Análisis de estrategias para clasificación de usuarios y post dentro de un hilo de discusión

Valeria Zoratto, Gabriela Aranda, Nadina Martinez Carod, Alejandra Cechich,
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: {vzoratto, gabriela.aranda, nadina.martinez}@fi.uncoma.edu.ar

RESUMEN

La Web actual se ha transformado en una plataforma que posibilita el encuentro de ideas y favorece la creación de debates en chat, blogs, foros de discusión, etc. En particular la comunidad informática suele aprovechar los medios disponibles en la Web de soporte grupal [1], tanto para solucionar problemas como para el aprendizaje de alguna tarea particular. Es por ello que este tipo de herramientas han tenido un gran auge en las últimas décadas, dentro de las cuales los foros de discusión se han convertido en los más utilizados para aprendizaje o como proveedor de soluciones de algún problema específico. Los foros de discusión generan contenido de manera continua lo que produce un gran volumen de información, que puede ser utilizado como fuente de conocimiento para un sistema de Information Retrieval (IR).

Las organizaciones actuales hacen cada vez más esfuerzos para reutilizar el conocimiento, definiendo estrategias para tener catalogadas y reutilizar soluciones ya probadas por lo que la disciplina de IR ha avanzado considerablemente.

El objetivo fundamental de nuestro proyecto es definir una herramienta que, a partir de información contenida en hilos de foros de discusión técnicos, pueda descargar dicha información de manera automática, la pueda clasificar de acuerdo a temas específicos, así como también poder establecer un ranking de soluciones posibles, teniendo en cuenta además a los usuarios involucrados en dichos foros.

CONTEXTO

La línea de investigación presentada se denomina “Reúso de Conocimientos en Foros de Discusión II” y forma parte del programa “Desarrollo de Software Basado en Reúso – Parte II”, con período de vigencia 2017-2020. El programa mencionado extiende el programa “Desarrollo de Software Basado en reúso” realizado durante el período 2013-2016.

1. INTRODUCCIÓN

La disciplina de Information Retrieval (IR) surge en la década de 1950 [2], ante la necesidad de reutilizar grandes volúmenes de información. 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.

Los foros son espacios web virtuales donde las personas pueden hacer preguntas, responder y participar en discusiones sin necesidad de estar en el mismo espacio geográfico ni en el mismo momento [3]. En los últimos tiempos han surgido foros

populares (específicos en el área de informática) como StackOverflow[1], Ubuntu Forum[2], etc. La información que se encuentra en dichos foros puede ser muy rica ya que puede ser utilizada por múltiples usuarios que tengan problemas similares.

La creación de un hilo comienza a partir de una serie de preguntas sobre un problema específico generada por un usuario de la comunidad del foro. Luego el usuario espera la respuesta del resto de los usuarios de la comunidad, estas respuestas pueden no ocurrir inmediatamente después de publicada la pregunta, ya que los participantes del foro no están físicamente presentes. De esta manera un usuario puede contestar a la pregunta estableciéndose así una comunicación entre las partes a fin de encontrar una solución.

Si bien las soluciones propuestas están focalizadas al usuario que abrió el debate, la totalidad de los mensajes queda disponible al público, y las soluciones pueden ser reutilizadas por participantes con problemas similares.

Muchos foros contienen diferentes tipos de usuarios, generalmente existen moderadores, administradores, usuarios registrados y usuarios anónimos. Algunos foros, asignan diferentes tipos de etiquetas a los usuarios registrados, dependiendo de la interacción que tenga este en la comunidad.

En general las técnicas actuales de IR se basan en medidas de similitud de palabra clave y no consideran algunas características importantes para analizar discusiones en hilos [4]. A menudo, los hilos que comparten palabras clave comunes discuten diferentes temas y, en tales casos, encontrar hilos relevantes se vuelve un desafío para los usuarios [5].

Este es un aspecto fundamental considerado en nuestro proyecto.

Existen varias propuestas de reúso de conocimiento disponible en foros de discusión: Por ejemplo, Elsas & Carbonell [6] fueron los primeros en revisar estrategias para la recuperación de hilos en una colección de prueba de pares <query, relevant document>. Por otro lado Seo et al. [7] describen cómo las estructuras de respuesta en los hilos de un foro pueden ser extraídos y utilizados para la recuperación a nivel de hilo y post. En otro trabajo, Bhatia & Mitra [3] emplean redes de inferencia para calcular evidencias de diferentes unidades estructurales. Este modelo utiliza múltiples atributos como el post inicial y las respuestas, longitud de los mensajes, medida de autoridad del post y los enlaces entre los posts en un marco de red de inferencia unificada basada en filas de clasificación en respuesta a las consultas de los usuarios.

Otros autores hacen foco en la jerarquía de información de los foros. Por ejemplo, Helic et al.[8], proponen clasificar los mensajes de foros de acuerdo a una jerarquía de temas pre-establecida. Luego, el enfoque de Nicoletti [9] clasifica los mensajes acordes a una jerarquía de temas obtenido de Wikipedia. Por otro lado, la investigación de Hecking et. at. [10] describe el análisis de la estructura social y semántica de los foros de discusión en cursos MOOC en términos de intercambio de información y roles de usuario.

Liu [11], en cambio, busca predecir si un autor de la pregunta estará satisfecho con las respuestas enviadas por los participantes de la comunidad.

