Clasificando información en sitios de CQA

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 cantidad de información disponible en Web crece día a día. En particular, los servicios de Community Question Answering (CQA) se han convertido en una forma popular de búsqueda de información en línea, donde los usuarios pueden interactuar e intercambiar conocimientos en forma de preguntas y respuestas. La información que contienen los foros de discusión y las CQA son muy valiosas para usuarios con los mismos intereses y más aún, cuando tienen un problema similar, ya que estos sitios son colaborativos, donde se presentan discusiones e intercambio de ideas sobre un tema específico. Esta información merece ser consultada y estructurada. y la recuperación de información (IR) es una tarea esencial para lograr estos objetivos.

Nuestro proyecto se enfoca en extraer y analizar la información que contienen los foros de discusión ya que estos sitios tienen base de conocimiento lo suficientemente completa para ser utilizada. Pero, para que dicha información sea de utilidad, debemos definir estrategias para clasificar las soluciones disponibles, y obtener de ellas las más confiables.

Para ello, nuestro objetivo principal es crear una herramienta que clasifique automáticamente la información que contienen los foros de discusión, utilizando diferentes estrategias de recuperación a nivel de hilo/foro, de clasificación a nivel de post y además, teniendo en cuenta la red de usuarios.

Palabras clave

Recuperación de Información, calidad de datos, Foros de discusión, CQA.

CONTEXTO

Nuestro proyecto se enmarca en el Programa "Desarrollo de Software Basado en reúso – Parte II", de la Universidad Nacional del Comahue, a realizarse en el periodo 2017-2021, el cual extiende al Programa "Desarrollo de Software Basado en reúso" realizado en el período 2013-2016. Dicho Programa está compuesto por tres subproyectos los cuales coinciden en el tratamiento del desarrollo de software basado en reúso, pero desde aspectos diferentes: orientado a dominios, orientado a servicios y orientado a foros de discusión. El proyecto actual, denominado "reúso de Conocimientos en Foros de Discusión – Parte II", continúa la línea de investigación enfocada en la recuperación de información y de conocimiento disponible en foros de discusión técnicos.

1. INTRODUCCIÓN

Ante la necesidad de procesar y reutilizar la información en grandes volúmenes de datos surge, en la década de 1950 [1], la Búsqueda y Recuperación de Información (del inglés Information Search and Retrieval (ISR)). Desde entonces, este campo de investigación fue creciendo logrando grandes contribuciones.

Se destacan dos grandes grupos de investigaciones, por un lado están las investigaciones que se enfocan en la recuperación 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 el proceso comienza con una búsqueda del usuario en el sistema, y las respuestas que retorna poseen diferentes grados de relevancia. En particular, los foros de discusión disponibles en la Web sobre temáticas relacionadas al desarrollo y mantenimiento de software, contienen un amplio conocimiento sobre diferentes problemáticas recurrentes.

Los foros de discusión además son herramientas colaborativas accesibles a todos los usuarios pero no todos permiten que cualquier usuario realice consultas, para ello deben estar registrados en dicho foro (la mayoría de los foros cumplen con estas características). Esto permite la generación constante de información, por lo que hacer un análisis de dicha información es algo deseable y valioso [2].

Una de las características que distingue a los foros de discusión es que su interacción es asincrónica, es decir, que no se necesita estar conectados al mismo tiempo para obtener la información que solucione un problema. Un usuario de la comunidad realiza una pregunta y espera que otro usuario conteste a su consulta, de allí surgen diferentes respuestas que pueden ser de ayuda para el usuario que pregunta o no. Incluso en la gran mayoría de las veces los participantes no se conocen personalmente, pero sí a través de sus nombres, alias o avatares (representaciones gráficas que se asocian a usuarios para identificarse). Un ejemplo de estos foros pueden ser Yahoo Answers! (YA)[1]. YA es un sitio web de preguntas y respuestas impulsado por la comunidad o un mercado de conocimiento de Yahoo!, que permite a sus usuarios tanto formular preguntas cómo responderlas. Para hacerlo, el usuario tiene que tener una cuenta Yahoo![2]. Otro foro a destacar es Stackoverflow[3] el cual está referido específicamente a problemas en entornos informáticos. Ofrece además la posibilidad de agregar código como respuesta a una pregunta, e incluye información de los usuarios, por ejemplo, la reputación, que, cuantos más puntos tenga el usuario, más cosas puede hacer en la comunidad.

Esto implica que al trabajar con foros, como cada uno tiene una estructura diferente, se complejiza la tarea de recopilación de información y del análisis a realizar sobre ella.

