Evaluación de estrategias para clasificar hilos de foros de discusión según su contenido

Valeria Zoratto, Gabriela Aranda, Nadina Martinez Carod, Facundo Otermin

Grupo GIISCo, Facultad de Informática
Universidad Nacional del Comahue. Neuquén, Argentina
{vzoratto, gabriela.aranda, nadina.martinez}@fi.uncoma.edu.ar

Resumen La información contenida en foros de discusión disponibles en la Web, suele ser considerada muy valiosa por usuarios que intentan resolver problemas similares. En función de esta necesidad, en los últimos años se ha generado un especial interés tanto en la búsqueda de técnicas de recuperación como en el análisis de la información que se extrae de hilos de discusión. La propuesta de investigación, a nivel general, es capturar y analizar hilos de discusión existentes en foros técnicos para, a partir de un problema particular, sugerir un conjunto de soluciones exitosas en menos intentos que utilizando buscadores multipropósito tradicionales, teniendo en cuenta diferentes técnicas para lograrlo. En este trabajo en particular, se utilizan métodos de Procesamiento del Lenguaje Natural (PNL) para evaluar la información textual de los hilos de discusión considerando sus categorías gramaticales y el rol de cada palabra en el contexto que es utilizada. Esta funcionalidad es lograda mediante la herramienta Stanford POS Tagger, y el agregado de sinónimos mediante el uso de WordNet. Respecto al caso inicial, se observa que agregar sinónimos de adjetivos, adverbios y verbos al hilo de discusión, permite clasificarlos de manera más precisa respecto a los documentos de referencia, que utilizando el texto sin el agregado de sinónimos.

1. Introducción

A medida que los sitios Web han evolucionado, han dejado atrás su formato estático (pensado solo para brindar información) dando un lugar cada vez más preponderante a las plataformas online para el trabajo colaborativo. El objetivo de dichas plataformas colaborativas es brindar los mecanismos para facilitar la comunicación entre distintos usuarios para que puedan trabajar en forma conjunta, sin importar que se encuentren en el mismo lugar físico ni que lo hagan en el mismo momento. Algunas de estas plataformas se enfocan en producir y compartir información en distintos formatos (como las wikis y weblogs), mientras que otras están enfocadas en el intercambio de opiniones y del conocimiento, como los foros de discusión. Estos últimos son ampliamente utilizados, tanto para solicitar ayuda como para mantener discusiones sobre contenidos relacionados a dominios específicos. En particular, los foros sobre temáticas relacionadas con el desarrollo y mantenimiento de sistemas software, suelen almacenar un gran volumen de contenido útil producido por sus usuarios, que es deseable y valioso que sea extraído y reutilizado [7, 11]. Además, una de las características más destacables de los foros de discusión es que tienen una estructura única, conformada por múltiples hilos,

compuestos a su vez por múltiples posts [4], lo que los hace ideales para la búsqueda de información de un dominio en cuestión.

Buscar soluciones en foros de discusión sobre un problema particular es una tarea cotidiana, pero pocas veces es una tarea sencilla ya que existe un gran volumen de información dispersa en la Web, por lo cual el usuario interesado debe hacer un análisis exhaustivo de las páginas disponibles para determinar cuáles de las soluciones presentadas sirven para resolver el problema que enfrenta. Incluso, a veces la primera solución encontrada no es la más adecuada para su problema y debe probar varias posibles soluciones hasta encontrar la correcta, transformándose en una tarea que le insume gran cantidad de tiempo.

Con el fin de minimizar la cantidad de trabajo manual requerido para encontrar la solución a un problema recurrente, inicialmente se investigaron y se presentaron las características de una herramienta que facilita dicha búsqueda estableciendo prioridades entre las soluciones encontradas [1]. Una componente de dicha herramienta, desarrollada por Zoratto et. al [22], evalúa la información de los hilos de discusión a partir del análisis léxico de los datos haciendo uso de técnicas de Recuperación de Información (RI). En este trabajo se presenta una extensión de la componente, que utiliza la base de datos léxica del idioma inglés WordNet [16] (que almacena relaciones semánticas entre conjuntos de sinónimos), en conjunto con Stanford POS Tagger[1], para obtener sinónimos adecuados según el rol de cada palabra en el contexto que es utilizada dentro del hilo de discusión. Al realizar esta extensión, se busca mejorar la performance de la clasificación de los hilos de discusión respecto a los documentos de referencia agregando sinónimos de sustantivos, adjetivos, adverbios y verbos a los post del foro de discusión.

En la Figura 1 se presenta el proceso en el que se basa esta investigación para la clasificación de hilos de discusión de acuerdo a un conjunto de entidades reconocibles, llamadas *documentos de referencia.* Dado que este ejemplo se basa en problemas del dominio del lenguaje de programación Java, los documentos de referencia utilizados en este trabajo fueron las especificaciones de clases Java de Oracle (versión 5)[2], aunque dicho conjunto de documentos puede variar de acuerdo a la temática de los hilos de discusión a clasificar.

