Búsqueda de Estrategias para la Clasificación del Contenido en Foros Técnicos de Discusión

Nadina Martínez Carod, Gabriela Aranda, Alejandra Cechich,
Valeria Zoratto, 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

En lás últimas décadas la disciplina Information Retrieval ha avanzado considerablemente. Esto se debe gran parte a que las organizaciones actuales hacen cada vez más esfuerzos para reutilizar el conocimiento, definiendo estrategias para tener catalogadas y reutilizar soluciones ya probadas [7].

Por el otro lado la evolución de la Web trajo consigo distintas herramientas de trabajo colaborativo. Dentro de dichas herramientas, los foros de discusión son ampliamente utilizados para plantear problemas, expresar sugerencias, así como para intercambiar conocimientos y experiencias. Dentro de un foro de discusión, un usuario de la comunidad puede realizar una pregunta, y el resto de los miembros de dicha comunidad pueden responder proponiendo soluciones al problema planteado. De esta manera, mediante el uso de esta herramienta, se genera un volumen de información bastante importante, que puede ser utilizado como fuente de conocimiento para un sistema de recuperación de información.

El objetivo fundamental de nuestro proyecto es definir una herramienta que, a partir de información existente en hilos de discusión de foros técnicos, la clasifique y establezca un orden entre soluciones posibles para problemas recurrentes del área de programación.

CONTEXTO

La línea de investigación presentada se denomina “Reuso de Conocimientos en Foros de Discusión II” y forma parte del programa “Desarrollo de Software Basado en Reuso – Parte II”, 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 que se encarga de recuperar información (Information Retrieval) surge en la década de 1950 [12], ante la necesidad de procesar y reutilizar la información almacenada en grandes volúmenes. Desde ese momento, este campo ha madurado y han ido surgiendo valiosos aportes en distintas ramas de investigación. Por ejemplo, algunos proyectos se han enfocado en utilizar la información almacenada en 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 consultas. En general el proceso de recupero de información comienza con la consulta de un usuario al sistema. Las respuestas a dicha consulta poseen diferentes grados de relevancia, y para organizarlas, se determina un ranking el cual evalúa el grado de respuesta a una consulta.

Si bien el conocimiento en la Web se encuentra diseminado en distintos tipos de aplicaciones, los foros de discusión en particular se caracterizan por ser herramientas colaborativas con grandes volúmenes de información, accesibles a la comunidad en general como fuente de consulta (la gran mayoría de los foros cumplen estas características). En estas herramientas se intercambia conocimiento constantemente. Esas son las razones por las cuales nos

enfocamos en los foros de discusión, y específicamente en aquellos que tratan temas técnicos.

En general, la mayoría de los métodos automáticos de IR se basan en analizar la ocurrencia de palabras en los documentos, lo que produce listas de palabras fuertemente relacionadas El principal problema detectado 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 el proyecto.

Dado que existen en la Web muchos foros de discusión sobre la misma temática, se pueden 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 correcta. Incluso muchas veces es necesario considerar características de calidad para evaluar soluciones [1][3][8], desafío que se intenta lograr al dar las posibles soluciones.

Existen varias propuestas de reuso de conocimiento disponible en foros de discusión, como [2] que implementa un sistema recomendador que busca y propone mensajes con contenido similar, en [4], los mensajes se clasifican de acuerdo a una jerarquía de temas preestablecida. El enfoque de Nicoletti [17] clasifica los mensajes acorde a una jerarquía de temas obtenido de Wikipedia. También existen propuestas de generación de algoritmos de ranking basados en la calidad de los atributos, como [11].

Bajo este prisma, nuestro proyecto tiene como objetivo principal favorecer el reuso de la información contenida en conversaciones existentes en la Web, con el valor agregado de un análisis de calidad de dichas 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 analizando la aplicación de sentiment analysis para mejorar las búsquedas. El análisis de lenguaje natural en foros de discusión permite analizar el tipo de fragmento dentro de un hilo de discusión, como en [10]. Por este motivo, nuestro proyecto está orientado a determinar un ranking de soluciones posibles, y cada línea dentro del proyecto se enfoca en esta acción desde diferentes aspectos.

2. LINEAS DE INVESTIGACION Y DESARROLLO

El proyecto de investigación se denomina “Reuso de Conocimientos en Foros de Discusión – Parte II” y está enmarcado dentro del Programa de Investigación “Desarrollo de Software Basado en Reuso – Parte II”, con período de vigencia 2017-2020.

El programa mencionado extiende la investigación realizada durante el programa denominado “Desarrollo de Software Basado en Reuso”, realizado en el período 2013-2016. 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.

Dicho programa 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, junto con asesoría y colaboración de otras universidades. En particular, este proyecto es desarrollado en colaboración con 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 con un trabajo conjunto y colaborativo.

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 datos y de información en la Web y en estándares para la calidad de datos software [9]. La validación de la selección de atributos y sub-atributos de dicho modelo se realizó mediante encuestas [13]. Además, durante 2014 se implementó una versión preliminar de una herramienta que permite la recuperación de información desde foros de discusión técnicos y su análisis mediante un conjunto preliminar de métricas de calidad, a partir de las cuales 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 sus resultados están presentados en una tesis de licenciatura con fecha de defensa a realizarse próximamente.

Además, durante 2015 y 2016 se avanzó en el análisis de casos de estudio más amplios, a partir de una cadena de búsqueda y en el estudio del orden esperado confrontado al 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 propias del dominio. Actualmente, se está avanzando en el uso de bases de datos léxicas (como WordNet [24]), en combinación con otras estrategias de Recuperación de Información. Esta línea de investigación está siendo desarrollada como parte de otra tesina que evaluará los resultados obtenidos al aplicar distintas funciones de las bases de datos léxicas [25] en la búsqueda de mensajes relacionados a una pregunta particular. Durante el desarrollo de esta tesina se establecerán nuevos corpus de hilos de discusión reales sobre los cuales aplicar las técnicas seleccionadas implementando una estrategia de validación empírica.

Por otra lado se está evaluando la aplicación de técnicas de Data Mining y de modelos de aprendizaje automático supervisados y no supervisados [18][19] como así también técnicas y herramientas disponibles de PLN [21] que puedan ser combinadas con las de aprendizaje automático para facilitar la incorporación de contenido sintáctico y/o semántico en la creación de ejemplos de forma automática. Este es el objetivo de una tesina de licenciatura que esta comenzando a desarrollarse.

Otro enfoque que está comenzando a estudiarse es el papel que juegan los distintos usuarios activos dentro de un foro (los que participan compartiendo opiniones y experiencias). Dicho análisis se está realizando para incorporarlo como posible mejora del recomendador de hilos de foros de discusión deseado. Bajo esta premisa, se han estudiado diferentes propuestas [20] [23] [22] y se está trabajando en una tesis en curso, 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, el cual tenía como objetivo definir un modelo de calidad a partir de información contenida en foros de discusión técnicos.

El proyecto actualmente se encuentra conformado por un grupo de docentes, asesores y alumnos desarrollándose en 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:

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.

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.