Calidad de información en comunidades virtuales

Valeria Zoratto, Gabriela Aranda, 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: {vzoratto, gabriela.aranda, nadina.martinez}@fi.uncoma.edu.ar

RESUMEN

La Web 2.0 se caracterizó por no ser simplemente un contenedor de información, sino en un plataforma de trabajo colaborativo donde se desarrollan servicios web, herramientas y aplicaciones que permitieron a los usuarios crear y compartir contenido en l web, como blogs, wikis, redes sociales y plataformas de video y música en línea, sin necesidad de conocimientos técnicos especializados. Al mismo tiempo, emergieron diversas plataformas para la creación de redes sociales con objetivos diversos, como el trabajo en colaboración a distancia, así como de intercambio de conocimiento técnico, académico, científico y social. Estas comunidades virtuales, construidas a través de la participación en línea, fueron reconocidas como un fenómeno sin precedentes en la historia de la comunicación digital.

La Web 3.0 es la siguiente etapa en la evolución de la web y se centra en la creación de una web semántica, en la que la información se organiza y se presenta de una manera más inteligible y útil para los usuarios. La Web 3.0 también se enfoca en la accesibilidad de la información, independientemente del dispositivo utilizado, y en la forma en que las personas interactúan con ella para obtener resultados precisos y relevantes.

Como parte de este tránsito hacia la Web 3.0, nuestro proyecto se centra en proporcionar modelos de calidad para sistemas de software que permitan la recuperación, análisis, clasificación y reutilización 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, con período de vigencia 2022-2025. Dicha línea extiende y avanza sobre temas desarrollados por el equipo de investigadores principales en proyectos pertenecientes al Programa “Desarrollo de Software basado en Reuso”, Parte I y II, de la Universidad Nacional del Comahue, llevado a cabo entre los años 2013 y 2021. En este nuevo proyecto se continúan 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 opinión pública.

1. INTRODUCCIÓN

En 1993, el término "comunidades virtuales" fue acuñado por Howard Rheingold para describir grupos sociales que se forman a través de la interacción en línea en espacios públicos de Internet [1]. Desde entonces una variedad de entornos colaborativos han surgido, incluyendo plataformas de redes sociales como Facebook, Twitter e Instagram, así como las Comunidades de Preguntas y Respuestas (CQA, por sus siglas en inglés).

Las CQA, como los foros de discusión, permiten a los usuarios buscar y compartir conocimientos a través de preguntas y respuestas. Estos sitios han crecido en popularidad en todo el mundo, y son utilizados diariamente por millones de usuarios para encontrar respuestas a preguntas complejas, subjetivas o específicas de un contexto [2, 3]. Sin embargo, con el creciente volumen de información en estos sitios, surge la necesidad de analizar y reutilizar esta información [4, 5]. La integración y reutilización de información (IRI, por sus siglas en inglés) juega un papel clave en la captura, representación, mantenimiento, integración, validación y extrapolación de información que se puede aplicar para mejorar la toma de decisiones en varios dominios de aplicación [6].

En relación a lo anterior, se han llevado a cabo diversos estudios para analizar, evaluar y extraer información de las CQA. Por ejemplo, algunos se centran en analizar los hilos de discusión para determinar los criterios que definen la calidad de las respuestas [7, 8]. Otros investigan técnicas de respuesta a preguntas en CQA y clasifican los sistemas de respuesta en dos categorías: sistemas basados en votación [3] y sistemas basados en modelos [27]. Además, algunos autores sostienen que la calidad de una respuesta se puede evaluar únicamente considerando las características textuales de los hilos [9, 10, 11], mientras que otros combinan dichas características con la red de usuarios [12, 13]. También se ha investigado la experticia de los usuarios que responden, estimando su experiencia basándose en evidencias como la calidad de sus intervenciones y su interrelación con otros usuarios [14]. En resumen, tanto el análisis del contenido textual de los mensajes como el estudio de las redes sociales subyacentes son temas muy relacionados con este tipo de investigación.

