Reuso de Conocimiento en Foros de Discusión:

Hacia una mejora en la recuperación de información

Nadina Martinez Carod, Gabriela Aranda, Alejandra Cechich,
Valeria Zoratto, Pamela Faraci, Carina Noda, Mauro Sagripanti

Grupo de Investigación en Ingeniería de Software del Comahue (GIISCo)
http://giisco.uncoma.edu.ar
Facultad de Informática. Universidad Nacional del Comahue
Buenos Aires 1400, (8300) Neuquén
Contacto: {nadina.martinez, gabriela.aranda, alejandra.cechich}@fi.uncoma.edu.ar

Resumen

Si bien consultar foros de discusión disponibles en la Web para obtener información sobre algún problema particular es una tarea cotidiana, pocas veces es una tarea sencilla ya que el volumen de información es muy grande, por lo cual se vuelve conveniente hacer un análisis exhaustivo de las páginas disponibles para determinar cuáles soluciones presentadas en ellas sirven para el problema que se enfrenta, y si son las más adecuadas.

Cada foro de discusión está mantenido por su comunidad, donde sus miembros actúan realizando preguntas y contestando u opinando sobre otras. Nuestro proyecto hace uso de la información que está disponible en dichos foros para generar una búsqueda más especializada que aquella que puede lograrse con motores de búsqueda multipropósito. Pese a que las consultas en estos buscadores retornan como resultado una lista de objetos de diferente formato (páginas web, artículos, blogs, etc); en muchos casos no se pueden distinguir las respuestas esperadas a los problemas consultados.

Pese a que los motores de búsqueda están superándose constantemente, la política de ordenamiento de las soluciones generalmente no siempre condice con lo que el usuario necesita para solucionar su problema, por lo que suele ser necesario visitar varias páginas y probar varias soluciones a problemas semejantes hasta encontrar aquella que verdaderamente funcione. Surge así el objetivo primordial de nuestro proyecto, el cual es facilitar esta tarea, utilizando como recurso los hilos de foros de discusión técnicos disponibles en la Web, donde la recolección de la información y su priorización se realizan en base a un análisis previo orientado a la calidad.

Palabras Clave

Foros de discusión técnicos, Reuso de conocimiento, Recuperación de información en la Web.

Contexto

Nuestra línea de investigación se encuentra inserta en el proyecto "*Reuso de Conocimiento en Foros de Discusión Técnicos"*, subproyecto del programa "*Desarrollo Orientado a Reuso*", de la

Universidad Nacional del Comahue. Periodo 2013-2016. La línea de investigación está enmarcada dentro de un acuerdo de cooperación con el Grupo Alarcos, Escuela Superior de Informática, Universidad de Castilla-La Mancha, Ciudad Real, España

Introducción

El advenimiento explosivo de herramientas digitales en la red trae aparejado un intercambio de conocimientos insospechados años atrás, donde las herramientas colaborativas posibilitan una nueva forma de comunicación colectiva, conformada por la suma de pensamientos y aportes individuales de los usuarios de las mismas, en un proceso de retroalimentación constante. La Web se puede considerar un lugar donde habitan aplicaciones de diferente naturaleza: puede ser de contenido cerrado como blogs, wikis, foros de discusión; o de contenido abierto como enciclopedias en red, libros digitales colaborativos entre otros.

Las organizaciones tienden a incentivar la colaboración entre sus empleados, mediante el uso de tecnologías, permitiendo compartir conocimiento, dudas, ideas e inquietudes, reutilizando soluciones ya probadas en problemas recurrentes [6]. En suma las estructuras colaborativas y cooperativas han resultado las más eficaces y enriquecedoras dentro de una organización, ya que permite que personas que están localizadas en puntos geográficos diferentes se puedan comunicar electrónicamente.

Nuestro proyecto se enfoca en los foros de discusión, espacios públicos en Internet donde las personas discuten acerca de algún tema en particular. Los foros constituyen una de las herramientas tecnológicas que favorece la interacción a distancia y asincrónica, la cual permite la discusión entre diferentes personas, sobre un tema particular. El proceso comienza a partir de una pregunta sobre un problema específico, generada por un usuario de la comunidad de un foro de discusión; a partir de realizada la pregunta el resto de los usuarios puede contestar estableciéndose una comunicación entre las partes con el fin de obtener la solución al problema planteado. Si bien las soluciones propuestas atienden a la pregunta propia del usuario que abrió la discusión, el conjunto de mensajes queda disponible al público en general, y la o las soluciones propuestas pueden ser reutilizadas al surgir problemas similares.

Dentro de un foro de discusión podemos distinguir los usuarios pertenecientes a la comunidad, que son aquellos usuarios que forman parte activa del mismo, realizando el intercambio de comunicación, y los usuarios externos que son los que utilizan o consumen la información, que además pueden o no ser miembros de la comunidad del foro. Nuestro enfoque se centra en el perfil del usuario que actúa como consumidor de información, esto es el que utiliza la información existente dentro de un foro sin importar su procedencia.

