# Una herramienta para el análisis de hilos de discusión técnicos

Gabriela N. Aranda, Nadina Martínez, Sandra Roger, Pamela Faraci, Alejandra Cechich

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

**Resumen** La información existente en la web puede ser reutilizada para la adquisición de nuevos conocimientos. Dentro de las herramientas colaborativas disponibles, los foros de discusión son ampliamente utilizados para plantear un problema y recabar experiencias que además pueden ser consultadas por otros usuarios en caso de problemas recurrentes. La herramienta que se propone tiene como fin capturar, mantener y analizar hilos de discusión existentes en foros técnicos para, dado un problema particular, sugerir un conjunto de soluciones exitosas. En este trabajo se presenta una herramienta para el procesamiento automático de hilos de foros técnicos y se muestra un caso de estudio sobre un problema específico.

## 1. Introducción

Durante las últimas décadas, la Web ha evolucionado convirtiéndose en un medio indispensable para la interacción y colaboración entre los miembros de comunidades virtuales. Nuevas plataformas han aparecido, como las wikis y los weblogs, y otras, como los foros de discusión que ya existían como herramientas colaborativas en la Web, han crecido y se han consolidado en su rol [1]. En especial, en este trabajo nos enfocamos en los foros de discusión técnicos, donde un usuario inicia un debate con una pregunta que representa su duda o problema y los miembros de la comunidad responden proponiéndole soluciones, de acuerdo a sus conocimientos y experiencias previas. Si bien las soluciones propuestas atienden a la pregunta que inició la discusión, el conjunto de mensajes queda disponible al público en general, y las soluciones propuestas pueden ser reutilizadas por otras personas que tengan inconvenientes similares. Para acceder a esta información, los usuarios utilizan motores de búsqueda multipropósito y suelen recorrer varias páginas buscando un problema similar al suyo, para luego ver si existe alguna solución propuesta. En general este proceso puede llevar al usuario a visitar distintas páginas antes de conseguir una que le proponga una solución válida, y a veces es necesario probar varias hasta lograr una solución exitosa. Nuestra propuesta es realizar un pre-procesamiento de la pregunta del usuario y, basado en el análisis previo de hilos de discusión clasificados por tema, ofrecerle un conjunto de soluciones con mayor probabilidad de éxito.

En la Figura 1 se muestra el proceso propuesto para seleccionar un conjunto de soluciones candidatas dada una pregunta ingresada por un usuario. Ante dicha pregunta, la primera tarea será identificar el problema al cual se refiere (*Reconocimiento del Problema*). En principio, toda pregunta estaría comprendida dentro de un tema en particular, por lo cual debería ser posible determinar una clasificación de temas, identificando y agrupando preguntas similares. Para lograr dicha clasificación se propone analizar los problemas recurrentes y crear una taxonomía que sirva como referencia para organizar los hilos de discusión recuperados de la web. Por medio de un agente inteligente, se obtendrá automáticamente la información desde un grupo de foros de la Web (a medida que la misma se encuentre disponible), y se los mantendrá de manera organizada en un repositorio (base de datos) al que pueda accederse más rápidamente. La definición de un cuerpo de información que contenga descriptores relacionados para cada problema de una taxonomía, se convierte así, en el eje principal para la clasificación, tanto de los hilos de discusión como de la pregunta del usuario. Sin embargo, los descriptores de cada tema no suelen funcionar de manera aislada sino que, es la aparición de ciertas combinaciones de los mismos lo que define la pertenencia de un hilo a un tema dentro de la taxonomía.