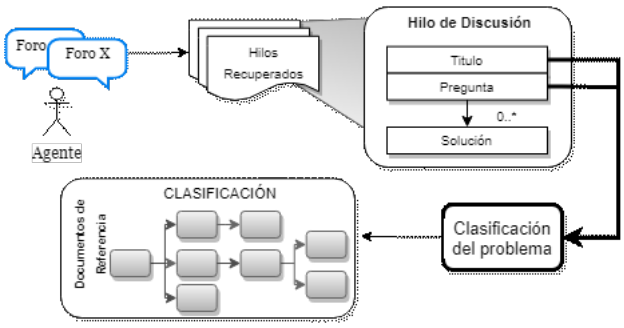

Figura 1. Proceso de clasificación de hilos de acuerdo a documentos de referencia

[1] https://nlp.stanford.edu/software/tagger.shtml
[2] http://docs.oracle.com/javase/1.5.0/docs/

El resto del artículo está organizado de la siguiente manera: en la Sección 2 se presentan algunas investigaciones relacionadas, en la Sección 3 se muestra el diseño de la estrategia empírica a realizar. Luego, en las Secciones 4 y 5 se explican las herramientas utilizadas y la metodología aplicada para llevar a cabo el caso de estudio en el que se enfoca este trabajo. A continuación, en las Secciones 6 y 7 se presentan y analizan los resultados obtenidos. Finalmente, conclusiones y trabajo futuro se presentan en la Sección 8.

2. Trabajos relacionados

Existen varias propuestas de reuso de conocimiento en foros de discusión. Algunos autores extraen pares de <*pregunta, respuesta*> analizando la información léxica y estructural del hilo [8, 12, 13]. Chen y Persen [6] en cambio, analizan automáticamente los hilos de un foro de discusión y proponen otros hilos con contenido similar a los anteriores. Otros autores clasifican los mensajes de un foro de discusión de acuerdo a una jerarquía de temas predefinida [10, 17], mientras que otros intentan identificar los tipos de post que existen dentro de un hilo de discusión [3, 19, 21], haciendo un análisis léxico del texto. Si bien estos trabajos apuntan a facilitar la búsqueda de soluciones dentro del texto por parte de un usuario humano, el contenido resumido de los hilos se puede utilizar para validar si el grado de similitud es más efectivo que con las otras opciones.

A diferencia de los trabajos mencionados anteriormente, este trabajo se enfoca en clasificar los hilos de un foro de discusión respecto a un conjunto de documentos de referencia, a partir del análisis sintáctico y gramatical de los posts. Es decir, que se agregan sinónimos a las palabras contenidas en los hilos de discusión teniendo en cuenta la gramática y luego se clasifican de acuerdo a un tema especifico.

3. Diseño de la estrategia empírica

Antes de avanzar en la definición de la metodología de clasificación, es necesario explicar las características de los documentos en los que se enfoca este estudio.

Por un lado, un foro de discusión está conformado por una colección de hilos. Cada hilo está compuesto por un título, una pregunta principal y una serie de mensajes que constituyen las respuestas (soluciones, pedidos de aclaración, agradecimientos, etcétera), relacionados tanto con la pregunta principal como con otros mensajes del hilo. Por otro lado, la estructura de un documento Oracle (documento de referencia) está conformado por el nombre de la clase (*Class Name*), las interfaces que implementa (*Implemented Interfaces*), las subclases conocidas (*Known Subclasses*), y también los métodos con los que cuenta la clase, indicando parámetros, tipo del resultado, comentarios que aclaran la funcionalidad del método y, en algunos casos, ejemplos de utilización de los mismos.

Inicialmente se propuso evaluar la performance del proceso de clasificación al incluir más o menos información proveniente de distintas secciones de los documentos (hilos de discusión y documentos Oracle), de acuerdo a la siguiente hipótesis:

- ⋆ HIPÓTESIS I: Utilizar mayor cantidad de información de los hilos de discusión permite clasificarlos de manera más precisa respecto a los documentos de referencia.

A efectos de evaluar dicha hipótesis, se desarrolló una metodología de clasificación para hilos de foros de discusión técnicos reales, conformada por la siguiente serie de pasos:

Fase 1 (Recuperación de documentos): Se recuperó un conjunto de hilos del foro de discusión StackOverflow[3] y el total de los documentos de especificación de clases Java del repositorio de Oracle para ser utilizados como documentos de referencia.

Fase 2 (Clasificación por expertos): Los hilos de discusión recuperados en la Fase 1 se analizaron por tres expertos que identificaron las clases Java más relacionadas a cada hilo.

Fase 3 (Procesamiento de los documentos): Los documentos de la Fase 1 se procesaron para eliminar código HTML irrelevante. Además se incorporaron etiquetas en formato XML para determinar comienzo y fin de cada sección (título, pregunta principal y cada respuesta en los hilos de discusión y nombre de clase, interfaces implementadas, etc., en los documentos Oracle).

Fase 4 (Indexación de documentos de referencia): Se utilizó la API Lucene [14] para indexar los documentos de referencia. En esta fase fue necesario redefinir el conjunto de *stopwords*[4] que necesita Lucene, para que no considere como tales las palabras reservadas del lenguaje Java (for, then, if y this).

