Análisis de la información presente en foros de discusión técnicos

Nadina Martínez, Gabriela N. Aranda, Mauro Sagripanti, Pamela Faraci, Alejandra Cechich

Grupo GIISCo, Facultad de Informática, Universidad Nacional del Comahue
Buenos Aires 1400 (8300) Neuquén, Argentina
{nadina.martinez|gabriela.aranda}@fi.uncoma.edu.ar

Resumen Los foros de discusión se han convertido en la herramienta colaborativa más utilizada por los practicantes informáticos para realizar preguntas y recibir propuestas de otros técnicos para solucionar o mejorar problemas técnicos particulares. Con el objetivo de construir un navegador especializado en encontrar dichas soluciones, este artículo introduce un modelo de la información contenida en foros de discusión técnicos y presenta los resultados preliminares de una encuesta realizada a usuarios de dichos foros, enfocada en la percepción de la adecuación de los hilos de discusión a un problema y la correctitud de las soluciones propuestas.

1. Introducción

La incorporación de la tecnología personal ha traído como consecuencia que más personas estén dispuestas a hacer uso de ella, obteniendo directamente beneficios tangibles [1]. Dentro de la comunicación virtual, las herramientas colaborativas como redes sociales, wikis, blogs y foros de discusión permiten el intercambio de información de forma rápida y eficaz, acortando las distancias entre las personas que tienen el conocimiento y aquellas que lo necesitan. Estas herramientas colaborativas no son utilizadas en forma personal solamente, sino también en entornos laborales e incluso académicos. Las herramientas pueden clasificarse en sincrónicas y asíncronas (según se requiera, o no, que las personas estén conectadas al mismo tiempo) [2]. Aunque en los ámbitos laborales y académicos pueden utilizarse ambos tipos de herramientas, se suelen aprovechar las asíncronas para diseminar conocimiento, ya que permiten que la información esté disponible aún cuando las personas que tienen el conocimiento no estén alcanzables.

Aún cuando hay varias herramientas colaborativas asíncronas disponibles en la web, existen algunas diferencias entre ellas. Por ejemplo, los blogs tienen una naturaleza no interactiva, donde una persona que es el dueño (autor del blog), escribe una bitácora o diario en línea, permitiendo que los visitantes participen agregando comentarios, no habiendo comunicación entre los participantes. Otra herramienta asíncrona colaborativa son las Wikis, que son páginas en las que

un autor publica algún tipo de información, a partir de ese momento los otros usuarios que acceden a dicha Wiki pueden modificarla (con la autorización del autor y dependiendo de la privacidad que ofrece el sitio). Por el contrario, los foros de discusión son canales de comunicación cuya finalidad suele ser intercambiar información, experiencias y conocimiento entre sus usuarios. En general son informales y dependen de un moderador (persona que mantiene el orden y naturaleza del foro). Una característica particular de los foros es que cualquier usuario puede comenzar un hilo de discusión quedando establecido un intercambio de información sobre un tema, permaneciendo de esta forma disponible para cualquier lector, por lo que un foro constituye una base de conocimiento al alcance del público en general. Particularmente, se han elegido los foros de discusión como base de este trabajo, dada su capacidad de representar problemas de los usuarios en general (no solo de los dueños de blogs o grupos de colaboradores en wikis), y dado que permite ver todos los comentarios y obtener conclusiones a partir de ellos (a diferencia de las wikis que esta información se mantiene oculta al público en general).

