Enriquecimiento Semántico de Contenidos en Redes Sociales basado en Ontologías y Vocabularios de Dominio

Diego M. López1

Edwin F. Caldón 1

Resumo: Muchos de los contenidos que se comparten en los sitios de redes sociales no tiene la calidad deseable, medida en términos de precisión, relevancia, completitud, actualidad, seguridad de la información etc. En el ámbito de las redes sociales en salud, la relevancia de los contenidos juega un papel muy importante. Herramientas inteligentes como buscadores semánticos, permitirían descubrir nuevo conocimiento a partir del contenido extraído de redes sociales, si la información estuviese adecuadamente estructurada y anotada semánticamente. Este trabajo propone un mecanismo basado en ontologías y/o vocabularios de dominio, desarrollado con el fin de enriquecer semánticamente la información compartida en plataformas de Redes sociales. Además, se describe en detalle definición de métricas para la evaluación de la calidad del contenido, especialmente la evaluación de la relevancia de la información.

1 Grupo de Ingeniería Telemática, Universidad del Cauca, Popayán, Colombia
{dmlopez, ecaldon @unicauca.edu.co}

Abstract: Most of the information currently shared in social networks presents several quality flaws: accuracy, relevance, completeness, timeless, security, etc. In the context of healthcare social networks, relevance is foremost important. Intelligent tools such as semantic searchers would allow discovering new knowledge, especially from the content extracted from social networks, when the information has been semantically structured and annotated. This paper proposes a mechanism based on domain ontologies and vocabulary, to semantically enrich the information shared in social networks. Furthermore, a detailed description of metrics for evaluating the quality of the content is presented, especially considering its relevance.

1 Introducción

Actualmente están teniendo una gran acogida los sitios de Internet denominados sitios de redes sociales. Estos sitios permiten establecer lazos entre personas distribuidas geográficamente, ofreciéndoles la posibilidad de interactuar y compartir recursos de acuerdo a un conjunto de actividades comunes. Sitios como Friendster, MySpace, Facebook o Twitter, han explotado este hecho para entretener a sus usuarios capturando información acerca de sus perfiles, intereses, hobbies, gustos y habilidades. También, esta idea se ha venido difundiendo en sitios especializados como LastFm para compartir gustos musicales, LinkedIn una vitrina de profesionales en todos los campos donde las empresas pueden buscar recurso humano calificado, Qdamos una red para buscar potenciales parejas, entre mucha otros ejemplos de este tipo de redes humanas [1].

En el ámbito de la salud, existen redes sociales como Sermo, PatientsLikeMe, Amerian Well, MyCancerPlace, y DailyStrength. Sermo (http://www.sermo.com/) es una de las redes más populares en Estados Unidos. Esta red es una comunidad de médicos que colaboran en casos difíciles, compartiendo puntos de vista clínicos y diferentes observaciones de sus prácticas médicas. Estas discusiones pueden llegar a refutar o aceptar ciertas tendencias en lo relativo a medicación, dispositivos y tratamientos, siempre teniendo en cuenta que el conocimiento generado colectivamente sirva para lograr mejores resultados para los pacientes. Esta red funciona con base en foros de discusión temática con sus respectivos comentarios. Además, ofrece ofertas de trabajo e incentivos por dirigir discusiones.

PatientLikeMe (http://www.patientslikeme.com/) es una comunidad de pacientes en torno a patologías específicas: Parkinson, SIDA, depresión, entre otras. Estos pacientes

comparten en sus perfiles, medicamentos y/o tratamientos, además de la evolución de sus enfermedades, con el fin de recibir apoyo por parte de otros pacientes que han sufrido el mismo problema de salud. American Well (http://www.americanwell.com/) es una red que aprovecha las tecnologías Web 2.0 para que la relación médico-paciente sea más interactiva mediante herramientas de Telemedicina, ya se por video chat, Mensajería Instantánea, email, o teléfono; además de una plataforma social que reúne a todos los pacientes. MyCancerPlace (http://www.mycancerplace.com/) es una comunidad de personas que padecen cáncer y que permite compartir información, dar y recibir apoyo, aprender de las experiencias de otros y crear una página personal donde se incluye texto, fotos y vídeos. DailyStrength (http://dailystrength.org) es una red que incluye más de 500 grupos de apoyo en diferentes enfermedades o estados de salud, comentarios sobre tratamientos, blogs, wikis, etc.