Fase 5 (Búsqueda de documentos relevantes): Se utilizó el índice generado en la fase anterior y la funcionalidad de búsqueda de Lucene para determinar la relación entre cada hilo y los documentos Oracle, mediante la fórmula de frecuencia de términos TF-IDF, que es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección [18].

Fase 6 (Evaluación): Se contrastaron los resultados de la clasificación propuesta por los expertos (Fase 3) con los retornados por la herramienta propuesta (Fase 5).

En la Figura 2 se muestra una vista de dichas fases, así como de las entradas y salidas de cada una de ellas. Si bien las fases se han numerado y presentado de manera secuencial, algunas de ellas pueden realizarse de manera paralela. Por ejemplo, la Fase 2 (clasificación por expertos) se puede realizar a la vez que la Fase 3, 4 y 5, dado que su evolución y resultado no depende de otras fases o de resultados intermedios.

A partir de la ejecución de un caso de estudio preliminar, se detectó que algunos documentos Oracle aparecían relacionados, con la mayoría de los hilos de discusión del conjunto bajo estudio. Al analizarlos, se descubrió que dichos documentos eran de clases cuyos nombres representan vocablos de uso común en el lenguaje natural del ambiente de la programación (como *Class*, *Error*, *Type*, entre otros), por lo que se repitió la fase de indexación (Fase 4), eliminando dichos documentos del conjunto de referencia. La lista completa de clases que se removieron del conjunto original de documentos de referencia, de ahora en mas *Clases Stopword*s, son las siguientes: *Any, Byte, Class, Doc, Error, Exception, HTML, Key, Method, Object, Operation, Option, Package, Parameter, Point, Result, Set, Source, System, Text, Time, Type, Types, View,* y *Void*.

Para efectuar el primer caso de estudio se capturó un conjunto de 150 hilos que tenían asociado el *tag* <*java*>(Fase 1). Se debe destacar que el tag es asignado por el

[3] http://stackoverflow.com/

[4] Palabras que carecen de significado por sí solas y no brindan información acerca del contenido del texto

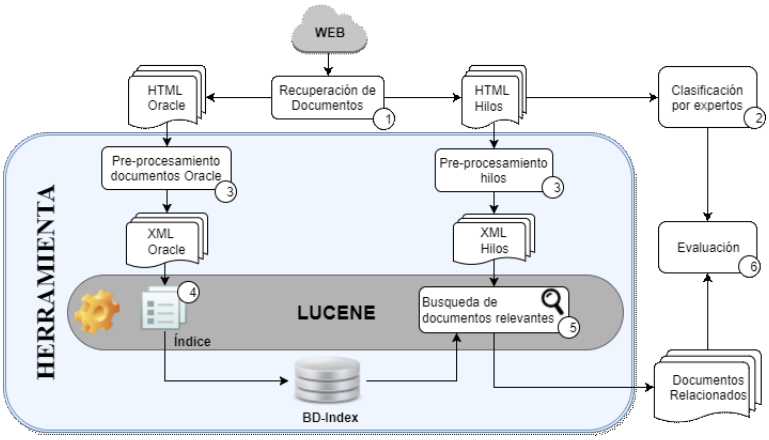

Figura 2. Fases del proceso de clasificación inicial

usuario que crea el hilo para hacer referencia al tema que trata su pregunta entonces, puede ingresar tags que no estén completamente relacionados a la consulta, por lo que, posterior a la clasificación realizada por los expertos (Fase 2), se descartaron los hilos que no referían a problemas específicos con clases Java, obteniendo un conjunto final de 50 hilos.

A fin de analizar la veracidad de la Hipótesis I, en la Fase 3 se generaron 3 versiones de los documentos Oracle, incluyendo información de subconjuntos de las distintas secciones de cada documento original, de la siguiente manera:

- O_a Nombre de la clase
- O_b Nombre de la clase y nombres de todos sus métodos
- O_c Todas las secciones del documento

De manera similar, se generaron 3 versiones de los hilos de discusión, agregando información de las distintas secciones de los hilos en cada una de las combinaciones, para poder evaluar la hipótesis propuesta:

- F_1 Título del hilo
- F_2 Título del hilo y pregunta principal
- F_3 Texto completo del hilo (título, pregunta principal y todas las respuestas)

En base a las 3 versiones obtenidas de cada tipo de documento, se establecieron 9 combinaciones que formaron la base del caso de estudio, tal como se muestra en la Tabla 1.

Tabla 1. Combinaciones de las versiones de documentos a analizar

	F_1	F_2	F_3
O_a	O_aF_1	O_aF_2	O_aF_3
O_b	O_bF_1	O_bF_2	O_bF_3
O_c	O_cF_1	O_cF_2	O_cF_3

A continuación se ejecutaron las fases restantes de la metodología propuesta y los resultados obtenidos fueron analizados diferenciando la cantidad de información en los documentos de referencia (ocurrencias O_1, O_2, O_3) y los hilos de discusión (ocurrencias F_1, F_2, F_3), considerando F-Measure. F-Measure es una de las medidas más utilizadas en RI [2] ya que brinda mayor información respecto a la clasificación, pues combina tanto la precisión como el *recall*. Dicha medida fue aplicada en este trabajo con distintos valores de corte (*cutoff*) sobre la cantidad de respuestas válidas obtenidas, es decir, considerando solamente los primeros N documentos adquiridos en el proceso de recuperación. La medida de corte inicial elegida fue N=2, ya que el mínimo conjunto de clases que tiene relacionado un hilo es 2 (de acuerdo a la clasificación por los expertos en la Fase 2), siguiendo la escala N=3, 4 y 5, tal como se muestra en la Tabla 2. En esta tabla se puede detectar que el mejor valor (0,571) se obtuvo al utilizar solo el nombre de la clase Oracle y el título y pregunta principal de los hilos de discusión (O_aF_2).

