositorio podría ser muy valioso para generar información estratégica, por ejemplo, seleccionando una o varias etiquetas se podrían construir redes de etiquetas. Gracias a estas redes los usuarios podrían tomar decisiones en función de las tendencias de las etiquetas y las relaciones existentes entre ellas.

En la Figura 1. se muestra la arquitectura del sistema descrito en esta sección y en el Apéndice A se puede observar la red de etiquetas asociada a la etiqueta crisis. La red presentada en el Apéndice A se construyó a partir de las noticias agregadas a Menéame y que contenían la etiqueta crisis, los datos se obtuvieron en el mes de julio del 2007 [12].

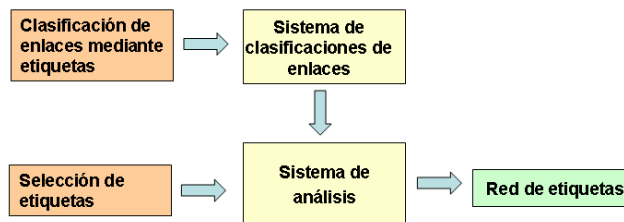


Fig. 1. Arquitectura basada en marcadores sociales.

4. Comunidades basadas en enlaces externos

Las páginas wiki son páginas que permiten construir contenidos de forma colaborativa. En principio estos sistemas son una herramientas excelentes para gestionar conocimiento, como muestra de ello podemos mencionar la Wikipedia, una enciclopedia que se está construyendo de forma colaborativa. Pero los textos no es lo único que puede aportar un usuario a la comunidad, también puede añadir imágenes, tablas, enlaces, etc. La característica más interesante que tienen los sistemas de páginas wiki para construir comunidades web de inteligencia es la gran cantidad de enlaces externos que contiene. Entendiendo los enlaces externos como los enlaces que hacen referencia a recursos que no se encuentran en el sitio web que los acoge.

El módulo de análisis del sistema de inteligencia puede aprovechar los enlaces externos para extraer redes léxicas. Estas redes permiten a los usuarios visualizar los centros de interés asociados a los enlaces externos y tomar decisiones estratégicas, como por ejemplo buscar nuevas tecnologías en las que invertir. Para extraer una red léxica suele ser necesario disponer de un gazetteer. Los gazetteers son las listas de palabras o términos asociadas a conceptos o entidades y que el módulo de análisis utiliza para generar la red. Estos gazetteers también pueden ser construidos utilizando páginas wiki. Para realizar el análisis de tendencias el módulo de análisis debería recibir las páginas wiki desde donde extraer los enlaces externos, bien indicando la página concreta o filtrando las páginas por palabras claves. También deberá recibir el gazetteer que utilizará en el análisis. Para facilitar las tareas al módulo de análisis haría falta disponer de un rastreador o crawler que indexe los contenidos de las referencias externas contenidas en las páginas wiki. Aunque esta sección ha estado centrada en como construir comunidades de inteligencia utilizando páginas wiki lo conceptos aquí presentados se pueden extrapolar fácilmente a los GSN.

En la Figura 2 se muestra el esquema de la arquitectura de las comunidades basadas en enlaces externos y en el Apéndice B se puede observar las red léxica obtenida al analizar los artículos de la revista Iberoamericana de IA [13].

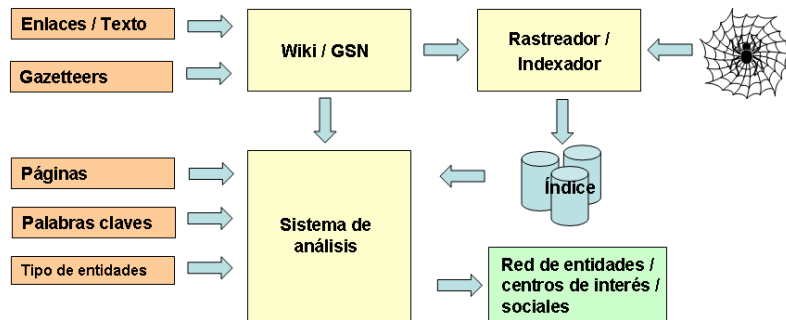


Fig. 2. Arquitectura basada en enlaces externos.

5. Comunidades basadas en rastreadores y extractores

Piggy Bank es una extensión para Firefox que permite a los usuarios extraer información de la web, almacenarla para uso futuro, etiquetarla con palabras claves, y compartirla. Para poder extraer información el usuario debe indicar a la extensión la información que quiere extraer y la extensión se encarga de definir el extractor. Esta tecnología podría ser utilizada para desarrollar comunidades web de inteligencia más complejas y cercanas a lo que se conoce como web semántica. En este caso los usuarios debería indicar al sistema que fuentes necesitan utilizar, estas fuentes podrían ser clasificadas con marcadores sociales para que otros usuarios pudiesen reutilizarlas (economía, política, tecnología, etc). Además para cada fuente habría que indicar como navegar por la web para acceder a la información, por ejemplo utilizando rangos de urls y como extraer la información, por ejemplo utilizando xpath. Con esta información el sistema podría ir navegando por la web, extrayendo información e indexarla, para su posterior utilización por el sistema de análisis. Para realizar el análisis el usuario debería indica al sistema de análisis las fuentes que se desea utilizar, las palabras claves que permitan describir los contenidos que debe buscar en los índices y el gazetteer que tiene que utilizar. En la Figura 3. se muestra el esquema de la arquitectura del sistema y en el Apéndice C se puede observar los resultados obtenidos al analizar la red social de mandatarios obtenidas a partir de los artículos internacionales publicados en varios periódicos [14].

