IFLA WLIC 2013
SINGAPORE

Enviado el: 15 de Junio de 2013

Una ontología basada en un sistema de clasificación de documentos totalmente automático utilizando un sistema semiautomático existente

Spanish Translation of the Original Paper: An ontology based fully automatic document classification system using an existing semi-automatic system

CHAAMINDA MANJULA WIJEWICKREMA
Biblioteca Central, Sabaragamuwa Universidad de Sri Lanka, Belihuloya, Sri Lanka.
manju@sab.ac.lk

RUWAN GAMAGE
Instituto Nacional de Biblioteconomía y Documentación de la Universidad de Colombo, Colombo, Sri Lanka.
mailruga@gmail.com

TRADUCTOR: Pascual Jiménez Huerta. Biblioteca Nacional del España

Resumen:

La clasificación automática de documentos se ha convertido en un área de investigación muy importante, debido al crecimiento exponencial de los contenidos digitales y a que la organización manual o semiautomática no es especialmente eficaz. Por una parte, la clasificación manual y semiautomática es muy minuciosa y laboriosa. Por otro lado, son inevitables en estos dos métodos los errores de clasificación debidos a las imprecisiones de los documentos y de los esquemas de clasificación.

Por tanto, el presente estudio trata de arrojar luz sobre estas cuestiones. Esta investigación propone un sistema automatizado que pueda realizar una clasificación completa de un documento de texto minimizando las ambigüedades del vocabulario. Uno de nuestros estudios anteriores ha desarrollado un sistema semiautomático para la clasificación de documentos y aquí proponemos extenderlo algo más, para obtener un sistema de clasificación de documentos totalmente automático.

Palabras clave: Clasificación automática, clasificación textual, Ontología, función de frecuencia de término -tf idf[1]

[1] ***Tf-idf*** *(del inglés Term frequency – Inverse document frequency), frecuencia de término – frecuencia inversa de documento (o sea, la frecuencia de ocurrencia del término en la colección de documentos), es una*

1 INTRODUCCION

La cantidad de información disponible tanto en formato impreso como electrónico se ha incrementado radicalmente en los últimos años. Este rápido crecimiento de la información en el mundo moderno lleva a la gente a organizar los materiales textuales en métodos de acceso más sencillos. Por otra parte, los esquemas de clasificación normalizados de las bibliotecas se utilizan para lograr este objetivo. Sin embargo, algunos defectos inherentes al lenguaje natural pueden crear graves problemas en el proceso de clasificación. En concreto, las ambigüedades del vocabulario que existen en el lenguaje natural pueden aumentar la inexactitud de la clasificación. Por ejemplo, el termino homónimo "ontología" en el campo de la informática puede ser clasificado también como "ontología" en el campo de la filosofía. Esto, a su vez, afecta a la probabilidad de que un documento pueda ser encontrado. El tiempo y el esfuerzo dedicados a la organización de los materiales es también muy alto en los métodos de clasificación tradicionales. Además, el volumen de documentos digitales también aumenta rápidamente, demandando métodos mecanizados sencillos y asequibles.

La Clasificación Automática de Textos (ATC) es un campo de estudio dedicado a la investigación sobre la categorización automática de documentos digitales en clases predefinidas. De hecho, el factor principal de la ATC es conocido como algoritmo de clasificación o clasificador de texto. Los algoritmos de clasificación que con más frecuencia se utilizan en el proceso de clasificación de textos están basados en máquinas de vectores de soporte, métodos probabilísticos, algoritmos genéricos, métodos de aprendizaje a distancia, modelos ocultos de Markov, métodos de árboles de decisión, métodos de regresión, métodos de reglas de decisión, métodos de redes neurales (Sebastiani, 2002; Tao, Ling, y Cheng, 2005) y funciones de frecuencia tf- idf (frecuencia de término – frecuencia inversa de documento)(Abbas, Smaili, y Berkani, 2010; Tao, Ling, y Cheng, 2005).

En este estudio, trataremos de mejorar aún más un sistema de clasificación de texto semiautomático (híbrido) basado en el tf-idf (denominado HTCS en este documento) presentado anteriormente por los autores (Wijewickrema y Gamage, 2012a, 2012b). La función básica tf- idf utilizada aquí se ha mejorado. Con el fin de reducir las ambigüedades de vocabulario, se ha utilizado una ontología de un dominio.

Aunque HTCS da mejores resultados que el método manual, tiene algunas limitaciones debido a la naturaleza semiautomática del proceso. Como no ha sido completamente eliminada del proceso de clasificación la intervención manual, la clasificación final todavía depende de decisiones humanas. Por una parte, el sistema semi-automático selecciona unas pocas materias candidatas, y el clasificador humano todavía tiene que seleccionar la materia más adecuada de entre todas ellas. Por otro lado, un método totalmente automático puede minimizar aún más el tiempo y el trabajo dedicado a una tarea determinada.