Tabla 2. F-Measure por corte sin agregado de sinónimos - Caso Base

	2	3	4	5
O_aF_1	0,489	0,489	0,489	0,489
O_aF_2	**0,571**	0,564	0,556	0,550
O_aF_3	0,541	0,516	0,473	0,450
O_bF_1	0,365	0,363	0,360	0,357
O_bF_2	0,481	0,447	0,416	0,392
O_bF_3	0,498	0,435	0,414	0,380
O_cF_1	0,213	0,190	0,189	0,187
O_cF_2	0,174	0,169	0,166	0,159
O_cF_3	0,211	0,223	0,224	0,214

Además, en la Figura 3 (obtenida a partir del valor promedio entre todos los cortes), se puede observar que, al enfocarse en la cantidad de información contenida en los documentos de referencia, los mejores resultados se obtuvieron al utilizar solo el nombre de la clase (versión O_a), mientras que al enfocarse en la cantidad de información proveniente de los hilos de discusión, la performance fue mayor para los documentos que contienen el título del hilo y la pregunta principal (versión F_2). Sin embargo al utilizar el texto completo del hilo (versión F_3), en todos las pruebas se mejoraron los resultados en comparación con aquellos donde solo se consideró el título del mismo (versión F_1).

En función de los resultados obtenidos se consideró que no podía rechazarse la Hipótesis I planteada[5] y se evaluó la posibilidad de mejorar la performance de dicha clasificación incorporando sinónimos de las palabras contenidas en los hilos de discusión. A tal efecto se estableció una segunda hipótesis de trabajo:

[5] La implementación de los casos de estudio realizados y el análisis de los resultados se encuentra detallado en [23].

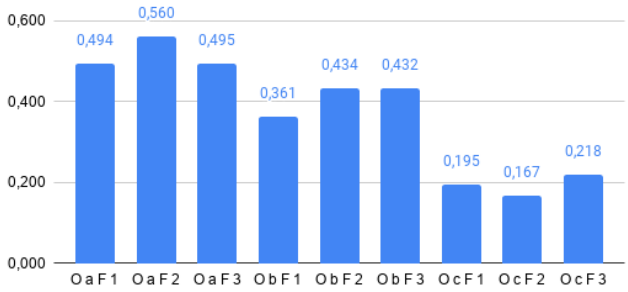

Figura 3. Análisis de performance para la Hipótesis I

⋆ HIPÓTESIS II: Utilizar mayor cantidad de información de los hilos de discusión, incluyendo sinónimos de las palabras utilizadas en el hilo, permite clasificarlos de forma más precisa respecto a los documentos de referencia.

En las secciones siguientes se explica cómo se modificó la preparación de los hilos de discusión para llevar a cabo dicho estudio y los resultados obtenidos tras su aplicación al mismo conjunto de datos, tomando como caso base los resultados presentados en esta sección.

4. Herramientas incorporadas

Stanford POS Tagger: es una compilación de software de código abierto publicada por el Grupo de NLP de la Universidad de Stanford [15] que permite analizar frases en diferentes idiomas y retorna un árbol con la estructura semántica de la oración (PST). Además del PST, esta herramienta proporciona dependencias tipificadas [9] que son tuplas que describen la correlación semántica entre las palabras de la oración. Stanford POS Tagger utiliza una implementación en Java del Modelo de Máxima Entropía optimizado (para tratamiento de palabras desconocidas, desambiguación de frases verbales, preposiciones y adverbios [20]), y el conjunto de etiquetas del estándar de Penn TreeBank[6]. Por ejemplo, para la frase "My dog also likes eating sausage." POS Tagger retorna el siguiente etiquetado: *"My/PRP$ dog/NN also/RB likes/VBZ eating/VBG sausage/NN ./."*, donde PRP$ indica el pronombre posesivo (My); NN indica los sustantivos (dog, sausage); RB indica adverbio (also); VBZ indica el verbo en tercera persona del singular (likes); y VBG indica el gerundio del verbo (eating). En este trabajo, se aplicó Stanford POS Tagger para procesar la estructura de las oraciones y determinar la semántica de las palabras utilizadas.

