Análisis de Foros de Discusión para 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)
htttp://giisco.uncoma.edu.ar
Facultad de Informática. Universidad Nacional del Comahue
Contacto: {nadina.martinez, gabriela.aranda, alejandra.cechich}@fi.uncoma.edu.ar

Resumen

La comunidad informática en general suele utilizar herramientas disponibles en la Web de soporte grupal, tanto para solucionar problemas como para el aprendizaje de alguna tarea particular. Es por ello que este tipo de herramientas de soporte han tenido un gran auge en las últimas décadas, dentro de las cuales los foros de discusión se han consolidado por ser la más utilizada.

El objetivo de los foros de discusión es el de compartir experiencias y soluciones a problemas de tópicos diversos. También funcionan como una importante fuente de información a la hora de realizar consultas y búsquedas de problemas particulares. De esta manera, cuando un usuario tiene una dificultad, realiza una pregunta sobre su problema, que suele ser respondida por diferentes usuarios, quienes proponen diversas soluciones, todo ello realizado en forma asíncrona. Bajo este prisma, los foros de discusión técnicos actúan como plataformas colaborativas para compartir soluciones que luego pueden ser reutilizadas en situaciones similares por otras personas.

En los foros de discusión los usuarios debaten un tema en común libre e informal, formando así una comunidad respecto a un interés común. Sin embargo no es necesario ser miembro de dicha comunidad para acceder a la información conformada por las preguntas y respuestas de los ítems de discusión, lo que permite libertad de acceso a la información a cualquier usuario externo.

La consulta en foros de discusión disponibles en la Web es una tarea sencilla, pero clasificar la información obtenida no lo es, ya que en la mayoría de los casos la cantidad de información a procesar es muy grande, y muchas veces está desordenada y repetida.

Surge así el objetivo de nuestro proyecto, el cual es facilitar la tarea de un motor de búsqueda para un problema particular de un usuario desarrollando técnicas y herramientas para mejorar el reuso basado en foros de discusión.

Contexto

Nuestra línea de investigación es un proyecto que se encuentra inmerso dentro del programa “Desarrollo Orientado a Reuso”, de la Universidad Nacional del Comahue, Periodo 2013-2016. Dicho programa cuenta con tres líneas de investigación. En particular, esta línea

mantiene un acuerdo de cooperación con el Grupo Alarcos de la Escuela Superior de Informática, Universidad de Castilla-La Mancha, Ciudad Real, España.

Introducción

La Web actual se ha convertido en una plataforma que posibilita el encuentro de ideas y favorece la creación de debates científicos en foros de discusión, chats, blogs, etc., que son en cierta forma, una nueva y exhaustiva forma de revisión realizada por toda la comunidad participante. Se puede destacar, entre las herramientas mencionadas, los foros de discusión, utilizados tanto para compartir conocimiento creado por una comunidad de aprendizaje como para identificar y opinar sobre un problema particular. La naturaleza de los foros de discusión maximiza el aprovechamiento de información cuando un tema se repite o un problema surge nuevamente.

Cuando a una persona le surge un problema, ésta trata de verificar si ese mismo inconveniente o uno similar ha sucedido previamente. Para ello busca consultas realizadas previamente sobre el problema en cuestión. En un foro de discusión, cada problema o consulta, se presenta como una pregunta principal. Luego, las respuestas de diferentes participantes a esa pregunta van formando un hilo (thread) de discusión que pueden ser reutilizadas, ya sea por las mismas o por otras personas, en situaciones similares.

Las preguntas generadas por los usuarios del foro de discusión, así como las respuestas que se generan a partir de esa pregunta se suceden hasta obtener alguna solución al problema planteado. Este intercambio de mensajes queda disponible no sólo para la comunidad del foro sino para el público en general, y puede ser reutilizado por ellos en problemas similares.

El proyecto se enfoca en hilos de foros existentes, haciendo uso de la información disponible en la Web. A partir de un análisis de calidad de la información existente en los foros, se clasifican los hilos de discusión de acuerdo a un orden de prioridad, definido mediante métricas, favoreciendo el reuso de dicha información. Esto se realiza definiendo cuando a una persona le surge un problema, ésta trata de verificar si ese mismo inconveniente o uno similar ha sucedido previamente. Para ello busca consultas realizadas previamente sobre el problema en cuestión. En un foro de discusión, cada problema o consulta, se presenta como una pregunta principal. Luego, las respuestas de diferentes participantes a esa pregunta van formando un hilo (thread) de discusión que pueden ser reutilizadas, ya sea por las mismas o por otras personas, en situaciones similares.

Si bien existen recomendadores que analizan automáticamente los hilos de un foro de discusión, como el recomendador de Chen and Persen [5], al igual que en Helic y Scerbakov [6], ambos se enfocan en un dominio de aprendizaje colaborativo. A diferencia de estas propuestas, nuestro recomendador apunta a un contexto más amplio, involucrando usuarios con distintos niveles de conocimientos del tema en cuestión. Tanto en estos trabajos como en [8] el foro utilizado es único, en cambio nuestra propuesta se enfoca en el análisis de un conjunto de foros, por lo cual se hace más

complejo con la utilización de diversos formatos de foros de discusión.

Líneas de Investigación, Desarrollo e Innovación

Este proyecto de investigación está desarrollado por el grupo de investigación de Ingeniería de Software (GIISCo), creado en el año 2002. GIISCo está compuesto por docentes y estudiantes de la facultad de Informática de la Universidad Nacional del Comahue con colaboración de otras universidades externas, nacionales y extranjeras, y su misión es el desarrollo conjunto y cooperativo de la investigación en el ámbito de la UNCo. Además se cuenta con una asesora perteneciente a la Escuela Superior de Informática, Universidad de Castilla-La Mancha, Ciudad Real, España.