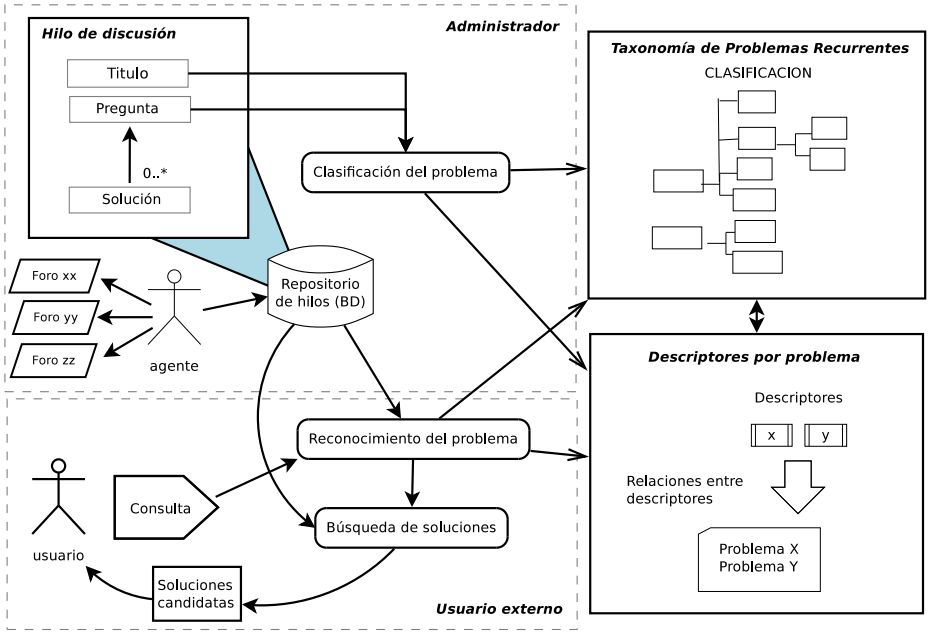


Figura 1: Proceso de selección de soluciones para un problema particular

El resto del artículo está organizado de la siguiente manera: Primero se introduce una herramienta que mantiene y gestiona la información contenida en foros de discusión técnicos como base para la gestión del conocimiento disponible en ellos. Luego se presenta un caso de estudio que ilustra la motivación de nuestro trabajo. Por último se presentan las conclusiones y líneas de trabajo futuro.

## 2. La herramienta propuesta

La arquitectura de nuestra herramienta está planteada en tres capas, siendo el principal objetivo la separación de la lógica del negocio respecto del diseño. En la Figura 2 se presenta el esquema general de dicha arquitectura y a continuación se mencionan los detalles más importantes de cada capa.

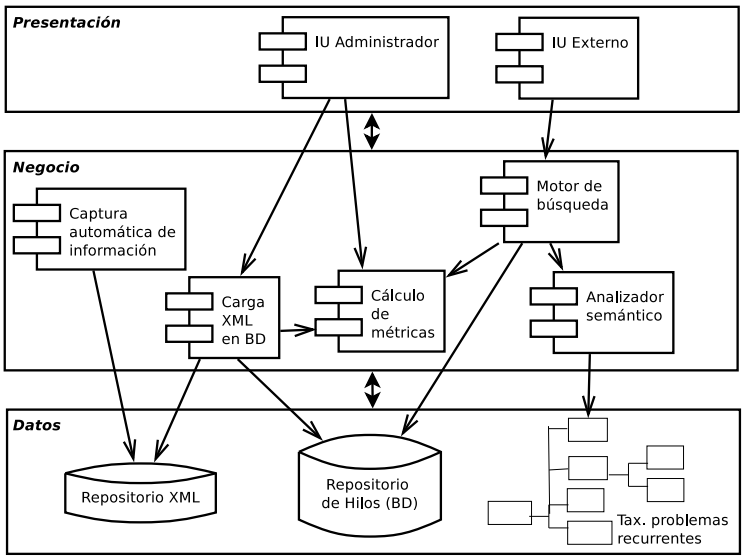


Figura 2: Arquitectura de la herramienta propuesta

### 2.1. Capa de datos

Esta capa permite el almacenamiento de tres cuerpos de información:

- *Repositorio de archivos XML*: En este repositorio se mantienen los archivos XML con el contenido de los hilos de discusión al momento de ser capturados desde la Web. Una vez procesados y cargados en la base de datos, se mantienen (en un lugar separado) con fines estadísticos y de control.
- *Taxonomía de problemas recurrentes*: Esta taxonomía mantiene la clasificación de problemas que ocurren frecuentemente, así como un conjunto de descriptores que lo representan y las relaciones entre los mismos.
- *Base de datos*: Para reutilizar el conocimiento contenido en las conversaciones de los foros de discusión técnicos, el primer paso fue la definición de un modelo de datos. En [2] se presentó el análisis de 36 hilos de discusión reales en 6 foros (en español e inglés), que se tomó como base para proponer un modelo conceptual de la información contenida en dichos foros desde el punto de vista del usuario externo, identificándose las entidades más importantes y sus atributos. Posteriormente, el modelo se extendió agregando el tipo de fragmento llamado *contenido multimedia* para considerar la aparición de figuras, audio y vídeo que no se tuvieron en cuenta en una primera instancia. En la Figura 3 se presenta el modelo conceptual

actualizado de acuerdo a los considerandos explicados previamente. Más información sobre el mismo puede encontrarse en [3,2]. Dicho modelo se plasmó en un conjunto de tablas que forman la base de datos de la herramienta propuesta. Para su implementación se utilizó el sistema gestor de bases de datos MySQL, mediante el motor de almacenamiento InnoDB[1].

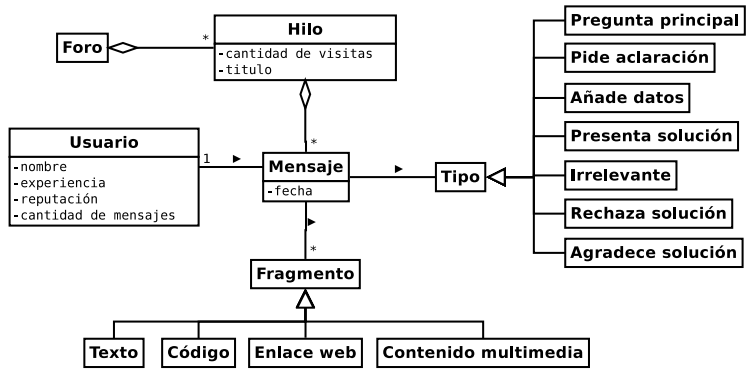


Figura 3: Modelo de la información contenida en foros de discusión técnicos

### 2.2. Capa de negocio

La capa de negocio de la aplicación incluye los siguientes componentes:

- *Captura automática de información*: Tiene como objetivo la búsqueda y descarga del contenido de hilos de discusión disponibles en la web para generar archivos XML de acuerdo al modelo conceptual previsto. Dado que cada foro puede tener un formato y un conjunto de información disponible distinta, se ha enfocado la tarea sobre un conjunto inicial de foros, que se irá ampliando paulatinamente. En la actualidad, este proceso se lleva a cabo de manera semi-automática, utilizando funciones provistas por la herramienta Google Drive Spreadsheets y el lenguaje Xpath para identificar las piezas de información dentro de los documentos HTML, mientras que el proceso de generación del documento XML asociado se completa manualmente agregando etiquetas de algunos atributos puntuales. Se encuentra en desarrollo una componente para automatizar en la totalidad dicho proceso de captura.
- *Carga de XML en la base de datos*: Este componente se encarga, a pedido de un usuario administrador que cuenta con una interfaz para tal fin, de recuperar uno o más archivos del repositorio de archivos XML y procesarlos para actualizar la base de datos (BD). En cuanto al funcionamiento sincronizado entre la componente anterior y ésta, se ha documentado una gramática que define exactamente la estructura esperada, etiquetas y atributos que deben contener los archivos XML a procesar.

