Estrategias para la clasificación de contenido y usuarios de Foros de Discusión Técnicos

Gabriela Aranda, Nadina Martínez Carod, Valeria Zoratto, Alejandra Cechich,
Facundo Otermin Sánchez, 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: {gabriela.aranda, nadina.martinez, valeria.zoratto, alejandra.cechich}@fi.uncoma.edu.ar

RESUMEN

Los foros de discusión son utilizados por muchos técnicos informáticos para plantear dudas y pedir sugerencias para resolver algún problema particular. Para ello, formulan una pregunta a partir de la cual se abre un hilo de discusión (thread), en el que suelen participar varios usuarios que analizan el escenario y proponen una o más soluciones al problema en cuestión. De esta manera, los foros se han convertido en plataformas colaborativas donde el conocimiento se explicita a la vez que se comparte.

Dado que existen muchos foros sobre las mismas temáticas (lenguajes de programación, aplicaciones específicas, etc.), es posible encontrar en la Web muchos hilos de discusión en diferentes foros que están relacionados al mismo problema. Cuando un técnico informático tiene un problema específico, suele utilizar un motor de búsqueda multi-propósito que le devuelve una lista extensa de páginas de varios tipos (blogs, foros, artículos, etc.), luego el técnico necesitar navegar por varias páginas hasta descubrir cuál es la que describe un problema más parecido al que tiene, y encontrar (si existe) una solución que pueda satisfacerle.

Para facilitar esta tarea periódica de los técnicos informáticos, nuestro proyecto tiene como objetivo la implementación de una herramienta que recupere la información disponible en hilos de discusión de foros técnicos de manera automática, y que a partir de un análisis basado en un modelo de calidad pertinente, permita clasificar dicha información y entregar a los usuarios un ranking de posibles soluciones para problemas recurrentes.

Palabras Clave
Foros de discusión, Reuso de conocimiento, Modelos de Calidad.

CONTEXTO

Nuestra línea de investigación se denomina “Reuso de Conocimientos en Foros de Discusión II” y forma parte del programa de investigación “Desarrollo de Software Basado en Reuso – Parte II”, de la Universidad Nacional del Comahue, con período de vigencia 2017-2020. El programa mencionado extiende el programa “Desarrollo de Software Basado en Reuso” realizado durante el período 2013-2016.

1. INTRODUCCION

La disciplina de Recuperación de Información (Information Retrieval) surge en la década de 1950 [12], ante la necesidad de procesar y reutilizar la información almacenada en grandes volúmenes. A partir de ese momento, este campo de investigación ha madurado y han surgido importantes aportes. Por un lado, varios proyectos se han enfocado en utilizar la información recuperada de documentos específicos, mientras que otros han desarrollado técnicas para generación automática de tesauros (lista de sinónimos, en conjunto con lista de antónimos, etc.) para su uso en distintos tipos de consultas. En general, la recuperación de información se realiza a partir de la consulta de un usuario. Luego, las posibles respuestas se organizan de acuerdo a un ranking que evalúa el grado de relevancia de cada respuesta con dicha consulta.

Si bien el conocimiento en la Web se encuentra diseminado en distintos tipos de sitios y documentos, nuestro proyecto pone el foco en los foros de discusión, que se caracterizan por ser herramientas colaborativas con grandes volúmenes de información, accesibles a la comunidad en general como fuente de consulta. En dichos foros se intercambia conocimiento entre los miembros de una comunidad de usuarios que comparte intereses y características similares.

En general, la mayoría de los métodos automáticos de IR se basan en analizar la ocurrencia de palabras en colecciones de documentos, a partir de lo cual se construyen listas de palabras fuertemente relacionadas. El principal problema detectado en estas técnicas es que no todas las palabras relacionadas con una palabra de consulta son significativas en el contexto de la consulta. Este es un aspecto fundamental considerado en nuestro proyecto.

Dado que en la Web existen muchos foros de discusión sobre la misma temática, es posible hallar preguntas y respuestas similares diseminadas en varios de ellos, por lo que generalmente es necesario navegar por varios hilos hasta dar con una solución adecuada. Incluso, a veces es necesario considerar otras características de calidad para evaluar distintas soluciones [1][3][8].

Existen varias propuestas de reuso de conocimiento disponible en foros de discusión: Por ejemplo Chen y Persen [2] implementan un sistema recomendador que busca y agrupa mensajes con contenido similar. Por otro lado, Helic y otros [4], propone clasificar los mensajes de foros de acuerdo a una jerarquía de temas preestablecida. Luego, el enfoque de Nicoletti [17] clasifica los mensajes acorde a una jerarquía de temas obtenido de Wikipedia. Finalmente, existen propuestas de generación de algoritmos de ranking basados en la calidad de los atributos, como el que se plantea en [11].