WordNet: es una herramienta desarrollada en la Universidad de Princeton [16]. Se trata de una gran base de datos léxica que ofrece una red semántica de amplia cobertura del idioma inglés, y es uno de los recursos de NLP más populares [5]. En WordNet las palabras se agrupan en conjuntos de sinónimos (*synsets*) y se almacenan varias relaciones semánticas entre ellos (hiperonimia, hiponimia, antonimia, derivación, entre otros). Dicha herramienta agrupa sustantivos, verbos, adjetivos y adverbios en conjuntos de sinónimos cognitivos, cada uno representando un concepto lexicalizado. Cada palabra en la base de datos puede tener varios sentidos que describen diferentes significados de la

[6] http://www.cis.upenn.edu/~treebank/

palabra. La base de datos no solo distingue entre las formas sustantivo, verbo, adjetivo y adverbio, sino que además categoriza cada palabra en subdominios. Estas categorías son, por ejemplo, en el caso de un sustantivo, artefacto, persona o cantidad. En este trabajo, WordNet se utiliza para obtener los sinónimos correspondientes a cada palabra, de acuerdo a la semántica de las mismas en las oraciones.

5. Mejora del procesamiento de los hilos

Para poder incorporar sinónimos en el mecanismo de clasificación, se modificó el proceso propuesto en [23] (Figura 2), como se muestra en la Figura 4. En la misma se especifica el número de fase con un valor numérico, donde la herramienta propuesta abarca las actividades de la Fase 3 hasta la salida a la Fase 5 explicadas en la Sección 3. El recuadro central de la figura muestra en qué etapas interactúan las herramientas WordNet y Stanford POS Tagger mencionadas anteriormente.

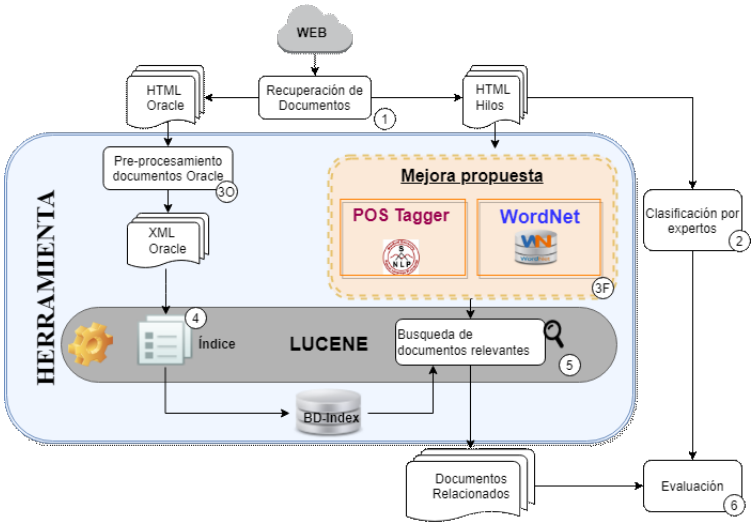

Figura 4. Fases del proceso de clasificación utilizando WordNet y Stanford POS Tagger

Asimismo, dentro de la estructura de la mejora propuesta, en la Figura 5 se muestra cómo interactúan las distintas partes, indicando las entradas y salidas de cada una, como así también el orden secuencial en el que se van realizando. Como se observa en dicha figura, después del pre-procesamiento de los documentos HTML (*Pre-procesamiento de Hilos*), se realiza un segundo procesamiento sobre los documentos XML retornados en dicha etapa, en el cual se utiliza Stanford POS Tagger (*Etiquetado Gramatical*) y se obtiene una segunda versión de los documentos XML (*XML Etiquetado*), marcando el contexto en el que las palabras se utilizan (verbo, adjetivo, adverbio, sustantivo, u otro). En la siguiente etapa (*Incorporación de Sinónimos*), se utiliza WordNet para agregar sinónimos solo de las palabras que corresponden a las clases gramaticales indicadas para el caso de estudio en cuestión (sustantivo, verbo, etc), exceptuando palabras *stopwords* y números.

Una vez obtenido el texto de los documentos XML con el agregado de los sinónimos necesarios (*XML extendido*), se procede a la clasificación y obtención de los documentos relevantes correspondientes a la Fase 5 en Figura 4.

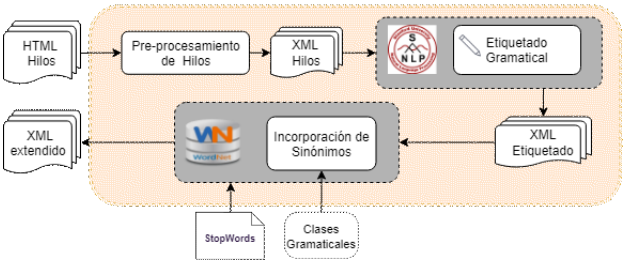

Figura 5. Ampliación de la mejora propuesta

En la próxima sección se muestran los resultados obtenidos luego de aplicar el proceso propuesto al caso de estudio.

6. Caso de estudio considerando sinónimos

A fin de evaluar la performance, se establecieron cuatro pruebas, considerando por separado la inclusión de sinónimos de sustantivos, adjetivos, adverbios y verbos. Luego, el proceso de clasificación se realizó teniendo en cuenta las 9 combinaciones presentadas en la Sección 3 (Tabla 1), según la cantidad de información proveniente de los hilos de discusión y los documentos de referencia.