[1] http://dev.mysql.com/doc/refman/5.0/es/innodb.html

- *Cálculo de métricas*: Analiza la base de datos para calcular métricas sobre uno o más hilos de discusión. Algunas de ellas pueden ser procesadas al momento de la carga del archivo XML en la base de datos, pero otras requieren (y dependen) de la cadena de búsqueda que será recibida desde el motor de búsqueda.
- *Analizador semántico:* Analiza la semántica de los fragmentos tipo *texto*. Por un lado es necesario que determine el tipo de mensaje, tal como fue presentado en el modelo de datos, para identificar si se trata de una propuesta de solución, aceptación, rechazo, etc. (Figura 3). Por otro lado, debe identificar el tipo de problema que se está discutiendo en el hilo. Para esta tarea de clasificación se utilizará una taxonomía de problemas recurrentes y un conjunto de patrones de correlación entre descriptores de cada tema, como fue presentado en la Sección 1.
- *Motor de búsqueda*: Tiene como objetivo la funcionalidad visible al usuario externo de la herramienta. Por un lado interactúa con la capa de presentación, solicitando al usuario que ingrese una cadena de búsqueda que represente su problema y devuelve una lista rankeada de soluciones candidatas. Internamente, requiere de los servicios prestados por los componentes *Cálculo de métricas* y *Analizador semántico*, para determinar qué soluciones devolver y en qué orden debe presentarlas.

### 2.3. Capa de presentación

La capa de presentación incluye dos componentes:

- *Interfaz de Usuario Externo*: Provee la interfaz Web para que el motor de búsqueda sea utilizado por usuarios interesados en encontrar soluciones a un problema técnico particular.
- *Interfaz de Usuario Administrador*: Provee la interfaz para que el o los usuarios de tipo administrador realicen tareas de mantenimiento, carga y actualización de la base de datos, así como la definición de nuevas métricas.

## 3. Caso de estudio

A continuación se mostrará, mediante un caso de estudio, la complejidad de la lógica asociada al componente *Analizador Semántico* de la herramienta propuesta.

### 3.1. Diseño del caso de estudio

Como base de este caso de estudio se tomó un problema recurrente entre los programadores del lenguaje Java, que es la separación de un texto en palabras aisladas (*tokens*), suponiendo que dicho texto se mantiene en una variable de tipo String. Para llevar a cabo el estudio, se utilizó como referencia el foro *Stack Overflow*[2], ampliamente utilizado por la comunidad de programadores. En dicho foro se buscó la cadena "*how to separate words from String in Java*", obteniéndose como resultado 288 hilos relacionados. La clasificación de los hilos se realizó

[2] http://stackoverflow.com/

manualmente. La mayor dificultad radicó en que la información disponible está expresada en lenguaje natural, por lo que la deducción no es directa y requiere cierto nivel semántico. Al ser una tarea de naturaleza no trivial, este estudio se restringió a los primeros 30 hilos recuperados. A fin de conseguir una clasificación más objetiva, se requirió el análisis consensuado de tres expertos. En la Tabla 1 se presentan las características básicas del caso de estudio realizado.

Tabla 1: Características del caso de estudio

| | |
|---|---|
| **Foro analizado** | Stack Overflow |
| **Idioma** | Inglés |
| **URL del foro** | http://stackoverflow.com/ |
| **Fecha de recuperación** | 23/07/2014 |
| **Cadena de búsqueda** | "how to separate words from String in Java" |
| **Hilos recuperados** | 288 |
| **Hilos analizados en el caso de estudio** | Primeros 30 resultados |

Durante el análisis de los hilos, se detectó que muchos de ellos no estaban directamente relacionados con la cadena de búsqueda ingresada, y que no ofrecían soluciones relevantes. Para clasificar la calidad de los resultados devueltos se propuso estudiar por separado los siguientes aspectos:

1. Si la pregunta principal del hilo está relacionada a la cadena de búsqueda.
2. Si alguna solución a la cadena de búsqueda se encuentra dentro del hilo

Para medir el grado de relación entre los puntos anteriores en cada hilo, se determinó la siguiente escala: *Total* (T) / *Parcial* (P) / *Ninguna* (N). Por ejemplo, la pregunta principal del hilo puede estar total/parcial/no relacionada a la cadena de búsqueda. Las soluciones en el hilo fueron analizadas de manera análoga. Además, se otorgó un peso a cada etiqueta (valor numérico entre 0 y 8), considerando en primer lugar la relación con las soluciones propuestas, es decir, se ponderó con un valor más alto al hilo que presenta soluciones adecuadas para la cadena de búsqueda, aún cuando la pregunta original del hilo no esté completamente relacionada a la misma. La Tabla 2 muestra las categorías posibles, el valor en la escala propuesta y la cantidad de hilos asignados a cada categoría.