De acuerdo al permiso que se le otorga a los participantes dentro de un foro, se pueden distinguir 3 (tres) tipos bien definidos: *los públicos*, donde todos los participantes pueden comunicarse o leer mensajes escritos por el resto sin necesidad de registrarse; los foros *protegidos* donde es necesario registrarse para luego poder enviar mensajes. Por último en los foros *privados* se exigen ciertas restricciones para participar y utilizar la información. Como el proyecto se centra en la información contenida en los foros de discusión, es necesario utilizar foros de discusión *públicos* o *protegidos*, donde se pueda acceder a la información de los mismos sin necesidad de registración.

Un foro de discusión está compuesto por

[1] https://answers.yahoo.com

[2] https://yahoo.com

[3] https://stackoverflow.com/

 RedUNCI - UNdeC. ISBN: 978-987-24611-3-3

múltiples hilos. Los hilos son creados por usuarios que tienen algún problema o duda respecto a un tema. El usuario abre un hilo con una consulta inicial y a partir de ese momento, los usuarios de la comunidad podrán debatir en función del tema y del problema enunciado. En base a este comportamiento es que luego, para obtener conocimiento proveniente de los hilos de discusión se utilizan diferentes técnicas y estrategias para establecer cuáles de las posibles soluciones obtenidas de los foros pueden ser relevantes para consultas sobre problemas similares.

El proyecto realiza, por un lado, el tratamiento del texto contenido en los hilos dentro del foro y por otro lado analiza la red de usuarios que existe en él.

Para ello se ha continuado con el enfoque de Elsas & Carbonelli [6], que fueron unos de los primeros en revisar las estrategias de recuperación de hilo. Utilizan la estructura de los hilos separando la pregunta inicial del resto del hilo, tratándolo como un par <pregunta, hilo>. Cong et al. [5] extraen pares de <pregunta, respuesta> utilizando un enfoque basado en grafos no supervisados. Al igual que Cao et al. [7] se enfocan en extraer contextos y respuestas para preguntas, asumiendo preguntas ya identificadas formando tuplas con formato <pregunta, respuesta, contexto>.
Otra orientación que se ha analizado es la de clasificar o estructurar temas mediante jerarquías, como el enfoque de Nicoletti [17] o el de Helic et al. [8]. O bien el trabajo de Gottipati et al. [15] que aplican técnicas de minería de texto para extraer conocimientos de un foro de discusión mediante la generación de resúmenes basados en temas.

También se ha estudiado la satisfacción del usuario, como lo hace Liu et al. [12] , intenta predecir si el autor de la pregunta estará satisfecho con las respuestas enviadas por los participantes de la comunidad, al igual que Agichtein et al. [13].

Existen además propuestas de generación de algoritmos de ranking basados en la calidad de los atributos, como en la investigación de Kuna et al. [4], o el enfoque que presenta Bathia y Mitra [9] que, a partir del análisis de expertitud de los usuarios, detectan niveles de conocimiento de los comentarios, para clasificar con mayor valoración los hilos en los cuales intervienen personas expertos o con altos conocimientos.
Por otro lado, la investigación de Hecking et at. [10] que combina varias de las técnicas mencionadas anteriormente, ya que analiza la estructura social y semántica de los foros de discusión en cursos MOOC en términos de intercambio de información y roles de usuario.

En base a estos antecedentes, nuestro proyecto tiene como objetivo principal hacer reuso de la información existente en foros de discusión de la Web, haciendo uso no solo de la información textual de los hilos sino que teniendo en cuenta además a los usuarios, utilizando la red que se forma con las distintas interacciones que tienen con el resto de la comunidad y analizando los roles que cumplen, 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. Ya que el análisis del lenguaje natural permite analizar el tipo de fragmento dentro de un hilo de discusión [11]. 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 con resultados favorables en su mayoría, permitiendo la

extensión de algunas líneas de avance y la elaboración de nuevas líneas a favor del objetivo del proyecto.

2. LÍNEAS DE INVESTIGACIÓN 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 2017-2021.

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 frecuentes. 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

Considerando que el objetivo de nuestro proyecto es la realización de de un recomendador de hilos de discusión teniendo en cuenta el ranking de las soluciones favorables, podemos mencionar los resultados obtenidos hasta el momento y los esperados a partir de ellos.

A partir del 2013 se investiga 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, En dicho modelo se determinaron métricas para medir la calidad de información contenida en un hilo y estándares para la calidad de datos software [14], que fueron validadas mediante encuestas [16]. Con el propósito de mejorar los resultados obtenidos en [23] se propone una variación del peso de cada métrica por medio de un sistema de parametrización ad-hoc.

Otra línea de investigación trabaja con el procesamiento del texto de los hilos de discusión, para ello 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, del cual se propone un ranking de soluciones posibles para una pregunta [18].

