Reuso de información en comunidades virtuales

Gabriela Aranda, Valeria Zoratto, Nadina Martinez Carod, Romina Schroeder
Andrés Flores, Natalia Baeza, Lucas Cavaliere, Sandra Lucero

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, vzoratto, nadina.martinez}@fi.uncoma.edu.ar

RESUMEN

Con el comienzo del nuevo siglo, la Web 2.0 permitió que cualquier usuario con conexión a internet y sin conocimientos especiales, pueda publicar contenido propio mediante herramientas colaborativas como blogs, wikis, y foros de discusión. Al mismo tiempo surgieron plataformas para creación de redes sociales de distinto tipo o con objetivos diferentes, concretamente se destaca el trabajo distribuido colaborativo, el intercambio de conocimiento técnico, académico, científico y lo social. Estas comunidades que fueron formándose a partir de la participación online recibieron el nombre de comunidades virtuales.

Durante los últimos años hubo una evolución hacia la Web 3.0 la cual se ocupa de, por un lado, facilitar la accesibilidad a la información sin depender del dispositivo utilizado, y por otro lado, el modo que las personas interactúan con ella para conseguir los resultados que desean. Además, se enfoca en que la información sea compartida de una forma inteligible, que sea útil para los usuarios que recurren a ella y a las necesidades especiales de sus usuarios de acuerdo a las circunstancias. Así, nuestro proyecto se enfocará en proveer modelos de calidad para sistemas software de recuperación, análisis, clasificación y reuso de la información proveniente de comunidades virtuales en la Web.

Palabras clave

Recuperación de Información, calidad de datos, comunidades virtuales, participación ciudadana.

CONTEXTO

Esta línea de investigación forma parte del proyecto de investigación “Reuso de información en comunidades virtuales”, de la Universidad Nacional del Comahue, en estado de evaluación, con período de vigencia 2022-2025. Dicha línea extiende y avanza sobre temas desarrollados por el equipo de investigadores principales en proyectos anteriores, pertenecientes al Programa “Desarrollo de Software basado en Reúso”, de la Universidad Nacional del Comahue, llevado a cabo entre 2013-2021. El nuevo proyecto continúa las líneas de investigación enfocadas en la recuperación de información disponible en foros de discusión y abarca nuevas tecnologías para soporte a comunidades virtuales con una mirada orientada a la participación ciudadana y la toma de decisiones basada en la opinión pública.

1. INTRODUCCIÓN

El término “comunidades virtuales” fue establecido en 1993 por Howard Rheingold para referirse a grupos sociales que emergen a partir de la interacción de personas en

espacios públicos de internet [1]. Dentro de esta definición se enmarcan distintos tipos de entornos colaborativos como las Comunidades de Preguntas y Respuestas (CQA, del inglés *Community Question Answering*), las plataformas para redes sociales (como Facebook, Twitter o Instagram), entre otras.

Particularmente, las CQA ofrecen a los usuarios la posibilidad de buscar y compartir conocimiento a partir de la formulación de preguntas, respondiendo las mismas y/o comentando. Entre las CQA se encuentran los foros de discusión, los cuales permiten el debate acerca de un tema en particular en forma de hilos de mensajes, teniendo una estructura similar aún entre distintas plataformas existentes. Por ejemplo, sitios como Quora (que permite hacer preguntas de todo tipo) y otros de dominios más específicos, como StackOverflow (para programación) o Mathematics StackExchange (para matemáticas), son sitios utilizados diariamente por millones de usuarios para encontrar respuesta a preguntas complejas, subjetivas o dependientes del contexto [2, 3]. Con el tiempo, creció considerablemente la cantidad de información acumulada en dichos sitios, por lo que analizar y reutilizar dicha información es algo "deseable y valioso" [4, 5]. Así, el reuso e integración de información (IRI, del inglés *Information Reuse and Integration*) juega un rol esencial ya que se enfoca en la captura, representación, mantenimiento, integración, validación y extrapolación de información, que luego puede ser aplicada para mejorar la toma de decisiones en varios dominios de aplicación [6].