Respecto a los hilos de la categoría TRTR (8), cabe mencionar que el nivel de similitud no significa una coincidencia total entre la cadena de búsqueda y la pregunta principal del hilo. En la Tabla 3, se presentan las preguntas principales de dichos hilos. Por ejemplo, en el hilo 4 la pregunta principal refiere a cadenas que son recuperadas desde un archivo de texto, que el usuario desea almacenar separadamente, cada palabra en una celda de un arreglo. En este caso, parte de la tarea deseada es separar las palabras contenidas en un String, aunque también realiza otras tareas que no están estipuladas en la cadena de búsqueda. En el hilo 13, la pregunta no menciona la separación de las palabras propiamente dicha, pero es una tarea necesaria para recuperar el valor numérico que se desea almacenar por separado. En el hilo 15, la restricción está dada por el tipo de separación entre palabras, al tratarse de un archivo de entrada de tipo CSV. De manera análoga se definió el tipo de similitud con las soluciones propuestas.

Tabla 2: Clasificación de los hilos según la relación con la cadena buscada

| Categoría | Relación con la cadena de búsqueda | Valor en escala | Cant. hilos |
|---|---|---|---|
| NRNR | Pregunta y soluciones no relacionadas | 0 | 10 |
| PRNR | Pregunta parcialmente relacionada y soluciones no relacionadas | 1 | 1 |
| TRNR | Pregunta totalmente relacionada y soluciones no relacionadas | 2 | 1 |
| NRPR | Pregunta no relacionada y soluciones parcialmente relacionadas | 3 | 2 |
| PRPR | Pregunta y soluciones parcialmente relacionadas | 4 | 6 |
| TRPR | Pregunta totalmente relacionada y soluciones parcialmente relacionadas | 5 | 1 |
| NRTR | Pregunta no relacionada y soluciones totalmente relacionadas | 6 | 0 |
| PRTR | Pregunta parcialmente relacionada y soluciones no relacionadas | 7 | 6 |
| TRTR | Pregunta y soluciones totalmente relacionadas | 8 | 3 |

Tabla 3: Hilos de la categoría TRTR para la cadena "*How to separate words from String in Java*"

| Hilo | Pregunta principal del hilo |
|---|---|
| 4 | I am trying to take in words from a file input that only contains strings and store each word separately into a single array ( `ArrayList` is not allowed).<br><br>My code below takes in the file input, but takes it in as one chunk. For example, if the file input was "ONE TWO THREE" I want each word to have its own index in the array ( `array[0] = "ONE"` , `array[1]="TWO"` and `array[2]="THREE"` ) but my code below just takes the sentence and puts it all in `array[0] = "ONE TWO THREE"` . How can I fix this? |
| 13 | I need to store users input from command prompt, for eg, if user types "hello 23"<br><br>I need to check "hello" and to get 23 to be stored in separate Integer |
| 15 | How to a match word in comma separated values (CSV) string with a single space after each comma.<br><br>Let's say:<br><br>`String = 'abc, def, ghijk, l, mn, opqr, stu';`<br><br>What would be the regex to match a complete word in the above string?<br><br>Edit: lets say i want to match ghijk from the given string. |

### 3.2. Análisis de los resultados

La evaluación se realizó utilizando como medida la *precisión*, definida como *"el porcentaje de ítems relevantes en el conjunto retornado"* [4]. Para realizar una mejor evaluación la precisión se calculó considerando los siguientes valores de corte (*cutoff*): 5, 10, 20 y 30; donde un valor de corte X indica la precisión obtenida dentro de los primeros X hilos del ranking.

La Tabla 4a presenta los valores de precisión para los hilos de categoría TRTR (valor 8 en la escala propuesta). Como se puede apreciar, sólo el 20 % de los 5 primeros hilos recuperados (1 hilo) presenta una solución relevante para la cadena de búsqueda. Este valor estaría indicando una muy baja precisión en el nivel de respuestas correctas. Incluso se debe recordar que, como se refirió anteriormente, dicho hilo no cumplía con el 100 % de exactitud en la similitud entre su pregunta

principal y la cadena de búsqueda. Este bajo nivel de respuestas correctas se decrementa más aún al considerar los 10 primeros hilos recuperados, dado que sólo 1 de ellos es relevante. Por el contrario, al tomar en cuenta los primeros 20 hilos recuperados, se puede apreciar un leve aumento en la precisión (15 %), esto se debe a que en las posiciones 13 y 15 se ubican los otros dos hilos que pertenecen a la categoría TRTR.