Existen además otros sitios que no son propiamente redes sociales, pero en los cuales se comparte información en salud. Por ejemplo Vitals (http://vitals.com/) y HealthGrades (http://www.healthgrades.com/) enfocadas a comentar sobre los médicos y los servicios que prestan, Pharma Surveyor (http://www.pharmasurveyor.com) dedicada a discutir sobre los medicamentos y su dosificación, Trusera (http://www.trusera.com/) y CarePages (http://www.carepages.com/) enfocadas a tratamientos e información, Google Health (http://health.google.com/) donde se presenta información personal de alergias, enfermedades, medicamentos o tratamientos que sigue con el fin de crear una historia clínica en línea; Estudiabetes (http://www.estudiabetes.com) que es una comunidad de diabéticos en español, entre muchas otras [2].

La importancia que las redes sociales tienen, y su potencialidad para soportar los procesos de cuidado, promoción y prevención de la salud (especialmente en el contexto de salud personalizada) esta recientemente siendo reconocida en el ámbito de la salud electrónica (eHealth) [3], y más concretamente en el de la salud electrónica personalizada (pHealth). En este sentido es el mismo paciente – y no la entidad prestadora de servicios de salud – quién es responsable de la gestión de información sobre su estado de salud. En consecuencia la calidad de la información, especialmente su confiabilidad, relevancia, seguridad juegan un papel preponderante.

Debido especialmente a la complejidad de la información en salud, el uso de tecnologías semánticas se constituye como una herramienta muy importante para soportar a los usuarios y gestores de información en redes sociales en los procesos de filtrado y selección de los contenidos de acuerdo a su relevancia. Desde el punto de vista semántico, el principal problema es que la mayoría de la información que se comparte en plataformas de redes sociales es entregada en documentos en texto plano, lo cual implica que no haya meta-datos o información adicional que ayude a los computadores a entender que los documentos presentes en la red están abordando un tema especifico y que, con esta información, puedan ayudar a los usuarios a tomar decisiones con base en varias fuentes de información que reciben. Para aportar información adicional a los datos existe la "web semántica", tecnología que entre otras cosas permite anotar formalmente los datos (dar más descripción acerca de los datos), mediante lenguajes especializados como XML, RDF u OWL para formar

ontologías o conceptos relacionados de acuerdo al dominio de conocimiento. Esto es, para cada área del conocimiento se agrupan conceptos y se relacionan de tal forma que describan el conocimiento existente de forma general o específica evitando ambigüedades conceptuales [4]. El uso de la Web semántica ayudaría a mejorar la calidad de los contenidos, especialmente su relevancia.

Algunas redes sociales ofrecen la posibilidad a los usuarios de enriquecer semticamente sus contenidos mediante el uso de tags (etiquetas). Las etiquetas se adicionan al texto de forma manual por los usuarios, siendo este proceso sujeto al conocimiento que el usuario tenga del dominio (el etiquetado es una tarea subjetiva). Uno de los sitios que hace uso de etiquetado comunitario (folcsonomia) es del.ici.us, donde cada persona crea unos "bookmarks" dando una descripción y unos tags para páginas que tratan de temas similares, agrupados bajo los mismas etiquetas. Sin embargo este etiquetado (enriquecimiento semántico del contenido) no tiene una estructura formal que pueda ser accedida o compartida desde cualquier otra aplicación que sea capaz de procesarlo.

Actualmente existen algunos esfuerzos para la anotación semántica (enriquecimiento formal) de manera automatizada. El proyecto Annotea (W3C [5]) permite adicionar ciertos metadatos directamente sobre las páginas web usando lenguajes como XML y RDF. SHOE Knowledge Annotator [6] provee mecanismos de anotación semi-automática, facilitando al usuario el marcado directo de documentos, ofreciendo una interfaz de edición similar a un editor de páginas HTML, además ofrece la posibilidad de enlazar el marcado con alguna Ontología de Dominio. Como mecanismo de anotación automática está OpenCalais [7]. Este es un proyecto que mediante servicios web ofrece análisis de textos planos e identifica elementos como personas, instituciones, lugares y hechos, de forma automática. Este servicio web se puede integrar en un sitio de red social o cualquier otro sitio que comparte información, y es un gran avance en el Procesamiento de Lenguaje Natural (NPL) y su anotación semántica. Sin embargo, la aplicación de estas técnicas de anotación semántica a las redes sociales en temáticas especializadas no es inmediata. Para dominios específicos no es suficiente con anotar los cuatro conceptos anteriores (personas, instituciones, lugares y hechos) sino que se necesitarían ontologías de dominio (en el caso de estudio abordado en este trabajo corresponde a una ontología en el dominio de la salud), que permitan la anotación semántica automatizada del contenido de la información.