Con referencia a esto último, existen varias investigaciones enfocadas en analizar, evaluar y extraer la información proveniente de CQA. Por ejemplo, Le et al. [7] y Amancio et al. [8] analizan los hilos de discusión para determinar cómo se constituye la calidad de las respuestas. En cambio, Neshati [3] evalúa la calidad del contenido desde la perspectiva del resultado de la votación. Por otro lado, algunos autores consideran que todo lo necesario para evaluar la calidad de una respuesta está en su contenido, es decir, tienen en cuenta sólo las características textuales de los hilos [9, 10, 11], mientras que otros combinan tanto las características textuales como la red de usuarios [12, 13]. Otros trabajos se enfocan en la calidad del contenido en base al nivel de experticia del usuario que responde, pero dado que no todas las CQA mantienen información de la reputación de un usuario como un puntaje o categoría asociada, una línea de investigación relevante es la de estimar su experticia basado en las evidencias disponibles, como la calidad de sus intervenciones y su interrelación con otros usuarios [14]. Es decir que tanto el análisis del contenido textual de los mensajes como el estudio de las redes sociales subyacentes son temas muy relacionados a este tipo de investigación.

Respecto al análisis de dichas redes sociales se requiere de un conjunto de métodos y técnicas en los que se mezclan teorías sociológicas y matemáticas. Para analizar las estructuras sociales de estas redes se utilizan los conceptos, el vocabulario y operaciones de la teoría de grafos que permiten probar teoremas sobre los grafos que las modelan, así como deducir y someter determinados enunciados a tests [15]. Algunos de los trabajos más actuales en esta área están enfocados en el estudio de información proveniente de redes sociales como Twitter [16], pero dado que cada red tiene intereses y características únicas, es pertinente evaluar el comportamiento de las personas en otros tipos de plataformas colaborativas; por ejemplo para detectar usuarios que emiten información irrelevante, los que son líderes de opinión (vistos como fuentes de conocimiento dentro de su comunidad) o los denominados *boundary spanner* (usuarios que permiten comunicación entre distintas comunidades) [17].

Como se ha mencionado, el gran volumen de información generado por las CQA y las redes sociales existentes en la Web es

propicio para estudiar y definir técnicas para reuso e integración de información [6], por lo que este proyecto se enfocará en definir técnicas para captura de información, análisis de contenido, detección de redes sociales y clasificación de perfiles de usuarios utilizando para su evaluación corpus de comunidades virtuales existentes como StackExchange[1], que es una red de webs para CQA que cuenta con más de 50 foros de discusión de diferentes temáticas, 14 millones de usuarios, 21 millones de preguntas y 31 millones de respuestas, cuyas bases de datos son accesibles de forma libre para investigación[2]. Este conocimiento puede ser aplicado en comunidades virtuales más específicas, como por ejemplo, las conformadas a partir de plataformas para la participación ciudadana. Este tipo de plataformas han surgido en los últimos años a partir de los desafíos que enfrentan las ciudades para garantizar la calidad de vida de sus habitantes y mejorar los procesos de toma de decisiones incorporando la opinión pública. Por ello, la participación ciudadana es una herramienta que mejora la gobernanza local y la toma de decisiones, que busca ser una forma directa de conocer las necesidades, demandas e ideas de los individuos que la componen [18]. En muchos casos la toma de decisión se realiza a partir de la opinión de la ciudadanía obtenida mediante herramientas colaborativas, como sitios web y/o aplicaciones móviles, que permiten conocer a corto plazo los efectos que pueden tener los cambios realizados, por ejemplo, en el espacio urbano. Dado que dichas tecnologías se basan en recuperar opiniones de ciudadanos, surgen temáticas de análisis y evaluación que, muchas veces son comunes a otras comunidades virtuales como las que hemos mencionado anteriormente. Con este objetivo, como parte de este proyecto se evaluarán productos existentes para participación ciudadana (como *decidim*[3], *CONSUL*[4] y *WeLive*[5] [19, 20]) y se trabajará en su adaptación para aplicarlos en el ámbito de barrios de la ciudad de Neuquén. Además, se planea hacer uso de las técnicas y herramientas elaboradas como parte de otras líneas de investigación, para hacer aportes al reuso de información como soporte para la toma de decisiones basada en la opinión de los ciudadanos.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