Al relajar el nivel de exactitud requerido entre la cadena de búsqueda y la pregunta principal (Tabla 4b), considerando como relevantes los hilos en las categorías TRTR (8) y PRTR (7), se observa que no existe variación con respecto al mismo corte en la Tabla 4a para las primeras 5 soluciones, pero sí se observa una mayor precisión al aumentar el valor de corte.

Finalmente, en la Tabla 4c se puede observar un alto grado de hilos no relevantes recuperados en todos los cortes. Estos son los hilos que no tienen ningún tipo de relación entre la cadena de búsqueda con la pregunta principal ni con las soluciones propuestas, y que representan casi el 50 % de los hilos recuperados.

Tabla 4: Precisión del ranking de resultados obtenido

(a) Cat. TRTR

| Corte | Precisión |
|---|---|
| 5 | 20 % |
| 10 | 10 % |
| 20 | 15 % |
| 30 | 10 % |

(b) Cat. TRTR y PRTR

| Corte | Precisión |
|---|---|
| 5 | 20 % |
| 10 | 20 % |
| 20 | 25 % |
| 30 | 30 % |

(c) Cat. NRNR

| Corte | Precisión |
|---|---|
| 5 | 40 % |
| 10 | 40 % |
| 20 | 50 % |
| 30 | 40 % |

A partir de la clasificación propuesta se estableció un orden deseado para los resultados, en donde los 3 hilos correspondientes a la categoría TRTR (valor 8) se encontrarían en los primeros 3 lugares, a continuación los hilos de la categoría PRTR (valor 7), y así sucesivamente ubicando los 10 hilos correspondientes a la categoría NRNR (valor 0) en las últimas posiciones. La Figura 4 muestra la variación en la ubicación en el ranking obtenido con respecto al deseado, en donde se observa una variación de hasta 20 posiciones de diferencia. Por ejemplo, en el caso del hilo 1 devuelto por el buscador, al estar en la categoría NRNR, bajaría a la posición 21. De manera similar, el hilo 29 en el *ranking* actual subiría 20 posiciones al lugar 9 en el *ranking* deseado.

### 3.3. Discusión

Buscar una solución a un problema dado en un foro de discusión puede resultar una tarea con cierto grado de dificultad. Como se ha observado en la sección anterior, los algoritmos de ponderación de los foros devolverían hilos con un bajo nivel de precisión. A partir del análisis realizado en la sección anterior, se observa que:

- La precisión con valor de corte 5, arrojó el 20 % de resultados relevantes (categoría 8). Esta precisión no aumentó aún cuando se relajó la restricción sobre la exactitud de la pregunta principal (agregando la categoría 7).

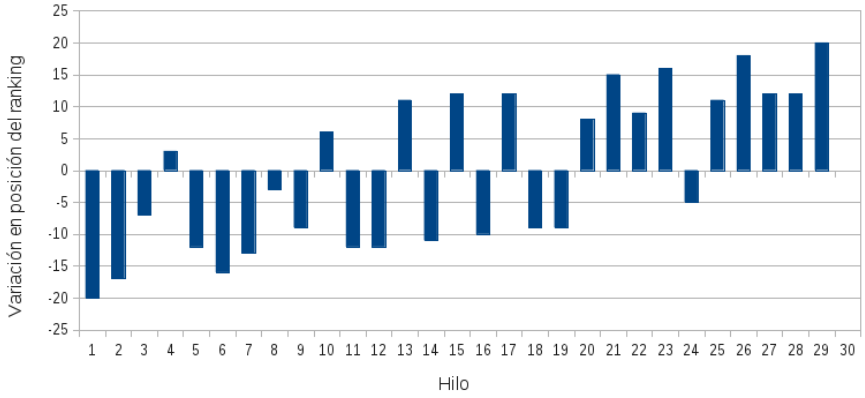


Figura 4: Comparación entre el orden obtenido y el deseado

- El porcentaje de hilos totalmente irrelevantes (categoría 0) es del 40 % o más independientemente del corte, lo cual representa un alto porcentaje de información recuperada que no es útil para el usuario.
- La variación entre la posición en el ranking obtenido y el ranking deseado demostró ser muy alta en algunos casos.