En la actualidad, cuando una persona tiene un problema técnico, ingresa una serie de palabras en algún buscador multipropósito, y éste devuelve una lista de enlaces a páginas Web de distinto formato (manuales, páginas de instituciones técnicas, blogs personales, foros de discusión, etc.) que contienen esas palabras. Luego la persona interesada va observando cada elemento de la lista, y debe visualizar el contenido de cada página para determinar si éste le sirve o no. Esta lista de elementos está ordenada de acuerdo a políticas del buscador, la cual puede no ser precisamente el orden de importancia que el usuario necesitaría. El objetivo futuro al que apunta nuestro trabajo es construir un navegador especializado en problemas técnicos que, a partir de un conjunto de palabras clave que representan la búsqueda inicial, retorne una lista ordenada de soluciones candidatas. Dichas soluciones se obtendrán a partir del análisis previo de varios hilos en foros de discusión técnicos. El orden otorgado a las soluciones candidatas será determinado por medio de un proceso de evaluación de calidad de la información. Con este objetivo en mente, nuestro trabajo actual se ha enfocado en analizar cómo los humanos acostumbran a buscar información en un hilo de un foro de discusión. Para ello, el resto del artículo está organizado de la siguiente manera: primero se introduce un modelo conceptual que representa la información contenida en los foros de discusión. Posteriormente se presenta el cuestionario preparado para recolectar conocimiento tácito de usuarios habituales de foros de discusión técnicos y la manera que ellos seleccionan qué soluciones probar. A continuación se presentan algunos resultados preliminares de la aplicación de dicho cuestionario. Por último se presentan las conclusiones y líneas de trabajo futuro.

2. Modelo conceptual para foros de discusión técnicos

Para definir el punto de vista de nuestro trabajo, se han tomado como base la clasificación de las necesidades de los actores relacionados con un sitio Web [3] y

la clasificación de los usuarios de foros de discusión de Roquet [4], destacando al usuario administrador como aquel usuario con mas privilegios que tiene el control total el foro; que determina quiénes serán los usuarios con roles de moderadores. También son importantes los usuarios moderadores ya que no sólo monitorizan las conversaciones sino aseguran que se cumplan las reglas de convivencia entre el resto de los usuarios. Los usuarios que publican, preguntan y contestan en los foros son los participantes, y por último se llaman participantes externos las personas que sólo pueden leer las conversaciones establecidas en los foros.

Con el objetivo de reutilizar el conocimiento contenido en las conversaciones entre usuarios participantes de una comunidad virtual como es un foro de discusión sobre temas técnicos, la primera instancia es definir un modelo de calidad para la información contenida en dicho tipo de foros. En este sentido, es pertinente que el modelo se plantee considerando sólo el punto de vista del *usuario externo,* es decir, enfocándose en la calidad desde el punto de vista de la información y no de la funcionalidad que el sitio pueda o necesite proveer para el resto de los tipos de usuario.

2.1. Esquema conceptual

Con el fin de establecer un marco teórico para el estudio de la información contenida en los foros de discusión, se realizó una revisión formal de 36 hilos de discusión reales en 6 foros distintos en idioma español e inglés. El resumen de dicho análisis se presentó en [5].

En base a dicho análisis se propuso un primer modelo conceptual de la información disponible en un foro de discusión desde el punto de vista del usuario externo, identificándose las entidades más importantes y sus atributos. A continuación se presenta una actualización de dicho modelo, al cuál se ha agregado el tipo de fragmento de mensaje *figura*. Dicha actualización del modelo surge a partir de la extensión de la revisión, al abarcar foros de discusión que permiten mostrar capturas de pantalla y otros tipos de imágenes en formato gráfico (los cuales no habían sido cubiertos en la primera revisión). Otra mejora al modelo se ha realizado en la definición de los atributos de los usuarios, para los cuales se ha diferenciado su *experiencia*, indicando el reconocimiento de la comunidad del foro respecto al nivel de pericia de dicha persona en el tema de discusión (generalmente expresado en una escala como novato, experto, gurú, etc), y por otro lado la *reputación*, que puede o no estar relacionado a la experiencia, y se suele representar en los foros como agradecimientos, pulgares arriba o abajo, indicadores de "me gusta", etc. Ambos atributos suelen ser expresados verbalmente o con imágenes tipo icono según el foro estudiado.

En la Figura 1 se presenta el modelo conceptual actualizado de acuerdo a los considerandos explicados previamente.