La anterior revisión de conceptos y tecnologías permite definir la pregunta de investigación abordada en este proyecto de investigación ¿cómo enriquecer y evaluar el contenido compartido en sitios de Redes sociales de forma automatizada e inteligente? Para responder a esta pregunta se ha planteado el siguiente objetivo general del proyecto: Definir y evaluar un mecanismo basado en ontologías y/o vocabularios de dominio para enriquecer semánticamente la información compartida en plataformas de Redes sociales.

2 Métodos

Para cumplir con el objetivo general del proyecto, se ha definido dos objetivos específicos: 1) desarrollar una arquitectura para la anotación semántica automatizada de contenidos en Redes sociales basada en ontologías y vocabularios de dominio, 2) evaluar la calidad de los contenidos en la solución propuesta en un caso de estudio de una red social en salud. Correspondientemente, para cada objetivo se han definido unos resultados específicos:

Para el objetivo 1:

A 1.1Una arquitectura de referencia para las plataformas de redes sociales definida.

A1.2. Un mecanismo para la anotación semántica automatizada de contenidos basado en ontologías y vocabularios, desarrollado.

A1.3. Un componente para la anotación semántica para la arquitectura de referencia en A1.1., diseñado e implementado.

Para el objetivo 2:

A2.1. Plataforma de red social y ontología del dominio de la salud definida y adaptada.

A2.2. Componente de integración implementado en A1.3., integrado en la plataforma de red social.

A2.3. Calidad de los contenidos en la red social del caso de estudio evaluada.

En el contexto del proyecto SALUS - qualidade em sites na área da saúde[10], se viene trabajando en la definición de una ontología para la evaluación de la calidad en sitios Web en salud. A continuación se presentan los resultados de este proyecto enmarcados dentro del proyecto SALUS, enfocado principalmente en la definición de métricas para la evaluación de la calidad del contenido, concretamente en la evaluación de la relevancia de la información.

3 Evaluación de la relevancia de los contenidos en redes sociales

El objetivo principal de cualquier iniciativa que busque el intercambio eficiente de información entre pacientes y profesionales de la salud, es que esta información sea relevante para cualquiera de los dos actores involucrados [9]. Sin embargo, evaluar la relevancia de la información en sitios Web no es una tarea fácil, especialmente porque no hay un acuerdo específico en lo que significa este concepto [10, 11], y por lo tanto es difícil definir métricas para él.

En el contexto del proyecto SALUS, se está definiendo una ontología para evaluar la calidad de sitios Web en salud. Los atributos de calidad definidos en la ontología están basados principalmente en tres de las cuatro dimensiones de calidad definidas en [12]:

Dimensión intrínseca, que evalúa si la información representa correctamente el mundo real y es consistente con ella;

Dimensión contextual, que evalúa la calidad del contenido basado en el contexto del usuario;

Dimensión de accesibilidad, que evalúa aspectos relacionados con la accesibilidad de los contenidos.

Estas dimensiones son, sin embargo, genéricas, requiriendo la definición de unas variables para cada una de las dimensiones. En el trabajo de presentado en [13], se sistematizan estas variables considerando las definiciones más comunes encontradas en la literatura. La tabla 1 resume estas aproximaciones.

Tabla 1. Variables definidas para las diferentes dimensiones de calidad

Variable	Dimensión intrínseca	Dimensión contextual	Dimensión accesibilidad
Accuracy	X		
Objectivity	X		
Consistency	X		
Timeliness	X		
Believability		X	
Completeness		X	
Understandability		X	
Relevancy		X	
Reputation		X	
Verifiability		X	
Amount of Data		X	
Availability			x
Response Time			X
Security			X

Considerando que uno de los objetivos de este trabajo es medir la mejora en la calidad de la información en redes sociales en términos de su relevancia, es necesario definir métricas concretas para esta variable. Según [13], las métricas para evaluar calidad de información pueden ser clasificadas como:

Métricas basadas en el contenido, que usan la información en sí misma para compararla con otra información de referencia;

Métricas basadas en el contexto, que usan meta-información creada sobre el contenido de la información;

Métricas basadas en pesos, que se valoran de cuerdo a pesos establecidos principalmente por usuarios o expertos en el dominio.