El proyecto de investigación actual se denomina “Reuso de información en comunidades virtuales” y su objetivo principal es definir cuáles son las mejores técnicas y algoritmos de recomendación para la asistencia inteligente a usuarios de comunidades virtuales en la búsqueda de información relevante

Este proyecto está desarrollado por integrantes del 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. El proyecto también involucra a docentes pertenecientes a otras áreas de la Facultad, como Programación, Ingeniería en Computación, Ingeniería en Sistemas 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

A partir del 2013 hemos realizado investigación en temas relacionados a

[1] https://stackexchange.com/
[2] https://archive.org/download/stackexchange
[3] https://decidim.org/es/
[4] https://consulproject.org/en/
[5] https://welive.eu/

comunidades virtuales de programadores, proponiendo un modelo de calidad para foros de discusión técnicos y determinando criterios para la calidad de información contenida en un hilo de discusión [22], que fueron validados mediante encuestas y se formularon variaciones en la parametrización para mejorar los resultados obtenidos [23].

También se trabajó en el procesamiento del texto de hilos de discusión de foros técnicos, implementando una herramienta ad-hoc (basada en Lucene) para recuperación de información y análisis según un conjunto propio de medidas de calidad para proponer un ranking de soluciones posibles para una pregunta [24]. Luego, se mejoró dicha herramienta incluyendo sinónimos con la base de datos léxica WordNet y el analizador morfológico Stanford POS Tagger para identificar el rol de las palabras en el contexto que eran utilizadas [25].

Además, se avanzó en la clasificación de roles de usuarios activos de un foro, para determinar la jerarquía de roles determinados por el nivel de conocimiento de los participantes en los hilos de discusión de acuerdo a los posts realizados por dichos usuarios [26].

Utilizando dichos conocimientos obtenidos por este grupo de trabajo en estudios anteriores, se espera aplicarlos en el análisis de información proveniente de comunidades virtuales creadas ad-hoc para los barrios de nuestra ciudad y aportar a la toma de decisiones basada en opinión pública.

4. FORMACIÓN DE RECURSOS HUMANOS

El proyecto se encuentra conformado por docentes de diferentes áreas debido a su naturaleza multidisciplinaria. Las personas que forman parte del proyecto, tanto como colaboradores, asesores o integrantes son:

- Tres docentes investigadores del Departamento de Programación e Ingeniería de Sistemas, con dedicación exclusiva, con Doctorado en Informática.
- Una docente investigadora del Departamento de Programación con beca del CONICET para realización de doctorado.
- Tres docentes con dedicación simple, pertenecientes a los Departamentos de Programación, Ingeniería de Sistemas e Ingeniería en Computación.
- Una profesora adjunta, asesora local, con dedicación exclusiva, del Departamento de Teoría de la Computación.
- 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 (UNCPBA), con experiencia en Sistemas de Recomendación y Recuperación de Información. Doctora en Ciencias de la Computación.
- Tres estudiantes de la carrera de Licenciatura en Ciencias de la Computación que realizan sus tesis dentro del proyecto.

De esta manera, se van incorporando actividades para extender líneas de investigación al proyecto inicial con nuevos enfoques.

5. BIBLIOGRAFÍA

[1] H. Rheingold. The Virtual Community, revised edition: Homesteading on the Electronic Frontier. MIT press, 2000.

[2] I. Srba and M. Bielikova. A comprehensive survey and classification of approaches for community question answering. ACM Trans. Web, 10(3), 2016.

[3] M. Neshati. On early detection of high voted qa on stack overflow. Information Processing Management, 53(4):780–798, 2017.

[4] G. Cong, L. Wang, C. Lin, Y. Song, and Y. Sun. Finding Question-answer Pairs from Online Forums. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pages 467–474, New York, NY, USA, 2008. ACM.

[5] S. Gottipati, D. Lo, and J. Jiang. Finding relevant answers in software forums. In 26th IEEE/ACM International Conference on Automated Software Engineering (ASE 2011), Lawrence, KS, USA, November 6-10, 2011, pages 323–332, 2011.

[6] M. Day, C. Ong, and W. Hsu. An analysis of research on information reuse and integration (2003-2008). International Transactions on Systems Science and Applications, 6(2):146–157, 2010.