El modelo conceptual puede resumirse de la siguiente manera: un foro de discusión técnico (*foro*) contiene varios hilos de discusión (*hilo*). Cada *hilo* se genera cuando un usuario participante de la comunidad (*usuario*) crea un nuevo tema de debate que surge generalmente a partir de una inquietud personal. Cada *hilo* se identifica por un *título*, que está generalmente relacionado con la

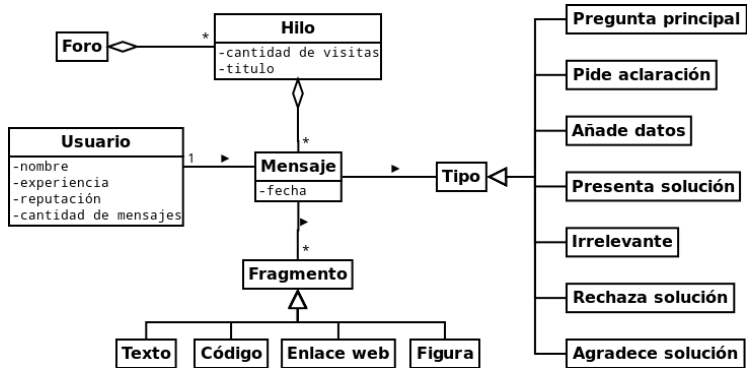

Figura 1. Diagrama de la información contenida en un foro de discusión técnico

pregunta principal, realizada por el usuario que inicia el hilo o tema de debate (esto suele ser un requisito mencionado en las reglas del foro y controlado por los moderadores de los foros). Si bien es cierto que no está presente en todos los foros, suele encontrarse disponible la información relacionada con la *cantidad de visitas* realizadas al *hilo*, es decir la cantidad de veces que la página fue accedida o visitada por un usuario participante o externo.

La estructura del *hilo* está formada por una serie de aportes. Cada aporte, llamado post o mensaje (*mensaje*), es realizado por un *usuario* participante en una *fecha* en particular. A fin de poder analizar el contenido de cada *mensaje*, se ha determinado que un *mensaje* consta de uno o más *fragmentos*, donde cada fragmento puede tratarse de lenguaje natural (*texto*), código que puede ser ejecutado en un sistema operativo o compilado en un lenguaje de programación (*código*), un enlace a una página Web donde una pregunta similar con posibles soluciones afines han sido propuestas (*enlace web*). También puede ser una *figura* (contenido en formato gráfico de tipo jpg, png, etc) que suele utilizarse para incluir esquemas, diagramas, capturas de pantalla, etc.

En base al análisis semántico de los fragmentos que componen el *mensaje*, se definió que existe un mensaje correspondiente a la pregunta principal y otros que cierran o completan la conversación: uno o más mensajes que proponen soluciones, que rechazan alguna solución, y otros que cierran la pregunta de manera positiva (agradeciendo alguna solución). Además de estos mensajes principales, se han reconocido los siguientes: mensajes de pedido de aclaración, de agregado de nuevos datos (que ayudan a los que responden a situarse en el problema), y mensajes irrelevantes (que a veces suelen ser eliminados por el moderador pero otras veces son parte del hilo de la conversación).

Respecto a los usuarios, para cada *mensaje* se puede saber el *usuario* que lo escribió, del cual se conoce su nickname o nombre dentro de la comunidad (*nombre*). Si bien no es un dato presente en todos los foros de discusión, habitualmente se cuenta con más información sobre el usuario como su *experiencia* y *reputación* (que fueron explicados anteriormente) y la *cantidad de mensajes* que ha emitido en la historia de su participación en la comunidad.

3. Encuesta a usuarios de foros de discusión técnicos

Con el objetivo de definir criterios para estimar la calidad de la información contenida en los foros de discusión técnico, se han definido dos características principales. La primera es la *pertinencia* de un hilo de discusión, es decir, el grado de proximidad entre el problema discutido en un hilo de discusión y el problema original definido por un usuario (expresado a partir de una cadena de búsqueda determinada). Y la segunda característica importante es cuán adecuada o *correcta* es una solución para el caso particular del usuario interesado.