En el contexto de este proyecto, la variable relevancia se ubica mejor en las métricas basadas en el contenido y en el contexto, pues estas métricas tienen en cuenta el contenido y su meta-información respectivamente, los cuales se espera que son mejorados utilizando las técnicas de anotación semántica con vocabularios y ontologías de dominio desarrolladas en el proyecto. La definición en términos cuantitativos de estas métricas hace parte del trabajo a desarrollar en el proyecto. Sin embargo, la alternativa más viable parece ser la utilización de modelos de vectores de espacio [14] para mediar la relevancia del contenido. Esta medida dependerá sin embargo de las ontologías y vocabularios utilizados. Las terminología de dominio como SNOMED-CT (Systematized Nomenclature of Medicine, Clinical Terms) [14] permite definir conceptos médicos y sus relaciones. SNOMED-CT sería alternativa más adecuada, dada su amplitud, difusión y aceptación en el dominio de eHealth.

4 Conclusiones

En este proyecto se ha descrito los resultados de un proyecto que pretende enriquecer semánticamente los de contenidos en redes sociales basado en ontologías y vocabularios de Dominio. Se busca demostrar que el uso de ontologías y la anotación automática permiten que los documentos compartidos en la red social tengan una alta calidad, no se presentan ambigüedades y apoyen a los usuarios de la misma en la toma de decisiones, con lo cual el proceso de compartir información es mucho más efectivo. La calidad de los contenidos será evaluada en términos de relevancia del contenido, para lo cual se han definido unas métricas.

5 Agradecimientos

Este trabajo ha sido soportado por el proyecto SALUS: - Qualidade em Sites na área da Saúde; financiado por el programa CYTED, y el proyecto MD-HELF: A Model-Driven Development Framework for Semantically Interoperable Health Information Systems, financiado por la Universidad del Cauca.

6 Referencias

[1] Boyd, D. M., y Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), article 11.

[2] El País Salud. (2009). La web social de los enfermos. Abril de 2009, número 24, pp. 12-13.

[3] Eysenbach GMedicine 2.0: Social Networking, Collaboration, Participation, Apomediation, and Openness. J Med Internet Res 2008;10(3):e22. URL: http://www.jmir.org/2008/3/e22/

[4] Cristóbal Cobo Romaní y Hugo Pardo Kuklinski. (2007). Planeta Web 2.0. Inteligencia colectiva o medios fast food. FLACSO México, La Universitat de Vic Catalunya, ISBN 978-84-934995-8-7. Disponible en http://www.planetaweb2.net/.

[5] Annotea Project . Disponible en http://www.w3.org/2001/Annotea/.

[6] Idoia Murua. (2006). Utilidad de las herramientas de anotación para el desarrollo de la web semántica. Revista / Vigilancia Tecnológica.

[7] Open Calais Document Viewer. Disponible en http://viewer.opencalais.com/.

[8] Proyecto SALUS. Disponible en http://www.cyted.org/cyted_investigacion/detalle_accion.php?un=05f971b5ec196b8c65b75d2ef8267331&lang=es.

[9] Eysenbach G. An Ontology of Quality Initiatives and a Model for Decentralized, Collaborative Quality Management on the (Semantic) World Wide Web J Med Internet Res 2001;3(4):e34 URL: http://www.jmir.org/2001/4/e34/

[10] Richard Wang, Veda Storey, and Christopher Firth. A Framework for Analysis of Data Quality Research. IEEE Transactions on Knowledge and Data Engineering, 7(4):623–640, 1995.

[11] Shirlee-Ann Knight and Janice Burn. Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science Journal, 8:160–172, 2005.

[12] Richard Wang and Diane Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4):5–33, 1996.

[13] Bizer, C. Quality-Driven Information Filtering in the Context of Web-Based Information Systems. 2007. Tesis de Doctorado. Unversidad Libre de Berlín, Alemania.

[14] David Grossman and Ophir Frieder. Information Retrieval:

[15] Algorithms and Heuristics. Springer, Berlin Heidelberg NewYork, 2nd edition, 2004.

[16] National Library of Medicine. SNOMED-CT. Disponible en: http://www.ihtsdo.org/snomed-ct/.