En base a estos antecedentes, nuestro proyecto tiene como objetivo principal favorecer el reúso de la información contenida en conversaciones existentes en foros de discusión de la Web, con el valor agregado de un análisis de los roles de usuarios para poder detectar usuarios expertos y darle un mayor peso a las respuestas candidatas de una pregunta. 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

[1] http://stackoverflow.com
[2] https://ubuntuforums.org

análisis del lenguaje natural permite analizar el tipo de fragmento dentro de un hilo de discusión [12]. 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. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

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

El programa mencionado extiende la investigación realizada durante el programa denominado "Desarrollo de Software Basado en Reúso", realizado en el período 2013-2016. Respecto a este proyecto en particular, el objetivo es extender los estudios realizados sobre reúso 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 frecuentes. Por otra parte, el programa está conformado por otros dos subproyectos que profundizan en las temáticas de Reúso Orientado al Dominio y Reúso 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 calidad de datos e información en la Web y estándares para la calidad de datos software [13]. La validación de los atributos y sub-atributos de dicho modelo se realizó mediante encuestas [14]. 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 exhibidos en [15].

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 [16][17]. Para ello se utilizó la herramienta Lucene, con mecanismos personalizados para establecer stopwords (palabras que no aportan significancia) propias del dominio. En 2017, se aplicaron estas técnicas en combinación con la base de datos léxica WordNet [18], cuyos resultados preliminares fueron presentados en [19]. En 2018, se apuntó a mejorar el proceso de recuperación de hilos mediante la incorporación de sinónimos de las palabras utilizadas en los hilos de discusión teniendo en cuenta a la estructura semántica de la oración. Para ello se utilizó WordNet de

manera conjunta con la aplicación Stanford POS Tagger[3] [21]. 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 [27] 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 [22][23], así como técnicas y herramientas de PLN [24] 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). Bajo esta premisa, se han estudiado las propuestas [20] [25] [10] y se está trabajando en una tesina, a partir de una estrategia empírica basada en la observación de hilos de discusión obtenidos de la web. Por otro lado se están estudiando las propuestas de [11][28][29] y se está trabajando en otra tesis para detectar la satisfacción de una respuesta del usuario.

4. FORMACIÓN 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: 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.

Dos docentes investigadores con dedicación simple, de los Departamentos de Ingeniería de Sistemas 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 el trabajo cooperativo de un grupo interdisciplinario. Además, la incorporación de estudiantes de la Facultad amplía los posibles tipos de desarrollo relacionados a la temática del proyecto.

5. BIBLIOGRAFÍA

[1] C. A. Ellis, S. J. Gibbs, and G. L. Rein, "Groupware: Some Issues and Experiences", Communications of ACM 34, 1 (1991), pp. 38-58.

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

[3] S. Bhatia and P. Mitra, "Adopting Inference Networks for Online Thread Retrieval.", in AAAI vol. 10, (, 2010), pp. 1300--1305.

[4] D. Feng, E. Shaw, J. Kim, and E. Hovy, "Learning to detect conversation focus of threaded discussions", in Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of … (, 2006), pp. 208--215.

[3] https://nlp.stanford.edu/software/tagger.shtml

[5] P. Biyani, S. Bhatia, C. Caragea, and P. Mitra, "Using Subjectivity Analysis to Improve Thread Retrieval in Online Forums", in European Conference on Information Retrieval (2015), pp. 495--500.

[6] J.L Elsas and J. G Carbonell, "It pays to be picky: an evaluation of thread retrieval in online forums", in Proceedings of the 32nd international ACM SIGIR (2009), pp. 714--715.

[7] J. Seo, W B. Croft, and D.A Smith, "Online community search using conversational structures", Information Retrieval 14, 6 (2011), pp. 547.

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

[9] M. Nicoletti, S. Schiafino, and D. Godoy. Mining interests for user profiling in electronic conversations. Expert Syst. Appl., Feb. 2013.

[10] 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.

[11] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering", 31st annual international ACM SIGIR (2008), pp. 483--490.

[12] 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.

[13] 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,

[14] N.Martínez Carod, G. Aranda. Análisis de la información presente en foros de discusión técnicos. In CACIC 2013, pp. 847- 856, 2013.

[15] 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, 2014.

[16] 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.

[17] V. Zoratto, G. Aranda, S. Roger, A. Cechich, Analyzing Discussion Forums ThreadsAbout Java Programming Language Usage, Electronic Journal of SADIO, 2016.

[18] 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.

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

[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] G. Aranda, V. Zoratto, N. Martínez Carod, Sandra Roger, F. Otermín, A. Cechich. Clasificación de contenido de hilos de discusión mediante análisis sintáctico y morfológico. CICCSI 2018.

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

[23] B. Liu. Web Data Mining. Exloring Hyperlinks, Contents, and Usage Data. Springer. 2008

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

[25] 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

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

[27] 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.

[28] Choi, E., Kitzie, V., & Shah, C. (2014). Investigating motivations and expectations of asking a question in social Q&A. *First Monday*, *19*(3).

[29] Agichtein, E., Liu, Y., & Bian, J. (2009). Modeling information-seeker satisfaction in community question answering. ACM Transactions on Knowledge Discovery from Data (TKDD), 3(2), 10.