En base a estas dos características se plantearon las siguientes preguntas principales para la investigación:

- *¿Cómo determinan los usuarios de los foros de discusión técnicos qué hilos leer (y cuales no)?*
- *¿Cómo seleccionan los usuarios de foros de discusión técnicos las soluciones a probar?*

Para resolver dicha pregunta principal, se propusieron las siguientes sub-preguntas:

- ¿Qué información consideran importante los usuarios para determinar la pertinencia de un hilo de discusión?
- ¿Con qué frecuencia consideran que un ítem de información es importante para determinar la pertinencia de un hilo de discusión?
- ¿Qué información consideran importante los usuarios para determinar la correctitud de una solución propuesta en un hilo de discusión?
- ¿Con qué frecuencia consideran que un ítem de información es importante para determinar la correctitud de una solución propuesta en un hilo de discusión?

3.1. Definición y aplicación del cuestionario

Para responder las preguntas planteadas, se definió un cuestionario destinado a usuarios habituales de foros de discusión de tipo técnico. La estructura del cuestionario fue la siguiente:

Primera Sección: además del nombre y rango de edad de los encuestados, se incluyeron las siguientes preguntas:

- ¿Qué rol o roles relacionados a la informática cumple habitualmente?
- ¿Con qué frecuencia accede a foros de discusión técnicos?
- ¿A qué temáticas se refieren los foros de discusión técnicos que visita?

Segunda Sección: en esta etapa se solicitó elegir una opción en la escala [Siempre, Casi siempre, A veces, Casi nunca, Nunca], para once afirmaciones que marcan la importancia de los ítems de información definidos en el modelo conceptual. Por cuestión de espacio, a continuación se exponen como ejemplo las dos primeras:

- Si el título del hilo tiene todas las palabras claves ingresadas, alcanza para saber si el tema en cuestión está relacionado con mi búsqueda.
- Si la pregunta principal (primer post) está relacionada en parte con lo que estoy buscando continúo leyendo el resto del hilo.

Tercera Sección: en esta sección se solicitó que los encuestados seleccionen uno o más ítems de información definidos en el modelo conceptual (título, mensaje principal, fecha del mensaje, etc), relacionados a las siguientes consignas:

- Para estimar si un hilo de un foro de discusión está relacionado con mi problema (es pertinente), observo...
- Suponiendo que se trata de un hilo que es pertinente para su problema, para estimar si una solución propuesta es correcta (correctitud), la información que observo es...

Finalmente, en la última sección se dejó espacio disponible para que los encuestados incluyeran comentarios o sugerencias de distinto tipo.

El cuestionario fue implementado mediante un formulario (*form*) en la plataforma de Google Drive[1]. Además, en el sitio web del proyecto[2] se publicó una página donde se explica el objetivo de la encuesta y se describen las secciones y pautas establecidas para el cuestionario. El enlace a dicha página se envió por correo electrónico a un conjunto de 40 personas que cumplen distintos roles relacionados a la informática en el ámbito de la Universidad Nacional del Comahue (docentes, estudiantes y graduados desempeñándose en el ámbito laboral local). Al momento de la redacción de este artículo se cuenta con 24 respuestas, cuyos resultados serán presentados en la siguiente sección.

3.2. Resultados preliminares

Perfil de los encuestados. A partir del análisis de la primera sección del cuestionario, se puede definir si el perfil de los encuestados abarca un amplio registro de tipos de usuarios de foros de discusión.

En primer lugar se observa que la mayoría de los encuestados son usuarios habituales de foros de discusión técnicos. De acuerdo a la Figura 2, el 75 % de ellos accede varias veces a la semana (58 %) e incluso algunos diariamente (17 %). Esto es importante para este estudio, dado que permite confiar en el conocimiento previo de los encuestados al momento de responder las preguntas planteadas.