En relación al análisis de redes sociales en CQA, es importante destacar que este tipo de análisis requiere de una combinación de métodos y técnicas que involucran teorías sociológicas y matemáticas. Para analizar las estructuras sociales de estas redes, se utilizan los conceptos, 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 pruebas [15]. Además, los estudios más recientes en esta área se enfocan en la información proveniente de redes sociales como Twitter [16]. Sin embargo, dada la diversidad de intereses y características de cada red, es importante evaluar el comportamiento de las personas en otros tipos de plataformas colaborativas. Por ejemplo, es relevante identificar a los usuarios que emiten información irrelevante, a los líderes de opinión que son considerados fuentes de conocimiento dentro de su comunidad, o a los boundary spanners, es decir, usuarios que permiten la comunicación entre distintas comunidades [17]. De esta manera, se puede entender mejor la dinámica social y de conocimiento en CQA y cómo estas pueden ser aprovechadas para mejorar la calidad de las respuestas y la experiencia de los usuarios.

Como se ha mencionado, el gran volumen de información generado por las CQA existentes en la Web es propicio para el estudio y definición de técnicas para el reuso e integración de información [6], por lo que nuestro proyecto se enfoca en definir métodos para capturar información, realizar análisis de contenido, así como detectar y clasificar 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 173 foros de discusión de diferentes temáticas, y que solo por mes recibe más de 100 millones de usuarios que realizan preguntas o responden a otros. Además, sus bases de datos son accesibles de forma libre para investigación[2]. Luego, el conocimiento adquirido puede ser aplicado en comunidades virtuales más específicas, como por ejemplo, las conformadas a partir de plataformas para la participación ciudadana, que 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 para 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 nos proponemos evaluar productos existentes para participación ciudadana (como *decidim*[3], *CONSUL*[4] y *WeLive*[5] [19, 20]) y trabajar 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.

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

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

El proyecto actual se denomina “Reuso de información en comunidades virtuales” y su objetivo principal es definir técnicas y algoritmos de recomendación para 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 la Facultad de Ciencias Exactas de la Universidad Nacional del Centro de la Provincia de Buenos Aires, y una colaboradora de la Facultad de Humanidades de la Universidad Nacional del Comahue que se desempeña en proyectos de investigación en participación ciudadana. Además, la participación de docentes pertenecientes a otras áreas de la Facultad (Programación, Ingeniería en Computación, Ingeniería en Sistemas y Teoría de la Computación), permiten abordar la investigación desde ópticas diferentes, enriqueciendo el desarrollo con un trabajo conjunto y colaborativo.

3. RESULTADOS OBTENIDOS/ESPERADOS

Desde 2013, nuestro equipo de investigación ha estado trabajando en temas relacionados con las comunidades virtuales de programadores. En particular, hemos propuesto un modelo de calidad para foros de discusión técnicos y hemos determinado criterios para evaluar la calidad de la información contenida en los hilos de discusión [22]. Estos criterios han sido

validados mediante encuestas, y hemos formulado variaciones en la parametrización para mejorar los resultados obtenidos [23].

También hemos trabajado en el procesamiento de texto de hilos de discusión en foros técnicos. Para ello, hemos desarrollado una herramienta ad-hoc basada en Lucene, que nos ha permitido recuperar información y analizarla según un conjunto de medidas de calidad que hemos definido, con el fin de proponer un ranking de soluciones posibles para una pregunta [24]. Posteriormente, hemos mejorado esta herramienta incluyendo sinónimos en la base de datos léxica WordNet y el analizador morfológico Stanford POS Tagger para identificar el rol de las palabras en el contexto en el que son utilizadas [25].

Además, hemos avanzado en la clasificación de roles de usuarios activos en un foro, con el objetivo de determinar la jerarquía de roles basados en el nivel de conocimiento de los participantes en los hilos de discusión, según los posts realizados por dichos usuarios [26].

Gracias a los conocimientos adquiridos en nuestros estudios anteriores, y a la incorporación de herramientas de participación ciudadana pretendemos aplicarlos en el análisis de información proveniente de comunidades virtuales creadas específicamente para los barrios de nuestra ciudad. De esta manera, esperamos poder contribuir a la toma de decisiones basadas en la 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 con dedicación exclusiva del Departamento de Programación, finalizando el Doctorado en Ciencias de la Computación.
- 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 de la Facultad de Humanidades de la misma universidad, investigadora del Instituto Patagónico de Estudios de Humanidades y Ciencias Sociales - CONICET.
- 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.

[27] Khot, T., Sabharwal, A., & Clark, P. (2017). Answering complex questions using open information extraction. arXiv preprint arXiv:1704.05572.