2 REVISIÓN DE LA LITERATURA

La existencia de ambigüedades en el lenguaje natural ha sido ampliamente discutida en la lingüística y en el procesamiento del lenguaje natural (Richardson y Smeaton, 1995). En 1985, la

---

medida numérica que expresa cuán relevante es una palabra para un documento en una colección. Esta medida se utiliza a menudo como un factor de ponderación en la recuperación de información y la minería de texto. El valor tf-idf aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección de documentos, lo que permite manejar el hecho de que algunas palabras son generalmente más comunes que otras. (Nota del traductor)

Universidad de Princeton comenzó el proyecto WordNet (Miller et al, 1990; Morato et al, 2004) para desarrollar un recurso léxico para el idioma inglés. Además, se han desarrollado muchos otros recursos léxicos electrónicos (Best, Nathan, y Lebiere, 2010; Prevot, Borgo, y Oltramari, 2005; Valitutti, Strapparava y Stock, 2004).

También son patentes varios intentos de utilizar recursos léxicos (u ontologías) para la clasificación automática. Prabowo et al. (2002) y Song et al. (2005) presentaron una ontología basada en sistemas de clasificación de páginas Web. Las antiguas ontologías utilizadas se basaban en la Clasificación de la Library of Congress (LCC) y los esquemas de Clasificación Decimal Dewey (DDC). Sin embargo, como comentan Song et al. (2005), Prabowo et al.'s (2002) el trabajo es menos productivo, ya que no se adapta a la creación de una clasificación complicada. Song et al. (2005) desarrollaron su ontología semiautomática en el ámbito de la economía. Sin embargo, como esta ontología no era tan descriptiva, no se puede esperar resultados muy precisos en la clasificación.

Además de a la clasificación de páginas Web, las ontologías se aplicaron también a la clasificación de correos electrónicos y noticias en formato digital. Taghva et al. (2003) formuló una ontología basada en sistemas de clasificación de correos electrónicos. Tenenboim, Shapira y Shoval (2008) presentaron una ontología basada en sistemas de clasificación para periódicos electrónicos.

3 METODOLOGÍA

El sistema propuesto se basa principalmente en tres fases esenciales de clasificación.

3.1 Primera Fase de Clasificación

3.1.1 Detección y eliminación de palabras vacías

La primera etapa del sistema se refiere a detectar y limitar los términos de indización a los términos fundamentales y a la eliminación de palabras vacías para reducir las palabras menos significativas del documento de entrada. Después de estos dos pasos, el documento refinado (ahora limitado a una lista de palabras) se toma como resultado del proceso. Utilizando este resultado, los primeros términos más frecuentes son seleccionados como las palabras clave para determinar la disciplina del documento (documento de entrada).

3.1.2 Colección de Entrenamiento

La colección de entrenamiento es una colección de documentos donde se almacenan los documentos de entrenamiento. Uno puede utilizarlos para compararlos con el documento de entrada a fin de determinar la materia del mismo. Esto se puede hacer mediante la evaluación de las similitudes entre las palabras clave previamente seleccionadas del documento de entrada y los términos en los documentos de entrenamiento. En nuestro estudio, hemos utilizado 385 documentos de entrenamiento dentro del dominio de la filosofía Estos han sido clasificados por clasificadores experimentados de acuerdo al esquema de DDC.

3.1.3 Algoritmo de Clasificación de Textos

En general, los algoritmos de clasificación textuales determinan numéricamente en qué medida un determinado documento se refiere a una materia concreta. En uno de nuestros estudios anteriores (Wijewickrema y Gamage, 2012b), se desarrolló un nuevo algoritmo clasificador de texto utilizando una función de peso de frecuencias de términos existentes llamada función de peso tf-idf (Salton y Buckley, 1988). Este mismo algoritmo clasificador se utiliza aquí también. Se comparan las similitudes entre el documento de entrada y cada uno de los documentos de entrenamiento. Por tanto, la disciplina del documento de entrenamiento que obtiene el valor numérico más alto determina la disciplina del documento de entrada. El nuevo clasificador utiliza

la fórmula que figura a continuación.

$$Document\ Score = \sum_{i=1}^{4} \frac{(tf-idf)_{i,D}}{\sqrt{\sum_{k=1}^{4}(tf-idf)_{k,D}^{2}}} \times (tf-idf)_{i,d} \quad (1)$$

$$Document\ Score = \sum_{i=1}^{4} \frac{(tf-idf)_{i,D}}{\sqrt{\sum_{k=1}^{4}(tf-idf)_{k,D}^{2}}} \times (tf-idf)_{i,d} \quad (1)$$

donde, $(tf\text{-}idf)_{i,D}$ y $(tf\text{-}idf)_{i,d}$ representan el factor de ponderación tf-idf para el término t_i en el documento de entrada D d el documento de entrenamiento d respectivamente. Aquí, limitamos el sumatorio sólo hasta cuatro, ya que consideramos sólo los cuatro primeros términos de mayor frecuencia para determinar la materia del documento de entrada.