Respecto a los roles que estos encuestados cumplen, puede observarse en la Figura 3 que se han visto representados todos los roles excepto el de tester, lo cual deberá ser tenido en cuenta al extender la muestra de usuarios en el futuro. Los porcentajes más altos corresponden a los roles de programador (20 %) y docente (15 %). Sólo dos encuestados informaron un rol distinto a la lista inicial: uno es

[1] https://drive.google.com/

[2] http://forumadvisor.wordpress.com/encuesta-calidad-de-la-informacion-en-foros-de-discusion-tecnicos/

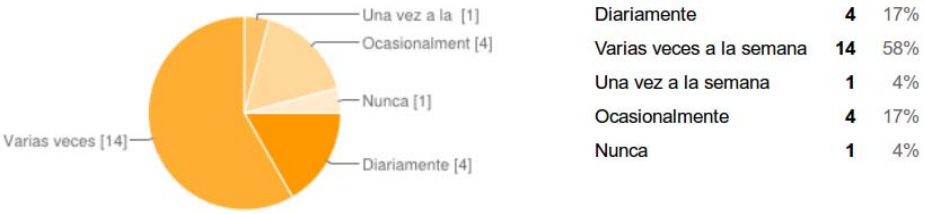

Figura 2. Frecuencia de acceso a foros de discusión

"Administrador de sistemas y redes" y el otro "Administrador de infraestructura de sistemas". Las tareas de dichos roles deberán ser analizadas y considerar la inclusión de ambos roles en una nueva aplicación del cuestionario.

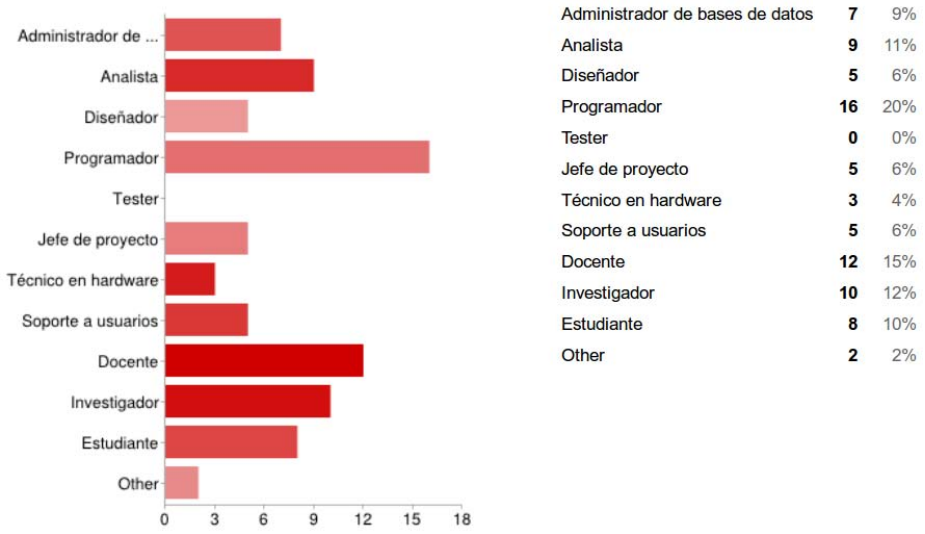

Figura 3. Roles de los encuestados relacionados a la informática

Analizados los tipos de foros visitados, puede observarse en la Figura 4 que todos los tipos propuestos han sido abarcados por los encuestados. Los porcentajes más altos corresponden a los foros sobre lenguajes de programación (31 %) y herramientas de software específicas (24 %). De las restantes, el rubro de foros sobre desarrollo de aplicaciones Web es el más utilizado (14 %). Otra vez, dos tipos de foro fueron agregados por los encuestados en la opción "Otros" y corresponden a los mismos usuarios que agregaron un nuevo rol a la lista. Los tipos de foros agregados son "Administración de sistemas" y "Active Directory". Dichos tipos de foros deberán ser analizados para considerar su inclusión en una nueva aplicación del cuestionario. Además, en el futuro planeamos extender el análisis de la información recolectada para comprobar estadísticamente la correlación entre los roles de los encuestados y los tipos de foros visitados por ellos.