La diversidad de foros de discusión trae como resultado que para una misma temática se puedan encontrar en la Web diversas preguntas y respuestas similares en diferentes foros de discusión; mostrando al usuario que necesita consultar algo, un gran número de hilos hasta obtener la respuesta apropiada a su pregunta inicial. El proyecto parte de esta premisa, y mediante un análisis previo de calidad de las fuentes mencionadas, favorece el reuso de la información

contenida en dichas conversaciones, clasificando los hilos de discusión de acuerdo a un orden de prioridad.

Los modelos de calidad definen una serie de criterios para satisfacer las necesidades de los desarrolladores y usuarios finales, los cuales a su vez, son utilizados para construir mejores productos y asegurar su calidad. De acuerdo a estos lineamientos se han considerado modelos de calidad enfocados en el producto y proceso software, los cuales descomponen la calidad jerárquicamente en una serie de características y subcaracterísticas que pueden usarse como lista de comprobación para los aspectos relacionados con la calidad. Parte de nuestro proyecto abarca la definición de un modelo conceptual para clasificar la información contenida en foros de discusión técnicos. El modelo de la información contenida en dichos foros de discusión se especifica con el objeto de construir un navegador especializado en encontrar soluciones a las preguntas realizadas. La validación del modelo se efectúa mediante encuestas realizadas a usuarios de los foros, enfocadas en la percepción de la adecuación de los hilos de discusión a un problema y la correctitud de las soluciones propuestas.

Existen otras propuestas que reutilizan el conocimiento respecto a los foros de discusión [2] [4] [10] [11], pero en general son dominios más restringidos, la potencialidad de nuestra propuesta radica en que apunta a la recolección de información de diferentes foros, por lo que la pluralidad de sus formatos constituye un desafío adicional.

Líneas de investigación, Desarrollo e Innovación

Este proyecto de investigación es parte del trabajo desarrollado por el grupo de investigación de Ingeniería de Software GIISCo, conformado por docentes y estudiantes de la Facultad de Informática de la Universidad Nacional del Comahue, junto con colaboradores de otras universidades nacionales y extranjeras. Su objetivo es trabajar para ofrecer soporte en investigación y transferencia de tópicos relacionados con la Ingeniería de Software.

Se puede destacar que el proyecto involucra más de un área de investigación permitiendo un trabajo colaborativo interdepartamental. Por un lado docentes del Departamento de Ingeniería de Software al cual pertenece el proyecto, por el otro docentes del Departamento de Programación, con una inclinación a foros de discusión en la temática de lenguajes de programación, sumado con otros investigadores del Departamento de Teoría de la Computación, los cuales se inclinan por el estudio y aplicación de algoritmos de análisis de lenguaje natural, sentiment analysis, aprendizaje automático, etc. El trabajo colaborativo interdepartamental establecido enriquece así a todas las áreas involucradas.

Específicamente, este trabajo está enmarcado en el subproyecto de investigación "Reuso de Conocimiento en Foros Técnicos" dentro del Programa de Investigación "Desarrollo de Software Basado en Reuso". Además de la línea de este subproyecto, el programa aborda otros aspectos del reuso: Reuso Orientado al Dominio y Reuso Orientado a Servicios.

Respecto al subproyecto "Reuso de Conocimiento en Foros de Discusión Técnicos", se está trabajando en dos objetivos principales: Por un lado, siguiendo la dirección de "Captura,

análisis y procesamiento de la información disponible en foros de discusión técnicos" se enfoca en la elaboración de un modelo de calidad para reutilizar el conocimiento disponible en la Web por parte de la comunidad de técnicos informáticos. Por el otro lado el objetivo de "Aplicación del conocimiento adquirido", cuya finalidad es utilizar el los resultados de nuestra investigación para analizar los foros de discusión de la Plataforma de Educación a Distancia de nuestra Universidad (PEDCO) y proponer mejoras en la enseñanza de los alumnos de nuestra Facultad.

Resultados y Objetivos

El objetivo de proyecto de investigación y sus lineamientos han sido presentados en WICC 2014[7]

En el año 2013 se avanzó en la definición de un modelo de calidad para foros de discusión (presentado en ASSE 2013 [8]) en base a modelos de datos y de información en la Web conocidos [3][5] [6] así como en estándares para la calidad de datos software [1]. Sumado a esto se realizó una encuesta a usuarios de foros de discusión técnicos, para validar la selección de atributos y sub-atributos de dicho modelo, cuyos resultados preliminares fueron presentados en CACIC 2013 [9].

Durante 2014 se avanzó en el desarrollo de una herramienta para analizar información recuperada de foros de discusión técnicos cuyos lineamientos fueron publicados en CACIC [12], donde también se presentó un caso de estudio particular con un análisis comparativo entre los resultados obtenidos a partir de una cadena de búsqueda en un buscador de un foro de discusión específico y el orden esperado.