No obstante, $(tf - idf)_{i,j}$ se define como sigue:

$$(tf-idf)_{i,j} = \frac{f_{i,j}}{\sum_k f_{k,j}} \times \left[1 + \ln\left(\frac{|N|}{|n_i|+1}\right)\right] \quad (2)$$

donde, $f_{i,j}$ indica el número de veces que aparece la palabra clave t_i en el texto del documento d_j, N es el número total de documentos de la colección y n_i es el número de documentos en los que aparece la palabra clave t_i . En general, la función de peso tf-idf tiene la capacidad de asignar valores numéricos a los documentos basados en muy pocos factores. Estos incluyen la frecuencia de términos, el número de términos total en ese documento, el número de documentos en los que una palabra concreta aparece en la colección, y el número total de documentos en el corpus. En consecuencia, la función de peso tf-idf se puede considerar también como un clasificador básico. Esta forma básica tiene algunos inconvenientes, como por ejemplo, la consideración de una palabra clave suelta para la clasificación sin tener en cuenta la importancia de otros términos con frecuencias igualmente altas. Sin embargo, nuestro nuevo clasificador es capaz de eliminar hasta cierto punto estas dificultades.

3.2 Segunda Fase de Clasificación

En la segunda etapa, se utilizó la ontología de un dominio para reducir las ambigüedades del vocabulario que puedan afectar durante el proceso intermedio de la clasificación. Además, se desarrolló un programa informático para dirigir los resultados generales de clasificación de la primera etapa a la ontología. Por lo tanto, al final de la segunda etapa del proceso de clasificación, se obtienen materias candidatas para el documento de entrada dado.

3.2.1 Ontología

Uno de los principales objetivos de esta investigación es la eliminación de ambigüedades de vocabulario en la clasificación de documentos de texto. Por consiguiente, es lo que se pretende lograr con el uso de una ontología bien estructurada. Para construir la ontología, hemos utilizado dos fuentes. Primero la ontología se construyó utilizando el esquema de DDC y luego enriquecida utilizando la lista de Sears. Nuestra ontología se limita al campo de la filosofía y materias

relacionadas con los fenómenos paranormales. Pertenecen a las clases 110 a 139 de la 21ª edición del esquema de DDC. Esta limitación se ha realizado debido a las dificultades prácticas de la construcción de una gran base de conocimiento. Por lo tanto, el estudio tiene la intención de clasificar los documentos sólo dentro de este rango. Sin embargo, el enfoque mapea las relaciones semánticas cuando existen asociaciones entre los términos que figuran en el dominio y los términos que se encuentran fuera del mismo. Además, un aspecto concreto de la nueva ontología es el desarrollo del tipo de relaciones como sinónimos, cuasi-sinónimos, hiperónimos e hipónimos.

Por ejemplo, la figura 1 muestra la forma en que el término 'cosmología' se presenta en la ontología.

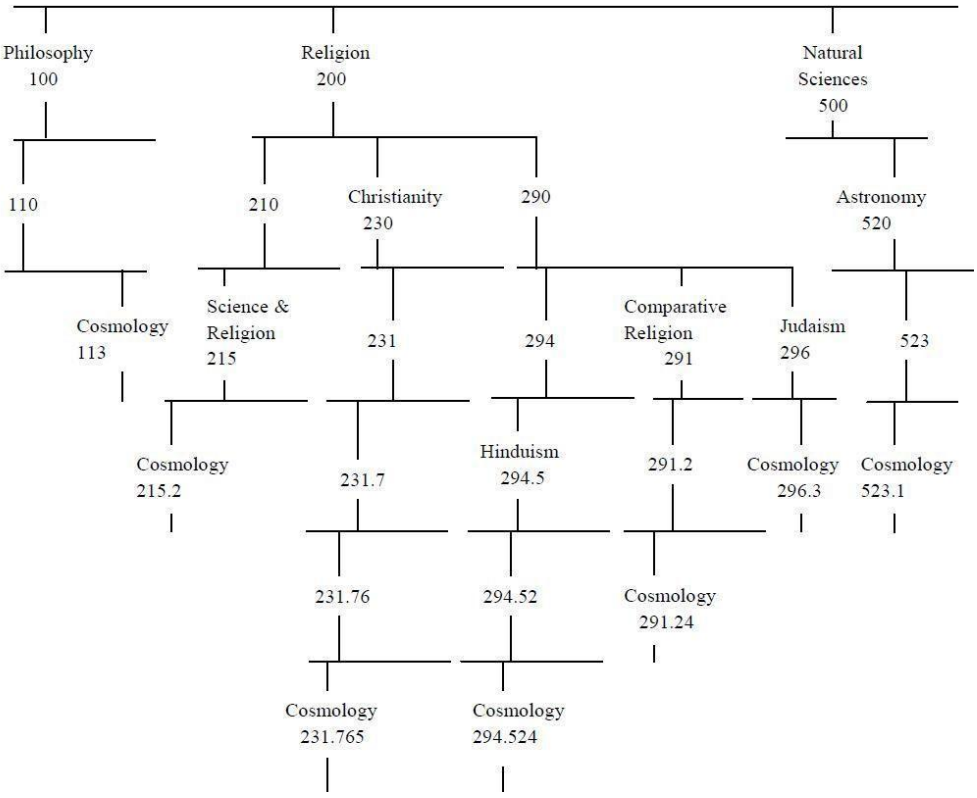

Figura 1: Disposición estructural del término 'cosmología' en la ontología del dominio

Como ilustra la figura 1, el término 'cosmología' aparece en más de un lugar de la ontología y se establecen relaciones entre ellos de manera apropiada. Así que, cada vez que consultamos el término 'cosmología' en la ontología, recupera las siguientes posibilidades con sus descripciones.