Una posible respuesta a estos resultados es que los algoritmos contemplan sólo las palabras de la cadena de búsqueda de manera aislada y con pocos recursos semánticos que ayuden a relajar el significado de las mismas (por ej.: sinónimos, relaciones semánticas, recursos específicos del dominio de la pregunta, etc.).

## 4. Trabajos relacionados

En cuanto a reuso en foros de discusión, existen varias propuestas: Chen et al [5] plantea un sistema para proponer mensajes con contenido similar, en un entorno de un curso universitario. Por otro lado, Helic & Scerbakov [6], clasifican los mensajes de un foro de acuerdo a una jerarquía de temas preestablecida. Ambas propuestas se enfocan en un único foro para aprendizaje colaborativo (e-learning), mientras que nuestro recomendador apunta a usuarios de foros técnicos en general. Además, a futuro, 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.

El trabajo de Tigelaar et al [7] se enfoca en simplificar el contenido de los hilos de discusión extensos, pero, a diferencia de nuestra propuesta, no intenta determinar si la información devuelta es de interés para quien realiza la consulta.

Finalmente, la propuesta de Kuna et al [8] se enfoca en obtener un *ranking* de publicaciones científicas a partir de buscadores de recursos académicos, evaluando la calidad de los mismos respecto a la fecha, tipo y lugar de publicación, así como a la trayectoria de sus autores. Se relaciona con nuestro trabajo en la utilización de algoritmos de ponderación aunque la naturaleza de la información recuperada (documentos científicos vs. hilos) es distinta.

## 5. Conclusiones y trabajo futuro

En este trabajo se han introducido los lineamientos generales de una herramienta, que se encuentra en desarrollo, para recuperar y analizar la información almacenada en foros de discusión técnicos. También se ha presentado un caso de estudio donde se analizó la relación entre una cadena de búsqueda con los mensajes recuperados de un foro específico, observándose que el orden de relevancia obtenido presenta un bajo grado de precisión en los hilos relevantes sumado a un alto grado de hilos irrelevantes. La selección de un único foro se debe a que los algoritmos para ponderar las respuestas pueden diferir de acuerdo al foro analizado, por lo que a futuro se planea repetir este experimento en otros foros y realizar un estudio comparativo de los resultados que se obtengan en cada caso. Además, se planea avanzar en el estudio y uso de técnicas para el análisis semántico de los mensajes de los foros. Complementariamente, se trabajará en un conjunto de métricas e indicadores de calidad que sirvan para determinar la correlación entre una pregunta que enuncia un problema y las soluciones diseminadas en distintos foros.

## Agradecimientos

Este trabajo está parcialmente soportado por el sub-proyecto "*Reuso de conocimiento en foros de discusión técnicos*" del Programa de Investigación 04/F001 "Desarrollo orientado a reuso" de la Universidad Nacional del Comahue.

## Referencias

1. J. Dorn, *Social Software (and Web 2.0)*, pp. 305–311. PA: Information Science Reference, 2010.
2. 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 42Â°-Jornadas Argentinas de Informática*, (Córdoba, Argentina), SADIO, 2013.
3. Martínez Carod, Nadina, Aranda, Gabriela, Sagripanti, Mauro, Faraci, Pamela, and Cechich, Alejandra, "Análisis de la información presente en foros de discusión técnicos," (Mar del Plata, Argentina), pp. 847–856, Oct. 2013.
4. C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.
5. 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.
6. 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.
7. 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.
8. Kuna, H, Rey, M, Martini, E, Solonezen, L, and Sueldo, R, "Generación de un algoritmo de ranking para documentos científicos del área de las ciencias de la computación," (Mar del Plata, Argentina), pp. 787–796, Oct. 2013.