Actualmente se encuentra en desarrollo una metodología de clasificación de hilos de discusión técnicos basados en análisis del texto recuperado, utilizando la herramienta Lucene [13] para indexación y recuperación de documentos, con el fin de determinar relación entre hilos de discusión y documentos Oracle de las clases Java.

A futuro se planea repetir las encuestas realizadas anteriormente, ajustando el tipo de preguntas y extendiendo el conjunto de personas encuestadas, a fin de considerar una mayor diversidad de usuarios que pueden utilizar la herramienta en desarrollo. También se planea continuar con el estudio y uso de técnicas para el análisis semántico de los mensajes de los foros y avanzar en un conjunto de métricas e indicadores de calidad para determinar la correlación entre una pregunta que enuncia un problema y las soluciones al mismo diseminadas en distintos foros.

Formación de recursos humanos

Para garantizar la pluralidad de puntos de vista y cooperación, el proyecto está conformado por un grupo interdisciplinario de docentes y estudiantes de la Facultad de pertenencia y otras universidades nacionales y extranjeras. Conforman el proyecto dos docentes del Departamento de Programación, con dedicación exclusiva, que han concluido en el año 2009 su Doctorado en Informática, dos docentes del Departamento de Ingeniería de Sistemas y de Programación con dedicación simple, que están comenzando a formarse en investigación. Dos estudiantes de Licenciatura en Ciencias de la Computación. que están desarrollando sus tesis de grado. También

colabora una docente del Departamento de Teoría de la Computación de la misma Facultad, que está desarrollando su tesis de Doctorado sobre técnicas de análisis de lenguaje natural y provee asesoramiento sobre algoritmos de aprendizaje automático. Se cuenta con la asesoría externa de una docente e investigadora de la Universidad de Castilla La Mancha, con experiencia en Gestión de Conocimiento, lo que permite asociar desarrollos y producciones entre ambas universidades.

Referencias

[1] ISO/IEC 25012:2008, Software product Quality Requirements and Evaluation (SQuaRE): Data quality model. 2008.

[2] W. Chen, R. Persen (2009), "A Recommender System for Collaborative Knowledge".

[3] C. Calero, A. Caro, M. Piattini (2008), "An Applicable Data Quality Model for Web Portal Data Consumers", World Wide Web, vol. 11, no. 4, pp. 465-484.

[4] D. Helic, N. Scerbakov (2003), "Reusing Discussion Forums as Learning Resources in WBT Systems".

[5] R. Wang, D. M. Strong (1996), "Beyond accuracy: What data quality means to data consumers", Journal of Management Information Systems, vol. 12, no. 4, pp. 5-33.

[6] Smith y Duffy (2001), Re-using knowledge: why, what and where. En Proceedings de 2001 International Conference on Engineering Design, Glasgow.

[7] G. Aranda, N. Martínez Carod, A. Cechich, P. Faraci, C. Noda, M. Sagripanti. *Avances en reuso de conocimiento en foros de discusión técnicos*. WICC 2014, XVI Workshop de Investigadores en Ciencias de la Computación, Ushuaia, Tierra del Fuego, 2014

[8] G. Aranda, N. Martínez Carod, P. Faraci, A. Cechich. *Hacia un framework de evaluación de calidad de información en foros de discusión técnicos*. ASSE 2013, 14th Argentine Symposium on Software Engineering, (JAIIO 2013, 42° Jornadas Argentinas de Informática), Córdoba, 2013.

[9] N. Martínez Carod, G. Aranda, M. Sagripanti, P. Faraci, A. Cechich. *Análisis de la información presente en foros de discusión técnicos*. WIS 2013, X Workshop en Ingeniería del Software. (CACIC 2013, XIX Congreso Argentino de Computación), Mar del Plata, pp. 847-856, 2013.

[10] A. Tigelaar, R. Op Den Akker and D. Hiemstra, *Automatic summarisation of discussion fora*, Natural Language Engineering, ISSN 1469-8110, Vol 16, Issue 02, pp. 161-192, 2010.

[11] H. Kuna, M. Rey, J. Cortes, E. Martini, L. Solonezen, R. Sueldo, *Generación de un Algoritmo de Ranking para Documentos Científicos del Área de las Ciencias de la Computación,* WIS 2013, X Workshop en Ingeniería del Software. (CACIC 2013, XIX Congreso Argentino de Computación), Mar del Plata, pp. 787-796, 2013.

[12] G. Aranda, N. Martínez Carod, S.Roger, P. Faraci, A. Cechich, V. Zoratto. *Una herramienta para el análisis de hilos de discusión técnicos. (*CACIC 2014 XX Congreso Argentino de Computación), Buenos Aires, pp.803-812,2014.

[13] M. McCandless, E. Hatcher, O. Gospodnetic. *Lucene in Action*, Second Edition, Manning Publications Co., ISBN 9781933988177, USA, 2010