Astronomía-Cosmología_En_Filosofía_de_la_Religión_215.2
Cosmología_En_Astronomía_523.1
Cosmología_En_la_Creación_En_el Cristianismo_231.765
Cosmología_En_ el_Hinduismo_294.524
Cosmología_En_la_Filosofía_113
Creación - Cosmología_En_la_Religión_291.24
Teología - Ética – Aspectos Sociales_En_el_Judaismo_296.3

3.3 Tercera Fase de Clasificación

En la etapa final, se ha seguido otra estrategia para filtrar todos los posibles resultados de la

clasificación que ofrece el sistema después de la segunda etapa para que el usuario obtenga sólo una sugerencia del número de materias. Como se muestra en la figura 2, se comparan los términos que figuran al final de la segunda filtración con los términos en el documento original de entrada. Por lo tanto, la disciplina que obtiene la más alta puntuación de coincidencia durante el proceso se puede considerar como la materia más adecuada para el documento de entrada.

Las principales etapas del proceso se muestran en la figura 2.

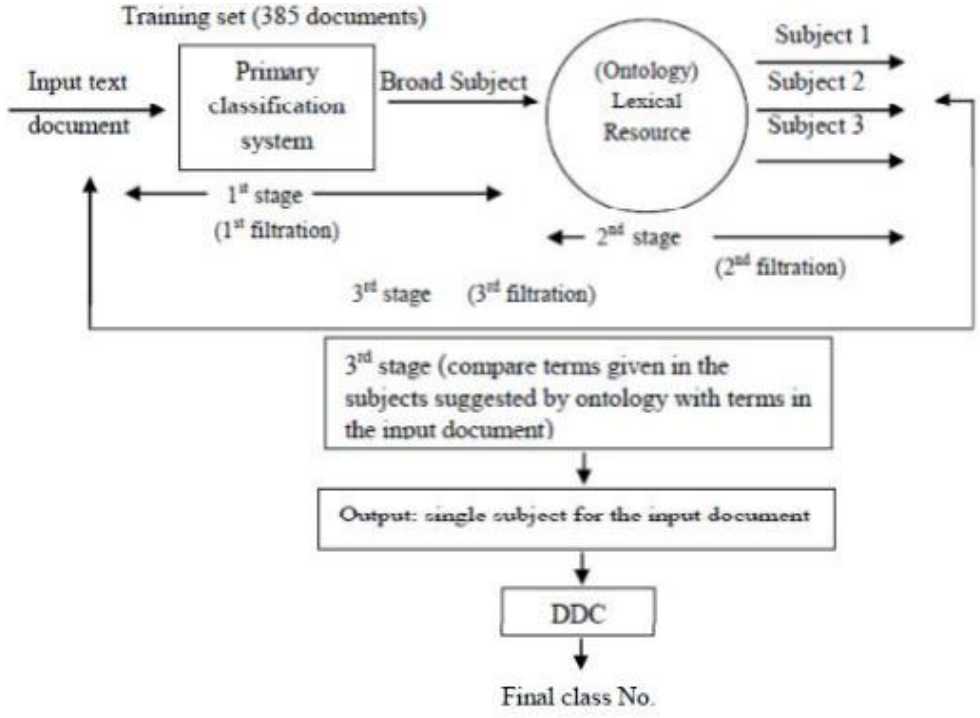

Figura 2: Todas las etapas del sistema de clasificación completamente automático.

3.4 Herramientas usadas para implementar el sistema

En la primera etapa, hemos utilizado la API Lucene (Application Programming Interface) para elegir la materia general más pertinente para el documento de entrada. Para iniciar la segunda etapa, la designación de la materia seleccionada fue enviada a una ontología OWL pre-construida. Esta ontología fue elaborada usando el editor de ontologías Protégé. Tras el desarrollo de la ontología en formato OWL, se combinó con la API Lucene, utilizando la API Protégé-OWL. En este caso la API Protégé-OWL se utiliza como una herramienta para recuperar la información deseada desde el fichero de la ontología OWL. En la salida de la segunda etapa, el sistema proporciona varias materias candidatas. Utilizamos esta información de salida como entrada de la tercera etapa para seleccionar la materia que sea más relevante para el documento de entrada. Para este propósito, se utilizó de nuevo la API Lucene.

El uso de cada herramienta y la forma en que cada una se posiciona en todo el sistema se muestra en la figura 3.