[7] L.T. Le, C. Shah, and E. Choi. Evaluating the Quality of Educational Answers in Community Question-Answering. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL '16, pages 129–138, New York, NY, USA, 2016. Association for Computing Machinery.

[8] L. Amancio, C. Dorneles, and D. Dalip. Recency and quality-based ranking question in CQAs: A stack overflow case study. Information Processing Management, 58(4):102552, 2021.

[9] G. Gkotsis, K. Stepanyan, C. Pedrinaci, J. Domingue, and M. Liakata. It's all in the content: state of the art best answer prediction based on discretisation of shallow linguistic features. In Proceedings of the 2014 ACM conference on Web science, pages 202–210, 2014.

[10] G. Burel, P. Mulholland, and H. Alani. Structural Normalisation Methods for Improving Best Answer Identification in Question Answering Communities. In Proceedings of the 25th International Conference Companion on World Wide Web, pages 673–678, 2016.

[11] Y. Pérez-Guadarramas, A. Rodríguez-Blanco, A. S. Cuevas, W. Hojas-Mazo, y J. A. Olivas. Combinando patrones léxico-sintácticos y análisis de tópicos para la extracción automática de frases relevantes en textos. Procesamiento del Lenguaje Natural, (59):39– 46, 2017.

[12] D. Kundu and D. Prasad Mandal. Formulation of a hybrid expertise retrieval system in community question answering services. Applied Intelligence, 49(2):463–477, 2019.

[13] H. Fu and S. Oh. Quality assessment of answers with user identified criteria and data-driven features in social qa. Information Processing Management, 56(1):14–28, 2019.

[14] M. Neshati, Z. Fallahnejad, and H. Beigy. On dynamicity of expert finding in community question answering. Information Processing Management, 53(5):1026–1042, 2017.

[15] L. Sanz-Menéndez. Análisis de redes sociales: O cómo representar las estructuras sociales subyacentes. Apuntes de Ciencia y Tecnología, 7:21– 29, 06 2003.

[16] R. Olivares, F. Muñoz, and F. Riquelme. A multiobjective linear threshold influence spread model solved by swarm intelligence-based methods. Knowledge-Based Systems, 212:106623, 2021.

[17] P. Matous and P. Wang. External exposure, boundary-spanning, and opinion leadership in remote communities: A network experiment. Soc. Networks, 56:10–22, 2019.

[18] D. García Castro, V. De Elizagarate Gutierrez, J. Kazak, S. Szewranski, I. Kaczmarek, and T. Wang. Nuevos desafíos para el perfeccionamiento de los procesos de participación ciudadana en la gestión urbana. retos para la innovación social. Management Letters/Cuadernos de Gestión, 20(1):41–64, 2020.

[19] I. Peña-López. Shifting participation into sovereignty: the case of decidim.barcelona. 03 2019.

[20] M. X. Rivera Rásury. Desarrollo de una herramienta de soporte metodológico a los procesos de e-participación. Master, Departamento de Sistemas Informáticos y Computación Universitat Politécnica de Valencia, Valencia, España, 2018.

[21] J. Levy Moreno and H. H. Jennings. "Statistics of Social Configurations." *Sociometry*, vol. 1, no. 3/4, pp. 342–374. *JSTOR*, 1938.

[22] 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, Buenos Aires, pp.803-812, 2014.

[23] G. Aranda, V. Zoratto, N. Martínez Carod, S. Roger, F. Otermin, A. Cechich. Clasificación de contenido de hilos de discusión mediante análisis sintáctico y morfológico. CICCSI 2018. ISBN 9789874568366. Mendoza, 2018, pp. 35-44.

[24] V. Zoratto, G. Aranda, S. Roger, A. Cechich. Analyzing Discussion Forums Threads About Java Programming Language Usage, Electronic Journal of SADIO, ISSN 1514-6774, 2016.

[25] V. Zoratto, G. Aranda, N. Martinez Carod, F. Otermin. Evaluación de estrategias para clasificar hilos de foros de discusión según su contenido, ASSE-JAIIO 2021, Argentine Symposium on Software Engineering, Argentina, 2021.

[26] N. Martínez Carod, G. Aranda, V. Zoratto, C. Murray (2019), Una propuesta para clasificación de roles de usuarios en foros de discusión técnicos. CACIC 2019, Argentina, 2019.