Con el agregado de sinónimos de sustantivos

En la Tabla 3 (a) se muestran los resultados obtenidos de F-Measure considerando solo los sinónimos de los sustantivos, donde se puede notar que el mejor rendimiento se obtiene (en la mayoría de los casos) para el corte N=2, y el valor más alto (0,536) se obtiene para la combinación O_aF_3, es decir cuando se considera solo el nombre de la clase de los documentos Oracle y el texto completo del hilo de discusión, lo cual apoya la Hipótesis I. Sin embargo, dicho valor no supera al mayor valor obtenido en el caso base (0,571), por lo que no apoyaría la Hipótesis II.

Con el agregado de sinónimos de adjetivos

Por otro lado, al agregar solo sinónimos de adjetivos, tal como se muestra en la Tabla 3 (b), los mejores resultados se obtienen para el corte N=3 (0,581) en las combinaciones que utilizan solo el nombre de la clase de los documentos Oracle y los documentos con más información de los hilos (O_aF_2 y O_aF_3), por lo tanto esta prueba también apoya la Hipótesis I. De manera similar a lo que ocurría al agregar sinónimos de los sustantivos, dicha performance va disminuyendo al agregar más información de

Tabla 3. F-Measure por corte utilizando solo sinónimos de sustantivos (a), y adjetivos (b)

(a) con sinónimos de sustantivos

	2	3	4	5
O_aF_1	0,509	0,519	0,517	0,517
O_aF_2	0,531	0,521	0,489	0,492
O_aF_3	**0,536**	0,527	0,489	0,501
O_bF_1	0,340	0,354	0,348	0,338
O_bF_2	0,417	0,381	0,358	0,352
O_bF_3	0,371	0,305	0,293	0,294
O_cF_1	0,229	0,214	0,191	0,164
O_cF_2	0,157	0,174	0,167	0,159
O_cF_3	0,110	0,123	0,136	0,138

(b) con sinónimos de adjetivos

	2	3	4	5
O_aF_1	0,509	0,522	0,522	0,522
O_aF_2	0,573	**0,581**	0,568	0,572
O_aF_3	0,573	**0,581**	0,568	0,567
O_bF_1	0,400	0,390	0,402	0,401
O_bF_2	0,494	0,456	0,429	0,397
O_bF_3	0,433	0,408	0,394	0,367
O_cF_1	0,218	0,212	0,199	0,176
O_cF_2	0,158	0,190	0,165	0,171
O_cF_3	0,158	0,175	0,159	0,153

los documentos de referencia. Además, en este caso, el valor obtenido para las combinaciones O_aF_2 y O_aF_3 (0,581) es mayor al valor obtenido sin el agregado de sinónimos (caso base = 0,571), lo que apoyaría la Hipótesis II.

Con el agregado de sinónimos de adverbios

De manera similar a la prueba anterior, cuando se incluyen sinónimos de adverbios solamente, los mejores resultados se obtienen con los cortes más pequeños (cortes 2 y 3), como muestra la Tabla 4 (a). Nuevamente, esto ocurre para las combinaciones que utilizan solo el nombre de la clase (en los documentos de referencia) y cuando los documentos con el texto de los hilos incluyen, o bien el texto completo del hilo o al menos el título y la pregunta principal (combinaciones O_aF_2 y O_aF_3), lo que apoya la Hipótesis I. Además, el valor obtenido en ambas combinaciones (0,593) supera al valor del caso base (0,571), como ocurría al agregar solo sinónimos de adjetivos, por lo cual este resultado también apoyaría la Hipótesis II.

Con el agregado de sinónimos de verbos

Finalmente, en la Tabla 4 (b) se muestran los valores obtenidos al clasificar los documentos agregando solo sinónimos de verbos. Como se puede observar, los mejores valores se presentan cuando se utiliza solo el nombre de la clase de los documentos de referencia; y, como ocurrió en las pruebas anteriores, a medida que se agrega más información de dichos documentos disminuye la performance de la clasificación. Por otro lado, al observar el comportamiento entre versiones de los hilos de discusión, a medida que se agrega más información (combinaciones O_aF_2 y O_aF_3) se obtienen los mejores valores, de modo que este resultado apoyaría la Hipótesis I. También, como ocurría al agregar solo sinónimos de adjetivos y adverbios, el valor obtenido (0,574) supera al valor del caso base (0,571), por lo cual esta prueba también apoyaría la Hipótesis II, aunque en menor medida que las dos anteriores.

Tabla 4. F-Measure por corte utilizando solo sinónimos de adverbios (a) y verbos (b)

(a) con sinónimos de adverbios

	2	3	4	5
O_aF_1	0,527	0,524	0,524	0,524
O_aF_2	**0,593**	**0,593**	0,581	0,587
O_aF_3	**0,593**	**0,593**	0,581	0,582
O_bF_1	0,408	0,407	0,412	0,404
O_bF_2	0,502	0,463	0,432	0,400
O_bF_3	0,443	0,428	0,396	0,370
O_cF_1	0,218	0,220	0,205	0,181
O_cF_2	0,158	0,182	0,166	0,154
O_cF_3	0,158	0,161	0,147	0,148

(b) con sinónimos de verbos

	2	3	4	5
O_aF_1	0,506	0,522	0,522	0,522
O_aF_2	**0,574**	0,567	0,549	0,555
O_aF_3	**0,574**	**0,574**	0,555	0,550
O_bF_1	0,366	0,354	0,359	0,366
O_bF_2	0,315	0,288	0,271	0,254
O_bF_3	0,269	0,283	0,254	0,235
O_cF_1	0,176	0,171	0,155	0,139
O_cF_2	0,140	0,142	0,142	0,138
O_cF_3	0,120	0,119	0,128	0,132