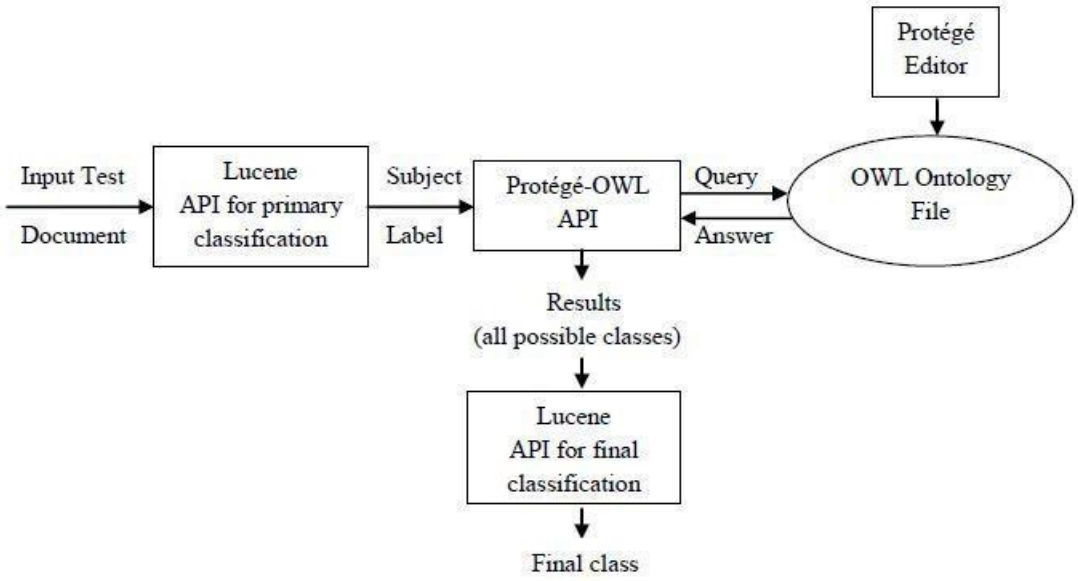

Figura 3: Implementación del sistema usando dichas herramientas

4 Resultados

En nuestro estudio anterior (Wijewickrema y Gamage, 2012a), se clasificaron según el sistema DDC un conjunto de 58 documentos por un clasificador experto. Además, se pidió al clasificador que estudiara los documentos y diera todas las materias/disciplinas posibles y cuál era la materia preferente. A continuación, se clasificó el mismo conjunto de documentos por el mismo clasificador de materias con la ayuda del sistema semi-automático. Una vez más el clasificador fue instruido para seleccionar la materia más relevante de entre todas las posibles dadas por el sistema. Siguiendo esta secuencia, obtuvimos la materia del documento de entrada por medios manuales y semi-automáticos al final de la segunda filtración. Como los documentos de entrada se pre-clasificaron, los resultados se pudieron comparar con exactitud.

Por ejemplo, los resultados de la clasificación obtenidos tras la segunda filtración se muestran en la figura 4. Este documento de entrada pertenece a la materia de "Otras religiones".

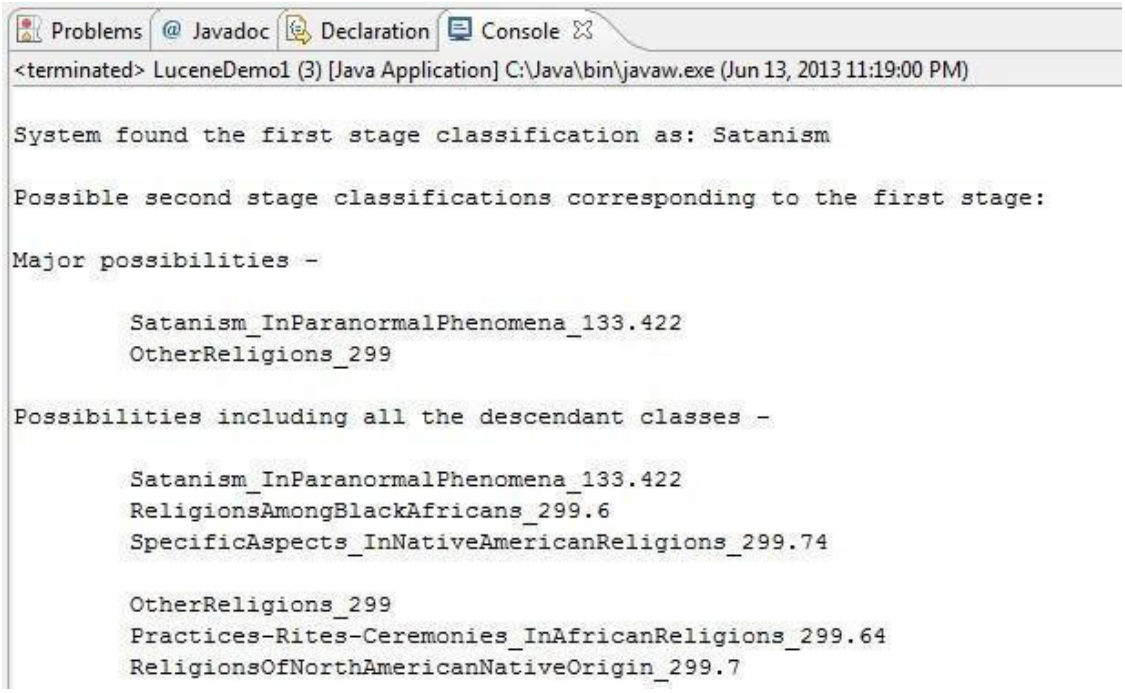

Figura 4: Resultados de la clasificación obtenidos después de la segunda filtración