En base a estos antecedentes, nuestro proyecto tiene como objetivo principal favorecer el reuso de la información contenida en conversaciones existentes en foros de discusión de la Web, con el valor agregado de un análisis de calidad de las fuentes de información. Además, se ha experimentado tanto con la aplicación de algoritmos de análisis de lenguaje natural como de aprendizaje automático, y se está evaluando la aplicación de *sentiment analysis* para mejorar las búsquedas. Por ejemplo, el análisis del lenguaje natural permite analizar el tipo de fragmento dentro de un hilo de discusión [10]. Teniendo esto en cuenta, nuestro proyecto está enfocado en determinar un ranking de soluciones posibles, y cada línea de

investigación dentro del proyecto lo hace desde ópticas diferentes.

2. LINEAS DE INVESTIGACION Y DESARROLLO

Como se ha mencionado, este proyecto de investigación, denominado “Reuso de Conocimientos en Foros de Discusión – Parte II”, está enmarcado en el Programa de Investigación “Desarrollo de Software Basado en Reuso – Parte II”, de la Universidad Nacional del Comahue, con período de vigencia 2017-2020.

Dicho programa extiende la tarea realizada entre 2013 y 2016 en el Programa “Desarrollo de Software Basado en Reuso”. Respecto a este proyecto en particular, el objetivo es extender los estudios realizados sobre reuso de conocimiento en foros de discusión técnicos, incorporando la definición de métodos y algoritmos de recomendación para la asistencia inteligente a usuarios en la búsqueda de soluciones a preguntas recurrentes. Por otra parte, el programa está conformado por otros dos subproyectos que profundizan en las temáticas de Reuso Orientado al Dominio y Reuso Orientado a Servicios.

El programa “Desarrollo de Software Basado en Reuso – Parte II” está desarrollado por el Grupo de Ingeniería de Software de la Universidad Nacional del Comahue, (GIISCo), formado por docentes y estudiantes de la Facultad de Informática de la Universidad Nacional del Comahue, y cuenta con la asesoría y colaboración de otras universidades. En particular, este proyecto se lleva a cabo con la colaboración de la Facultad de Ciencias Exactas de la Universidad Nacional del Centro de la Provincia de Buenos Aires. Aunque el objetivo del Grupo GIISCo es brindar soporte en investigación y transferencia de tópicos relacionados con la Ingeniería de Software, el proyecto también involucra a docentes pertenecientes a otras áreas de la Facultad, como Programación y Teoría de la Computación, lo que permite abordar la investigación desde ópticas diferentes, enriqueciendo el desarrollo.

3. RESULTADOS OBTENIDOS/ESPERADOS

Como antecedentes de este proyecto de investigación, en el año 2013 se presentó un modelo de calidad para foros de discusión en base a modelos de calidad de datos e información en la Web y estándares para la calidad de datos software [9]. La validación de los atributos y subatributos de dicho modelo se realizó mediante encuestas [13]. Durante 2014 se implementó una herramienta para la recuperación de información de foros de discusión técnicos y su análisis mediante un conjunto preliminar de métricas de calidad, a partir del cual se propone un ranking de soluciones posibles para una pregunta. Dicha herramienta fue aplicada en varios casos de estudio con hilos de discusión reales y algunos de sus resultados están presentados en [27].

Entre 2015 y 2016 se avanzó en el análisis de casos de estudio a partir de una cadena de búsqueda y en el estudio del orden esperado comparado con el orden obtenido por medio de las herramientas de análisis de texto [15][16]. Para ello se utilizó la herramienta Lucene, con mecanismos personalizados para establecer *stopwords* (palabras no significativas de búsqueda) propias del dominio . En 2017, se aplicaron estas técnicas en combinación con la base de datos léxica WordNet [24], cuyos resultados preliminares fueron presentados en [28].

Esta línea de investigación se sigue desarrollando en una tesis de doctorado en la cual se evalúa distintas funciones de las bases de datos léxicas [25] para la búsqueda de mensajes relacionados a una pregunta particular.

Por otra lado, se continúan evaluando técnicas de Data Mining y modelos de aprendizaje automático supervisados y no supervisados [18][19], así como técnicas y herramientas de PLN [21] que puedan ser combinadas con las ya aplicadas.

Otra línea en marcha se enfoca en el rol de los usuarios activos de un foro (los que participan compartiendo opiniones y experiencias). Dicho análisis tiene el objetivo de incluir nuevas métricas de calidad en el recomendador de hilos de discusión en construcción. Bajo esta premisa, se han estudiado las propuestas [20] [23] [22] y se está trabajando en una tesina, a partir de una estrategia empírica basada en la observación de hilos de discusión reales obtenidos de la web.