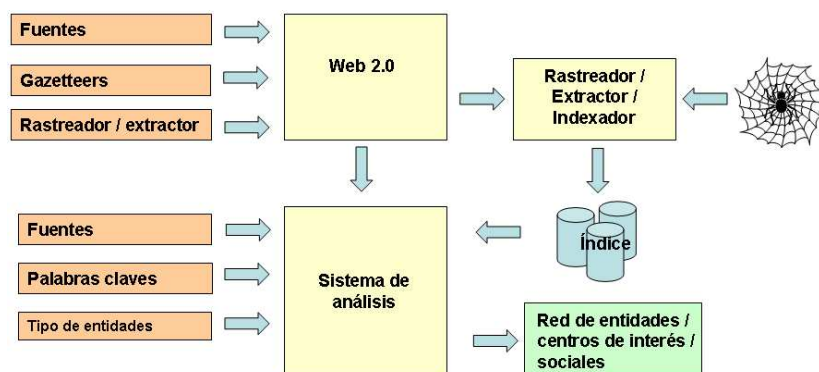


Fig. 3. Arquitecturas basada en navegadores y extractores.

6. Conclusiones y trabajos futuros

Las comunidades web de inteligencia son una herramienta excelente para implementar procesos de inteligencia competitiva dentro de departamentos de I+D+I, además con un coste bajo. Y aunque los sistemas propuestos están basados en sistemas conocidos y fáciles de utilizar implican un cambio en la cultura de la organización y esto puede provocar reacciones adversa a su implantación.

Se han presentado sistemas web de inteligencia basados en el marcado social, en enlaces externos y en rastreadores y extractores, pero también es posible adaptar estos modelos para utilizar otras fuentes de información como blogs, archivos RSS o listas de distribución. También se propone utilizar gazetteers para seleccionar los términos que el sistema de análisis tiene que tener en cuenta. El proceso de construcción de los gazetteer no tiene que verse como una simple lista de palabras introducidas por los usuarios, actualmente existen algoritmos que utilizan técnicas de procesamiento de lenguaje natural para reconocer de entidades, de esta forma se puede utilizar un proceso asistido para generar gazetteers [6]. A su vez tampoco se tiene que considerar que la única información que pueden extraer estos sistemas sean redes léxicas, existen algoritmos que pueden ser utilizados para extraer redes sociales [15, 16, 17].

Aunque parte de los experimentos presentados están implementados con la herramienta Tredar (Technology & REsearch raDAR) que está en fase beta. Esperamos que en el futuro dicha herramienta proporcione la infraestructura necesaria para poder implementar todos los procesos necesarios para implantar una comunidad web de inteligencia. Otra línea de trabajo en estudio implica la realización de análisis que permitan observar como evolucionan las tendencias en el tiempo.

Referencias

1. S. Golder and B.A. Huberman (2005) The Structure of Collaborative Tagging Systems, Information Dynamics Lab, HP Labs.
2. Tim Berners-Lee, James Hendler and Ora Lassila (May 17, 2001). "The Semantic Web". Scientific American Magazine.
3. S. Chakrabarti. Mining the Web: Analysis of Hypertext and Semi Structured Data. Morgan Kaufmann, 2002. ISBN: 1-558-60754-4
4. B. Liu (2007) Web data mining: exploring hyperlinks, contents and usage Data. Springer, Heidelberg. ISBN 3-540-37881-2
5. Escorsa P., Maspons, R. (2001), De la Vigilancia Tecnológica a la Inteligencia Competitiva. Madrid. Prentice Hall. ISBN: 8-420-53057-3
6. C.D. Manning and H. Schütze. (1999) *Foundations of Statistical Natural Language Processing*. MIT. Press. ISBN: 0-262-13360-1
7. Callon, M., Law, J., and Rip, A. (1986). "Mapping the dynamics of science and technology: Sociology of science in the real world". London: Macmillan.
8. Callon, M., Courtial, J.P. and Laville, F. (1991). "Co-Word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry". *Scientometrics*, vol. 22, nº 1, pp. 155-205.
9. Coulter, N., Monarch, I. and Konda, S. (1998). "Software engineering as seen through its research literature: A study in co-word analysis". *Journal of the American Society for Information Science*, 49(13), pp. 1206-1223
10. T.M.J. Fruchterman and E.M. Reingold, (1991). Graph Drawing by Force-Directed Placement. *Software: Practice and Experience*, 21(11).
11. E. Orduña-Malea y J-A. Ontalba-Ruipérez (2009). "Propuesta de indicadores métricos para gestores sociales de noticias: análisis de la prensa digital española en Menéame". *Information Research*, 14(3) paper 406.
12. F. de la Rosa T. (2007). Trabajo Menéame.
<http://www.lsi.us.es/~ffrosat/index.php/Ffrosat/TrabajoMeneame>

13. F. de la Rosa T. (2008). Trabajo sobre la Revista Iberoamericana de IA
<http://www.lsi.us.es/~ffrosat/index.php/Ffrosat/MapaListaIAIBEREs>
14. F. de la Rosa T. (2008) Trabajo noticias de periódicos.
<http://www.lsi.us.es/~ffrosat/index.php/Ffrosat/NoticiasSep2008>
15. Matsuo, J. Mori, M. Hamasaki, H. Takeda, T. Nishimura, K. Hashida and M. Ishizuka. (2006) 'Polyphonet: an advanced social network extraction system, Proceedings of WWW2006.
16. Mika. (2005) 'Flink: Semantic web technology for the extraction and analysis of social networks', Journal of Web Semantics, Vol. 3, No. 2.
17. F. de la Rosa T. and R. M. Gasca (2008) Automatic extraction of social networks by topics of interest. IJCAT , Vol. 33 , Nr. 4, p. 292-299.

Apéndice C: Red de relaciones sociales