El mismo conjunto de 58 documentos de entrada se utilizó de nuevo para obtener los resultados del estudio actual. Se clasificaron por el sistema automático; el sistema dio sólo una opción de la materia que mejor coincidía con el documento de entrada. Los resultados se obtuvieron de dos maneras. En primer lugar, se comparó la precisión de la clasificación de los métodos manual, semiautomático y completamente automático. En segundo lugar, se midió la relación entre la imprecisión de los documentos y la inexactitud de la clasificación. Estos resultados también se obtuvieron para los tres tipos de métodos de clasificación manual, semiautomático y totalmente automático.

El resultado de la clasificación obtenida por el nuevo sistema para el mismo documento que pertenece a "Otras religiones" se muestra en la figura 5.

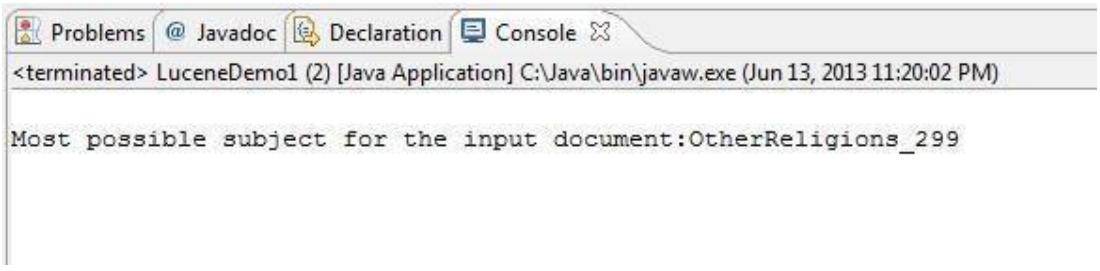

Figura 5: Resultados de la clasificación obtenidos por el sistema completamente automático

4.1 Resultados obtenidos de la precisión de la clasificación.

La Tabla 1 muestra los resultados que se obtuvieron al comparar la precisión de la clasificación con respecto a los métodos de clasificación manual, semiautomático y totalmente automático. La primera columna de la tabla muestra el número de documentos de cada materia que fueron seleccionados para probar la exactitud de la clasificación. Las columnas segunda, tercera y cuarta muestran el número correcto de las clasificaciones que se obtuvieron mediante el método de

clasificación correspondiente.

Tabla 1: Clasificaciones correctas en los métodos manual, híbrido y automático

Materia	Total de documentos	Clasificaciones manuales correctas	Clasificaciones híbridas correctas	Clasificaciones totalmente automáticas correctas
Apariciones	1	0	1	1
Aries	1	1	1	1
Atributos-Facultades	1	0	0	0
Axiología	1	1	1	1
Causalidad	1	0	1	1
Cosmología	4	2	3	0
Epistemología	2	1	2	0
Espíritus malignos	1	0	0	0
Feng Shui	1	0	0	1
Geomancia	1	1	1	1
Leo	1	1	1	1
Libra	1	1	1	1
Amor	1	1	1	1
Mente	1	0	0	1
Ontología	4	3	3	4
Otras Religiones	1	1	1	1
Quiromancia	2	2	2	2
Frenología	2	2	2	2
Piscis	1	1	1	1
Poltergeists	3	3	2	2
Precognición	2	2	2	2
Fenómenos Psíquicos	2	2	2	0
Psicoquinésis	3	2	2	3
Reencarnación	3	2	2	2
Espacio	1	1	1	1

Fenónemos Mediúmnicos Específicos	3	2	0	1
Conjuros-Maldiciones-Encantamientos	4	3	3	0
Espiritismo	1	0	1	1
Tauro	2	2	2	2
Teleología	1	1	1	1
Telepatía	3	3	2	2
Tiempo	2	2	2	2

De acuerdo con los resultados de la Tabla 1, cinco documentos diferentes han sido clasificados de forma más precisa por el sistema totalmente automatizado que por el clasificador humano con la ayuda del sistema automatizado. Estos documentos pertenecen a las materias de Feng Shui, Mente, Ontología, Psicoquinesis y Fenómenos mediúmnicos específicos. Por otro lado, hay diez documentos que fueron clasificados más correctamente por el sistema híbrido que por el sistema totalmente automatizado. Pertenecen a las materias Cosmología, Epistemología, Fenómenos psíquicos, y Conjuros-Maldiciones- Encantamientos. No obstante, de todas las materias que se utilizaron, el 3,125 % fueron clasificadas más correctamente por el sistema totalmente automatizado que por el método híbrido.

La figura 6 presenta una comparación entre los porcentajes de clasificación semi-automática de forma correcta y los porcentajes de clasificación completamente automática de forma correcta.