4. FORMACION DE RECURSOS HUMANOS

El proyecto avanza en la línea del proyecto comenzado en 2013, cuyo objetivo era definir un modelo de calidad a partir de información contenida en foros de discusión técnicos.

Actualmente, el proyecto se encuentra conformado por un grupo de docentes, asesores y alumnos de las áreas de Ingeniería en Sistemas, Programación y Teoría de la Computación, trabajando en forma colaborativa e interdisciplinaria.

Las personas que colaboran, asesoran y forman parte del proyecto son:

- Dos docentes investigadores del Departamento de Programación, con dedicación exclusiva, ambos con título de Doctor en Informática.
- Un docente investigador del Departamento de Programación, con una beca doctoral otorgada por el CONICET.
- Dos docentes investigadores con dedicación simple, de los Departamentos de Ingeniería de Computadoras y de Programación.
- Tres estudiantes de Licenciatura en Ciencias de la Computación que están desarrollando sus tesis de grado dentro del proyecto.
- 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, asesorando en temas de aprendizaje automático y lenguaje natural.
- Una docente investigadora externa, perteneciente al Instituto Superior de Ingeniería del Software (ISISTAN) de la Universidad Nacional del Centro de la Provincia de Buenos Aires. Dicha docente tiene un doctorado y experiencia en modelado de usuarios, sistemas de recomendación y Recuperación de Información.

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

5. BIBLIOGRAFIA

[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] I. Rafique et al(2012), "Information Quality Evaluation Framework: Extending ISO 25012 Data Quality Model", International Journal of Computer and Information Sciences, vol.6.

[6] 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.

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

[8] P. Di Maio (2009), Toward Pragmatic Dimensions of Knowledge Reuse and Learning on the Web. Proceedings of I-KNOW'09 and I-SEMANTICS'09, Graz, Austria.

[9] 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,

[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, et al. , *Generación de un Algoritmo de Ranking para Documentos Científicos del Área de las Ciencias de la Computación*, , CACIC 2013, XIX pp. 787-796, 2013.

[12] Singhal,. *Modern information retrieval: A brief overview*.IEEE Data Eng. Bull., 2001, vol. 24, no 4, p. 35-43

[13] N.Martínez Carod et al. *Análisis de la información presente en foros de discusión técnicos.* In CACIC 2013, pp. 847- 856, 2013.

[14] G. Aranda, N. Martínez-Carod, S. Roger, P. Faraci, and A. Cechich. *Una herramienta para el análisis de hilos de discusión técnicos*. In CACIC 2014, pages 803 - 812, Oct. 2014.

[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, pp. 176-190.

[16] V. Zoratto, G. Aranda, S. Roger, A. Cechich, *Analyzing Discussion Forums ThreadsAbout Java Programming Language Usage*, Electronic Journal of SADIO, 2016 .ISSN (versión online): 1514-6774. En revisión. Publicación estimada Noviembre 2016.

[17] M. Nicoletti, S. Schiafino, and D. Godoy. *Mining interests for user profiling in electronic conversations.* Expert Syst. Appl. , 40(2):638-645, Feb. 2013.

[18] I. Witten, E. Frank and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier. 2011

[19] Bing Liu. *Web Data Mining. Exloring Hyperlinks, Contents, and Usage Data*. Springer. 2008

[20] M. Lui and T. Baldwin. *Classifying user forum participants: Separating the gurus from the hacks, and other tales of the internet*. In Proceedings of Australasian Language Technology Association Workshop , pages 49-57, 2010.

[21] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[22] T. Hecking, I. Chounta, and H. U. Hoppe. *Investigating social and semantic user roles in MOOC discussion forums*. In LAK, pages 198-207. ACM, 2016.

[23] S. Bhatia and P. Mitra. *Classifying user messages for managing web forum data*. In Z. G. Ives and Y. Velegrakis, editors, WebDB , pages 13-18, 2012

[24] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. 1990. *WordNet: An online lexical database*. Int. J. Lexicograph. 3, 4, pp. 235–244.

[25] A. Gangemi, R. Navigli, P. Velardi. *The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet*, In Proc. of ODBASE 2003, Catania, Sicily (Italy), 2003, pp. 820–838.

[26] R. Navigli, S. P. Ponzetto. *BabelNet: Building a Very Large Multilingual Semantic Network.* Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden, July 11–16, 2010, pp. 216–225.

[27] N. Martínez Carod, P. Faraci, G. Aranda *Análisis de métricas de calidad en foros de discusión técnicos*, CACIC 2017, pp.650-659

[28] V. Zoratto, N. Martínez Carod, F. Otermín, G. Aranda: *Análisis de estrategias para clasificar contenidos en foros de discusión*, CACIC 2017, pp. 640-649