7. Análisis comparativo de los resultados

De acuerdo a lo presentado en las secciones anteriores, se observa que la clasificación consiguiría la performance más alta para las combinaciones que consideran solo el nombre de la clase en los documentos de referencia para los cortes 2 y 3 (ocurrencias O_a). Respecto a la cantidad de información de los hilos de discusión, los mejores valores se obtienen también para los cortes 2 y 3, y, en la mayoría de los casos, se mantienen iguales tanto para los documentos que contienen el título y la pregunta principal (O_aF_2) como aquellos que consideran el texto completo del hilo (O_aF_3). Sin embargo, cuando se agregan sinónimos solo de sustantivos, el valor más alto de la clasificación se obtiene al considerar el texto del hilo completo (0,536 para la combinación O_aF_3 en Tabla 3(a)).

Al hacer un análisis más detallado de estos casos, en la Tabla 5 se presenta el promedio de F-Measure calculado solo para los cortes 2 y 3 de las combinaciones O_a, donde se observa que el promedio de la performance para la prueba de sinónimos de sustantivos, es más alto para la versión de los hilos con mayor cantidad de información (0,532 para O_aF_3) y ocurre lo mismo al considerar el promedio de la performance al agregar sinónimos de verbos (0,574 para O_aF_3). Por otro lado, en las pruebas que se agregan sinónimos de adjetivos y adverbios, la performance es igual para las combinaciones O_aF_2 y O_aF_3, por lo que se puede inferir que a mayor cantidad de información de los hilos de discusión mejoraría la performance de la clasificación respecto al conjunto de documentos de referencia (Hipótesis I).

También se puede notar que, en todos los casos, agregar sinónimos al texto completo de los hilos de discusión (combinación O_aF_3) permite clasificarlos de forma más precisa que en el caso base, lo que apoya la Hipótesis II planteada como objetivo de este trabajo. Esto ocurre también en todos las combinaciones O_aF_1 y en la mayoría de las combinaciones O_aF_2, excepto al agregar solo sinónimos de sustantivos (0,526), que se observa una leve disminución de la performance respecto al caso base (0,568).

A continuación, la Figura 6 muestra gráficamente los valores obtenidos en la Tabla 5 donde también se pueden observar ambas tendencias: (1) a medida que aumenta la

cantidad de información (secciones) de los hilos de discusión mejora la performance de la clasificación (Hipótesis I) y (2) en todas las pruebas se percibe una mejora al agregar sinónimos en comparación con el caso base (Hipótesis II).

Tabla 5. Promedio de la performance para cortes N=2 y N=3

	Caso base	Sustantivos	Adjetivos	Adverbios	Verbos
O_aF_1	0,492	0,513	0,516	0,526	0,514
O_aF_2	**0,568**	0,526	**0,577**	**0,593**	0,571
O_aF_3	0,529	**0,532**	**0,577**	**0,593**	**0,574**

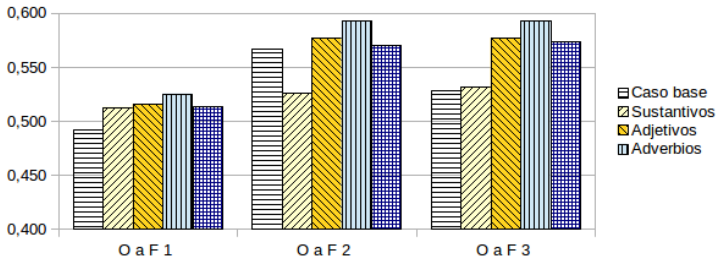

Figura 6. Comparación del promedio de la performance para documentos O_a

Finalmente, al analizar el porcentaje de mejora respecto al caso base, se observa que:

- En las combinaciones F_1 la inclusión de sinónimos de sustantivos mejora el 4,2 %, de adjetivos 4,7 %, de adverbios 6,8 % y de verbos 4,5 %.
- En las combinaciones F_2 disminuye con la inclusión de sinónimos de sustantivos el 7,3 %, pero mejora para verbos (0,5 %), adjetivos (1,6 %) y adverbios (4,5 %).
- En las combinaciones F_3 la inclusión de sinónimos mejora la performance para todas las pruebas. Si bien la diferencia es menor para sustantivos (0,5 %), para el resto de las pruebas se obtienen valores más significativos (9,1 % para adjetivos, 12,2 % para adverbios y 8,6 % para verbos).

8. Conclusiones y trabajo futuro

En este trabajo se presentan los resultados de un caso de estudio, realizado con el fin de analizar estrategias para clasificar hilos de conversaciones recuperadas de un foro de discusión técnico respecto a un conjunto de documentos de referencia. En esta etapa se restringió el análisis a un foro de discusión técnico (StackOverflow) y se tomó como corpus un subconjunto de hilos con el *tag* <*java*> y como documentos de referencia se utilizaron los documentos Oracle de las clases Java versión 1.5.