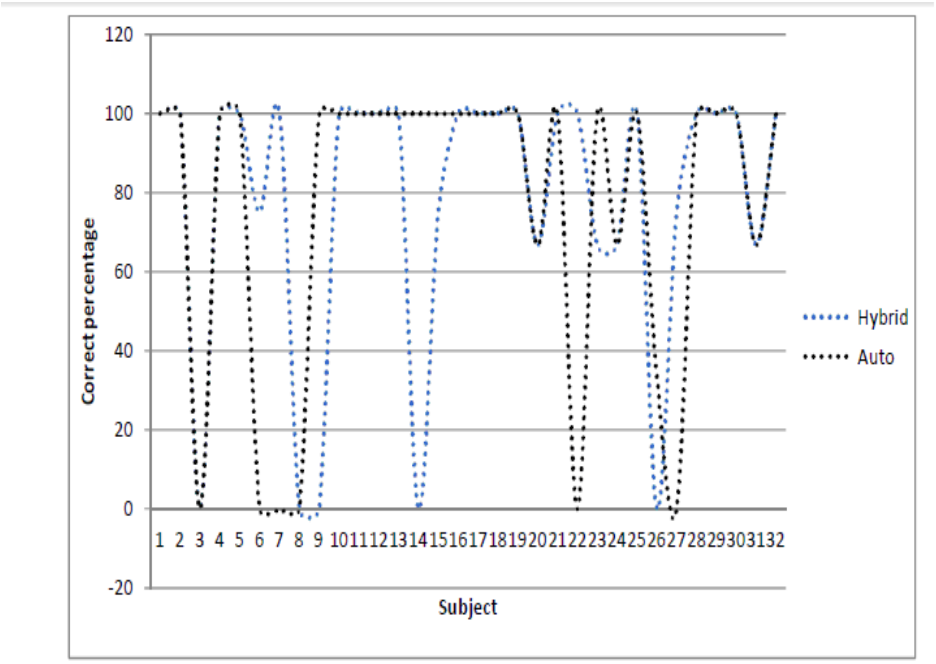

Figura 6: Porcentaje de clasificaciones semiautomática y totalmente automática correctas

Aquí, se puede ver que no hay mucha diferencia en la desviación típica de la media de clasificaciones correctas de las dos curvas en la figura 6. Aunque hay una mejora en la asignación correcta del nuevo sistema, no parece en la figura 6 que sea una diferencia verdaderamente significativa.

4.2 Resultados obtenidos para las imprecisiones y clasificaciones erróneas del documento.
Se analizó el mismo conjunto de documentos para comprobar si existe una relación entre la imprecisión del contenido y los tres métodos de clasificación. Además, se pidió al clasificador que mencionara si los documentos eran ambiguos o no. Los documentos ambiguos se han marcado como 1, mientras que los otros se han marcado como 0. Después de la clasificación, también se marcaron. Si un documento se clasificó incorrectamente, se marcó con un 1, mientras que se le dio un 0 a las clasificaciones correctas. Los resultados obtenidos se muestran en la tabla 2.

Tabla 2: Ambigüedad del documento y clasificaciones inexactas realizadas por los métodos manual, híbrido y totalmente automático.

Materia	Ambigüedad	Clasificación manual incorrecta	Clasificación híbrida incorrecta	Clasificación completamente automática incorrecta
Apariciones	1	1	0	0
Aries	0	0	0	0
Atributos-Facultades	1	1	1	1
Axiología	0	0	0	0
Causa lidadtion	1	1	0	0
Cosmología	1	1	0	1
	1	0	0	1
	1	1	1	1
	1	0	0	1
Epistemología	1	1	0	1
	0	0	0	1
Espíritus malignos	1	1	1	1
Feng Shui	0	1	1	0
Geomancia	0	0	0	0
Leo	0	0	0	0
Libra	0	0	0	0
Amor	1	0	0	0
Mente	1	1	1	0
Ontología	1	0	0	0
	1	1	1	0

	1	0	0	0
	1	0	0	0
Otras Religiones	1	0	0	0
Quiromancia	0	0	0	0
	0	0	0	0
Frenología	0	0	0	0
	0	0	0	0
Piscis	0	0	0	0
Poltergeists	0	0	0	0
	0	0	1	1
	0	0	0	0
Precognición	0	0	0	0
	0	0	0	0
Fenómenos Psíquicos	1	0	0	1
	1	0	0	1
Psicoquinésis	0	0	0	0
	0	0	0	0
	1	1	1	0
Reencarnación	0	0	1	1
	1	1	0	0
	0	0	0	0
Espacio	1	0	0	0
Fenómenos Medñiunnicos específicos	1	0	1	0
	1	0	1	1
	1	1	1	1
Conjuros-Maldiciones-Encantamientos	0	0	0	1
	0	0	0	1
	0	0	0	1
	1	1	1	1
Espiritismo	1	1	0	0

Tauro	0	0	0	0
	0	0	0	0
Teleología	0	0	0	0
Telepatía	0	0	1	1
	0	0	0	0
	0	0	0	0
Tiempo	1	0	0	0
	1	0	0	0

El análisis de los resultados anteriores se ha realizado mediante regresión logística binaria. Aquí, la razón de probabilidades[2] se ha utilizado para determinar la existencia de una relación entre la ambigüedad de los documentos, y las clasificaciones incorrectas obtenidas mediante los métodos manual, semiautomático y totalmente automático. De hecho, la razón de probabilidad se utiliza para medir el tamaño del efecto, describiendo la fuerza de asociación entre dos valores de datos binarios.