Para poder manipular la información de foros, se trabajó en el análisis de textos, utilizando la herramienta Lucene[4] [19] con mecanismos personalizados para las *stopwords* (palabras que no aportan significancia) propias del dominio, haciendo tratamientos de recuperación de información para lenguaje específico de Java [20]. Se continuó con la

[4] https://lucene.apache.org

utilización de sinónimos, mediante el uso de la base de datos léxica en inglés WordNet[5] y un analizador morfológico como Stanford POS Tagger[6] [21, 22].

Otra de las líneas de investigación se enfoca en el rol de los usuarios activos de un foro (los que participan compartiendo opiniones y experiencias). Para ello se trabajó con una estrategia para determinar la jerarquía de roles determinados por el nivel de conocimientos de los participantes en los hilos de acuerdo a los posts realizados [24].

Teniendo en cuenta los resultados obtenidos hasta el momento, se continúa trabajando, por un lado analizando la satisfacción del usuario que pregunta a partir de la propuesta de Liu et al. [25] y de Agichtein [13], por otro lado se está trabajando en las respuestas de calidad siguiendo la investigación de Burel et al. [26].

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 [9, 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.

4. FORMACIÓN DE RECURSOS HUMANOS

El proyecto se encuentra conformado por docentes de diferentes áreas debido a su naturaleza multidisciplinaria. En particular en el área de Ingeniería en Sistemas, Programación, y Teoría de la Computación. Las personas que forman parte del proyecto, tanto como colaboradores, asesores o integrantes son:

- Dos docentes investigadores del Departamento de Programación, con dedicación exclusiva, ambos con Doctorado en Informática.
- Un docente investigador del departamento de Programación con beca del CONICET para realización de doctorado.
- Tres docentes con dedicación simple, uno de ellos del Departamento de Ingeniería de Sistemas y dos del Departamento de Programación.
- 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.
- Seis estudiantes de la carrera de Licenciatura en Ciencias de la Computación realizando 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] Singhal,. Modern information retrieval: A brief overview.IEEE Data Eng. Bull., 2001, vol. 24, no 4, p. 35-43

[2] 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, pp. 323332, 2011.

[3] Chen, W., & Persen, R. (2009, June). A Recommender System for Collaborative Knowledge. In AIED (pp. 309-316).

[4] Kuna, H. D., Rey, M., Martini, E., Solonezen, L., & Sueldo, R. (2013). Generación de un algoritmo de Ranking para Documentos Científicos del Área de las Ciencias de la Computación. En el XVIII Congreso Argentino de Ciencias de la Computación.

[5] https://wordnet.princeton.edu/

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

[5] G.Cong, L. Wang, C. Lin, Y. Song, and Yueheng (2008). 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). Association for Computing Machinery, 467–474.

[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] Cao, Y., Yang, W. Y., Lin, C. Y., & Yu, Y. (2011). A structural support vector method for extracting contexts and answers of questions from online forums. Information processing & management, 47(6), 886-898.

[8] Helic, D., & Scerbakov, N. (2003). Reusing discussion forums as learning resources in wbt systems. In IASTED International Conference Computers and Advanced Technology in Education,(Rhodes, Greece) (p. 223).

[9] 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

[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] 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.

[12] Liu, Bing. Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media, 2007.

[13] 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), 1-27.

[14] 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.

[15] Gottipati, S., Shankararaman, V., & Ramesh, R. (2019, October). TopicSummary: A Tool for Analyzing Class Discussion Forums using Topic Based Summarizations. In 2019 IEEE Frontiers in Education Conference (FIE) (pp. 1-9). IEEE.

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

[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] 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.

[19] 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.

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

[21] Zoratto, V., Martínez Carod, N., Otermin, F., & Aranda, G. N. (2017). Análisis de estrategias para clasificar contenidos en foros de discusión. In XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017).

[22] 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. In CICCSI 2018, pp 35-44, 2018.

[23] N. Martínez Carod, V. Zoratto, G. Aranda, P Faraci. Aplicación de Métricas de calidad a hilos de discusión. In CICCSI 2018, pp 151-160, 2018.

[24] N.Martínez Carod, G. Aranda. Valeria Zoratto, Christian Murray. Una propuesta para clasificación de roles de usuarios en foros de discusión técnicos. In CACIC 2019, pp. 836- 845, 2019.

[25] Liu, Y., Bian, J., & Agichtein, E. (2008, July). Predicting information seeker satisfaction in community question answering. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 483-490).

[26] Burel, G., Mulholland, P., & Alani, H. (2016, April). Structural normalisation methods for improving best answer identification in question answering communities. In Proceedings of the 25th International Conference Companion on World Wide Web (pp. 673-678).