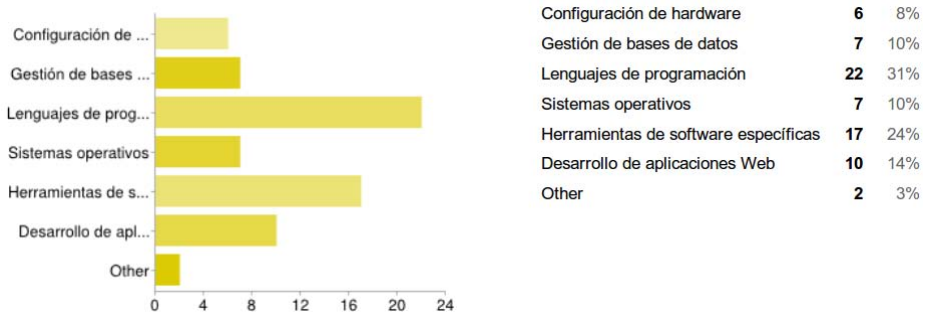

Figura 4. Temática de los foros visitados por los encuestados

Análisis de la pertinencia de la consulta. A continuación se analiza la información de la primera pregunta de la tercera sección del cuestionario.

Esta pregunta pedía a los encuestados que seleccionaran aquellos ítems (uno o más) que chequean para definir si un hilo de discusión está relacionado con su problema particular. Los resultados obtenidos se muestran en la Figura 5.

Figura 5. Información analizada para definir la pertinencia

Los ítems más seleccionados son el título (19 %), la pregunta principal (21 %) y los mensajes que proponen las soluciones (17 %). Al contrario de nuestra expectativa previa, 3 encuestados no seleccionaron la pregunta principal como lo más importante para definir si el hilo era pertinente o no, e incluso el título fue elegido por más personas (22) que la pregunta principal (21). Sin embargo, a partir de análisis de cada caso, debe destacarse que todos los encuestados seleccionaron el título o la pregunta principal como importantes para definir la pertinencia del hilo. Algo que también debe destacarse es el bajo porcentaje otorgado a la experiencia previa y la reputación de los usuarios que proponen la

solución, siendo apenas un 4 % o 5 %, mientras que los mensajes que agradecen una solución son utilizados como referencia más a menudo (9 %).

Discusión. En base al análisis de un conjunto preliminar de encuestas, se puede resumir que el perfil de los usuarios encuestados abarca un amplio espectro de roles relacionados a la informática. Ellos, a su vez, visitan habitualmente varios tipos de foros de discusión técnicos. Esta característica es importante para nuestro trabajo, dado que permitirá obtener una visión amplia del comportamiento de técnicos que buscan soluciones en foros de discusión. Respecto a los ítems de información que los encuestados indican importantes para identificar la pertinencia de un hilo en cuestión, es importante destacar que el título y la pregunta principal son los dos ítems más mencionados, y que, aunque la expectativa previa era que todos respondieran que la pregunta principal era siempre la más utilizada, 3 personas (4 %) no lo perciben así. Otro resultado interesante es que muy pocos encuestados (4 %) respondieron que consideran importante la experiencia o reputación de quienes proponen las soluciones. Estas tendencias deberán ser tenidas en cuenta al avanzar en la extensión del estudio.

4. Trabajos relacionados

La propuesta de Tigelaar et al [6] se enfoca en simplificar el contenido de los hilos de discusión extensos, resumiéndolos automáticamente con un prototipo de implementación basado en lenguaje natural. Este trabajo es un gran aporte para el análisis de los hilos de ejecución, pero no se enfoca en determinar si el conocimiento será de interés para la persona que lo consulta o no, como es el interés de nuestra propuesta.