La razón de probabilidad para la ambigüedad de los documentos y la clasificación incorrecta de los documentos realizados manualmente fue de 29,00, mientras que la misma proporción para la ambigüedad de los documentos y clasificaciones incorrectas realizadas por el método semi-automático fue de 3.61. Por otro lado, la razón de probabilidad de la ambigüedad de los documentos y la clasificación incorrecta de los documentos realizados por el sistema completamente automático fue de 2.46. Por tanto, la razón de probabilidades obtenida para el efecto de ambigüedad en las clasificaciones incorrectas hechas de forma automática es la más baja.

5 Conclusiones

Los resultados de la clasificación obtenidos por los métodos manual, semiautomático y totalmente automático fueron comparados para determinar el trabajo de clasificación. Aunque el sistema híbrido muestra clasificaciones más precisas de los documentos que la clasificación totalmente automática, el sistema totalmente automatizado clasificó un 3.125 % más de materias de forma correcta que el sistema híbrido. La razón podría ser el error de interpretación por el nuevo sistema de algunas disciplinas, como la Cosmología y los Conjuros-Maldiciones-Encantamientos. Por tanto, cuantos más documentos de entrada pertenecen a estas materias, el número de documentos mal clasificados también sube. Aparte de esos casos especiales, el rendimiento general da mayor precisión en la clasificación al nuevo sistema que al método anterior. Este estudio sugiere, además, que la ambigüedad de los documentos tiene un mayor efecto sobre la clasificación semiautomática, que en el método totalmente automático de clasificación.

Referencias

Abbas, M., Smaïli, K., & Berkani, D. (2010). Efficiency of TR-Classifier versus TFIDF. *2010 First International Conference on Integrated Intelligent Computing*. doi: 10.1109/ICIIC.2010.60

Best, B. J, Nathan, G., & Lebiere, C. (2010). Extracting the Ontological Structure of OpenCyc for Reuse and Portability of Cognitive Models. *Proceedings of the 19tConference on Behavior*

[2] La odds ratio no tiene una traducción concreta en español aunque hace referencia a la razón de momios que es una medida de tamaño de efecto. También se puede traducir como razón de momios, razón de posibilidades o razón de oportunidades. (Nota del traductor)

Representation in Modeling & Simulation (BRiMS 2010). Disponible en: http://www.adcogsys.com/pubs/Brims2010-best-gerhart-lebiere- opencyc.pdf

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*. Disponible en: http://courses.media.mit.edu/2002fall/mas962/MAS962/miller.pdf

Morato, J., Marzal, M. A., Llorens, J., & Moreiro, J. (2004). WordNet Applications. *Proceedings of the 2nd International Conference on Global WordNet*. Disponible en: http://www.fi.muni.cz/gwc2004/proc/105.pdf

Prabowo, R., Jackson, M., Burden, P., & Knoell, H. D. (2002). Ontology-Based Automatic Classification for the Web Pages: Design, Implementation and Evaluation. *Proceedings of the 3rd International Conference on Web Information Systems Engineering*. Disponible en: http://portal.acm.org/citation.cfm?id=674083

Prevot, L., Borgo, S., & Oltramari, A. (2005). Interfacing Ontologies and Lexical Resources. *Proceedings of OntoLex 2005*. Disponible en: http://www.loa-cnr.it/Papers/%5B22%5DprevotBorgoOltramari-3.pdf

Song, M. H., Lim, S. Y., Kang, D. J., & Lee, S. J. (2005). Automatic Classification of Web Pages based on the Concept of Domain Ontology. *Proceedings of the 12th Asia- Pacific Software Engineering Conference (APSEC'05)*, 645-651. doi: 10.1109/APSEC.2005.46

Tenenboim, L., Shapira, B., & Shoval, P. (2008). Ontology-based Classification of News in an Electronic Newspaper. *Proceedings of the International Conference on Intelligent Information and Engineering Systems*. Disponible en: http://www.foibg.com/ibs_isc/ibs-02/IBS-02-p12.pdf

Valitutti, A., Strapparava, C., & Stock, O. (2004). Developing Affective Lexical Resources. *PsychNology Journal*, 2(1), 61-83. Disponible en: http://www.psychnology.org/File/PSYCHNOLOGY_JOURNAL_2_1_VALITUT TI.pdf

Wijewickrema, P. K. C. M. & Gamage, R. C. G. (2012). Automatic Document Classification Using a Domain Ontology, *Proceedings of the 09th National Conference on Library and Information Science (NACLIS 2012)*, ISSN 978-955-9075-17-2, 85-107.

Wijewickrema P. K. C. M. & Gamage, R. C. G. (2012). An enhanced text classifier for automatic document classification, *Journal of the University Librarians' Association of Sri Lanka, 16* (2), ISSN 1391-4081, 138-159.