La clasificación se realizó considerando la información contenida en distintas secciones de ambos tipos de documentos, excluyéndose de los documentos de referencia las clases Java cuyos nombres son de uso común en lenguaje natural (como Error,

Message, etc.), dado que generaban ruido al ser seleccionadas por el algoritmo como relacionadas a un hilo cuando los posts no estaban mencionando a dichas clases.

Según la clasificación automática realizada con la herramienta Lucene, combinada con las herramientas Stanford POS Tagger y Wordnet para la selección de sinónimos según su rol gramatical, los resultados obtenidos apoyarían la primera hipótesis planteada, respecto a que a mayor cantidad de información tomada de los hilos de discusión se logran mejores resultados en la clasificación.

En cuanto a la inclusión de sinónimos, si bien al agregar sinónimos de sustantivos considerando solo el título y la pregunta principal del hilo no se obtiene una performance mayor que en el caso base, se observa que la clasificación mejoraría en todas las pruebas restantes, especialmente al agregar sinónimos de adjetivos, adverbios y verbos por separado, lo cual apoyaría la segunda hipótesis planteada.

En trabajos futuros se planea evaluar cómo se comporta la performance de la clasificación si se realizan combinaciones de las diferentes categorías gramaticales (es decir, incluyendo sinónimos de sustantivos y adjetivos en un mismo caso de estudio, o de verbos, adjetivos y adverbios a la vez, etc.), como así también, dado que este estudio proviene de un conjunto de hilos restringido, se espera replicar los experimentos con mayor cantidad de hilos y usando otras técnicas en la fase de recuperación de información, para asegurar la generalidad de los resultados.

Agradecimientos

Este trabajo está parcialmente soportado por el subproyecto "*Reuso de Conocimiento en Foros de Discusión, Parte II*", correspondiente al Programa de Investigación 04/F002 "*Desarrollo Orientado a Reuso, Parte II*"de la Universidad Nacional del Comahue (Neuquén, Argentina), Periodo 2016-2021.

Referencias

1. Aranda, Gabriela, Martínez Carod, Nadina, Roger, Sandra, Faraci, Pamela, and Cechich, Alejandra. Una herramienta para el análisis de hilos de discusión técnicos. In *CACIC 2014, XX Congreso Argentino de Ciencias de la Computación*, pages 803 – 812, San Justo, Argentina, October 2014.
2. Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
3. Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. Identifying the role of individual user messages in an online discussion and its use in thread retrieval. *Journal of the Association for Information Science and Technology*, 67(2):276–288, 2016.
4. Sumit Bhatia and Prasenjit Mitra. Adopting inference networks for online thread retrieval. In *AAAI*, volume 10, pages 1300–1305, 2010.
5. François-Régis Chaumartin. Upar7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 422–425. Association for Computational Linguistics, 2007.
6. Weiqin Chen and Ricard 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*, pages 309–316, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press.

7. Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng 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.
8. Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng 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.
9. Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Genoa Italy, 2006.
10. D.Helic and N. Scerbakov. Reusing discussion forums as learning resources in wbt systems. In *IASTED International Conference Computers and Advanced Technology in Education*, pages 223 – 228, Rhodes, Greece, 2003.
11. Swapna Gottipati, David Lo, and Jing 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.
12. Liangjie Hong and Brian D Davison. A classification-based approach to question answering in discussion boards. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 171–178. ACM, 2009.
13. Jizhou Huang, Ming Zhou, and Dan Yang. Extracting chatbot knowledge from online discussion forums. In *IJCAI*, volume 7, pages 423–428, 2007.
14. Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.
15. James H Martin and Daniel Jurafsky. Speech and language processing. *International Edition*, 710:25, 2000.
16. George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.
17. Matias Nicoletti, Silvia Schiaffino, and Daniela Godoy. Mining interests for user profiling in electronic conversations. *Expert Syst. Appl.*, 40(2):638–645, 2013.
18. Anand Rajaraman and Jeffrey David Ullman. *Data Mining*, pages 1–17. Cambridge University Press, 2011.
19. Almer S. Tigelaar, Rieks Op Den Akker, and Djoerd Hiemstra. Automatic summarisation of discussion fora. *Natural Language Engineering*, 16:161–192, 4 2010.
20. Kristina Toutanova and Christopher D Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference EMNLP/VLC*, volume 13, pages 63–71. Association for Computational Linguistics, 2000.
21. Eric Utrera Sust, Alfredo Simon-Cuevas, Jose A Olivas, and Francisco P Romero. An approach of a personalized information retrieval model based on contents semantic analysis. *Procesamiento de lenguaje Natural*, (61):31–38, 2018.
22. Valeria Zoratto, Gabriela N Aranda, Sandra Roger, and Alejandra Cechich. Análisis de estrategias para clasificar contenidos en foros de discusión: Un caso de estudio. In *Simposio Argentino de Ingeniería de Software (ASSE 2015)-JAIIO 44*, pages p. 176–190, Rosario, 2015. SADIO.
23. Valeria Zoratto, Gabriela N Aranda, Sandra Roger, and Alejandra Cechich. Analyzing discussion forums threads about java programming language usage. *Electronic Journal of Informatics and Operations Research*, 15(1), 2016.