En cuanto a reuso de conocimiento en foros de discusión, Chen et al [7] proponen un sistema recomendador para conocimiento desarrollado de manera colaborativa, analizando automáticamente los mensajes de un foro de discusión de un curso de Inteligencia Artificial para proponer mensajes con contenido similar, escritos por estudiantes de dictados anteriores del mismo curso. Otra propuesta existente es la de Helic y Scerbakov [8], que presenta un método de clasificación de los mensajes de un foro de discusión de acuerdo a una jerarquía de temas preestablecida. En primer lugar, nuestro enfoque se diferencia de las propuestas anteriores porque ambas están desarrolladas para dominios de aprendizaje colaborativo (e-learning), mientras que nuestro recomendador apunta a un dominio más amplio, que involucra usuarios con distinto conocimiento previo (background). Además, en dichos trabajos el foro utilizado es único, lo que permite asegurar que la información a analizar se encuentra en un formato estándar y que cualquier modificación puede ser prevista y gestionada a priori. Por el contrario, nuestra propuesta apunta a recolectar información de distintos foros, por lo tanto la heterogeneidad de formatos de la información a capturar y la posibilidad de cambios no programados es un desafío extra.

5. Conclusiones y trabajo futuro

En este trabajo se ha presentado una mejora del modelo conceptual para la información contenida en foros de discusión técnicos presentada en [5]. Luego, se ha introducido un cuestionario para usuarios habituales de foros de discusión técnicos y se han presentado los resultados preliminares sobre el perfil de los encuestados y el análisis de una pregunta del cuestionario, que considera la pertinencia de un hilo de discusión en relación con un problema particular. Como trabajo a futuro se planea avanzar en el análisis semántico de los mensajes en los foros, así como en establecer una serie de métricas e indicadores de calidad que sirvan para la detección automática de soluciones a problemas técnicos.

El objetivo a futuro es trabajar en el desarrollo de un buscador especializado en soluciones a problemas técnicos. Dicho buscador está previsto que mantenga una base de datos de las experiencias de los usuarios (después de seleccionar y aplicar las soluciones candidatas), como un mecanismo de mejora constante a partir de la retroalimentación realizada por los mismos usuarios.

Agradecimientos

Este trabajo está parcialmente soportado por el subproyecto "Reuso de conocimiento en foros de discusión técnicos", correspondiente al Programa de Investigación 04/F001 "Desarrollo orientado a reuso", de la Universidad Nacional del Comahue, y por el Proyecto PICT-2012-0045 "Mecanismos de soporte para grids híbridos orientados a servicios y técnicas de desarrollo de aplicaciones".

Referencias

1. S. Poltrock and J. Grudin, "CSCW, groupware and workflow: experiences, state of art, and future trends," in *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '99, (New York, NY, USA), pp. 120–121, ACM, 1999.
2. C. A. Ellis, S. J. Gibbs, and G. L. Rein, "Groupware: Some issues and experiences," *Communications of ACM*, vol. 34, no. 1, pp. 38–58, 1991.
3. L. Mich, M. Franch, and G. Cilione, "The 2QCV3Q quality model for the analysis of web site requirements," *Journal of Web Engineering*, vol. 2, pp. 105–127, Sept. 2003.
4. G. Roquet García, "Los foros de discusión en educación," *Siglo XXI: Perspectiva de la Educación desde América Latina*, no. 4, pp. 69–78, 1998.
5. G. Aranda, N. Martinez, P. Faraci, and A. Cechich, "Hacia un framework de evaluación de calidad de información en foros de discusión técnicos," in *ASSE 2013-Simposio Argentino de Ingeniería de Software, JAIIO 42Âº-Jornadas Argentinas de Informática*, (Córdoba, Argentina), p. a publicarse, SADIO, 2013.
6. A. S. Tigelaar, R. Op Den Akker, and D. Hiemstra, "Automatic summarisation of discussion fora," *Natural Language Engineering*, vol. 16, pp. 161–192, 4 2010.
7. W. Chen and R. Persen, "A recommender system for collaborative knowledge," in *2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, (Amsterdam, The Netherlands, The Netherlands), pp. 309–316, IOS Press, 2009.
8. D.Helic and N. Scerbakov, "Reusing discussion forums as learning resources in wbt systems," in *IASTED International Conference Computers and Advanced Technology in Education*, (Rhodes, Greece), pp. 223 – 228, 2003.