El trabajo está enmarcado en el subproyecto de investigación "Reuso de Conocimiento en Foros Técnicos", que a su vez está incluido en el Programa de Investigación "Desarrollo de Software Basado en Reuso". Dicho programa está basado en el reuso como una estrategia de desarrollo para dar respuesta a las demandas de un menor costo tanto en la producción y mantenimiento como en la calidad del software y consta de tres líneas que convergen en el tratamiento del desarrollo de software basado en reuso. Las líneas incluidas son: Reuso Orientado al Dominio y Reuso Orientado a Servicios.

El proyecto es un trabajo colaborativo interdepartamental, en el cual están incluidos docentes del departamento de Ingeniería de Software, del Departamento de Programación y de Teoría de la Computación. Si bien el proyecto pertenece al departamento de Ingeniería de Software, éste analiza foros de discusión con la temática específica de lenguajes de programación, incorporando este área. Sumado a esto el proyecto incluye la utilización tanto de lenguaje natural, de aprendizaje, de sentiment analysis y aprendizaje colaborativo, por lo cual se complementa con asesoría de docentes del Departamento de Teoría de la Computación.

En particular la línea de nuestro proyecto es "Reuso de Conocimiento en Foros de Discusión Técnicos" la cual por un lado determina un modelo de calidad de utilización del conocimiento existente en los foros de discusión técnicos, y por el otro define estrategias para obtener y mantener, manejar y mejorar el conocimiento existente previamente en los foros de discusión".

Resultados y Objetivos

La problemática del proyecto de investigación, los objetivos y sus características se presentan en [3] e incluye la definición de un modelo de calidad y de gestión de conocimiento a partir de información contenida en foros de discusión técnicos. La mejora al modelo se introduce en [1] [2], lo que realiza basado a modelos de datos, de información, y de calidad de datos software [4] [7] [14].

El proceso al cual apunta nuestra propuesta se describe en [1], donde se compara la captura y el análisis de hilos de discusión, entre el buscador especializado y un navegador multipropósito estándar.

Los resultados de la encuesta a usuarios de foros técnicos, sobre su percepción en la correctitud de las soluciones sugeridas en dichos foros, se ha publicado en [9].

En [2] se presenta la herramienta para el procesamiento automático de hilos así como un caso de estudio sobre un problema particular, donde se compararon resultados obtenidos a partir de una cadena de búsqueda en un buscador de un foro de discusión específico.

En [15] se analizan distintas estrategias para clasificar hilos de discusión. Para ello se han utilizado hilos de discusión reales, recuperados del foro de discusión Stack Overflow, referidos a problemas sobre el uso del lenguaje de programación Java y, como conjunto de documentos de referencia, un repositorio de especificación de las clases Java de Oracle. El texto recuperado ha sido indexado, para determinar la relación entre los hilos y los documentos Oracle de las clases Java, con una herramienta API de recupero de información mediante indexación y búsquedas, de software abierto llamada Lucene [10]. Se está avanzando en las técnicas de análisis semántico (incluyendo stopwords, stemming, sinónimos), sobre los hilos de discusión, complementando con el diseño de un conjunto de métricas de calidad que ayuden a definir un orden en el grupo de las soluciones provistas en distintos foros para un problema particular.

Actualmente se está trabajando en generalizar los resultados publicados en [15], replicando el experimento con mayor cantidad de hilos y aplicando técnicas diferentes durante la fase de recuperación de información, así como la realización de nuevos casos de estudio. También se están estudiando alternativas para extender el trabajo realizado, utilizando un conjunto de documentos de referencia considerando su grado relevancia.

Formación de recursos humanos

Forman parte del proyecto los siguientes miembros:

- Dos docentes investigadores del Departamento de Programación, con dedicación exclusiva, ambos con doctorado en Informática.
- Un docente investigador del Departamento de Programación, con una beca doctoral otorgada por el CONICET en diciembre de 2015.
- Dos docentes investigadores con dedicación simple, de los Departamentos de Ingeniería de Sistemas y de Programación.
- Dos estudiantes de Licenciatura en Ciencias de la Computación que están desarrollando sus tesis de grado dentro del mismo.
- 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. Dentro del proyecto, asesorando en temas de aprendizaje automático y lenguaje natural.
- Una docente investigadora externa, perteneciente al Grupo Alarcos de la Universidad de Castilla La Mancha, dicha docente

tiene su doctorado y una amplia trayectoria en gestión de conocimiento.

La conformación del equipo con docentes de distintos departamentos, sumado a la asesoría externa mencionada, permite el trabajo cooperativo de un grupo interdisciplinario. Además, la incorporación de estudiantes de la Facultad amplia los posibles tipos de desarrollo relacionados a la temática del proyecto.

Referencias

[1] 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), Córdoba, 2013.

[2] 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), Buenos Aires, pp.803-812,2014.

[3] 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), Ushuaia, Tierra del Fuego, 2014

[4] 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.

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

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

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

[8] 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,* (CACIC 2013, XIX), Mar del Plata, pp. 787-796, 2013.

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

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

[11] M. Nicoletti, S. Schiaffino, and D. Godoy, *Mining interests for user profiling in electronic conversations,* Expert Syst. Appl., vol. 40, pp. 638-645, Feb. 2013.

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

[13] 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.

[14] 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.

[15] V. Zoratto, G. Aranda, S. Roger, A. Cechich, *Análisis de estrategias para clasificar contenidos en foros de discusión: Un caso de estudio* ASSE 2015, 16º Simposio Argentino de Ingeniería de Software- pp. 176-190.