Título
Desarrollo de un sistema de clasificación automática de contenidos en medios de comunicación hispano-mexicanos

Resumen
El objetivo de la investigación es desarrollar un sistema de clasificación automática para los contenidos recuperados a través de la plataforma Resync, especializada en la investigación de fuentes de información en medios de comunicación. Se justifica su desarrollo debido a la falta de métodos automatizados para organizar la información recopilada por medio de dicha plataforma. Por otro lado, debido a la necesidad de estudiar en profundidad las categorías temáticas abordadas por los medios de comunicación según el país. Para resolver estos problemas, se transforma el tesauro multilingüe Eurovoc, en una pseudo-ontología, que es utilizada como vocabulario clasificatorio del corpus documental, compuesto por más de 400.000 noticias publicadas durante los meses de junio-julio de 2011, entre medios Mexicanos y Españoles. Por otro lado, se diseñan y prueban 5 algoritmos de clasificación automática, de consulta precisa y genérica, que emplean el vocabulario clasificatorio anteriormente mencionado, para su equiparación con la colección de prueba. Se obtienen todos los resultados cuantitativos del experimento, concluyendo un escalonamiento progresivo en el porcentaje de contenidos clasificados, dado por el grado de precisión del algoritmo y su condicionamiento. Finalmente se sientan las bases para evaluar cualitativamente la clasificación efectuada por el sistema, con el objetivo de perfeccionar el proceso aquí descrito.

Palabras clave
Clasificación automática, ontologías, tesauros, automatización, sindicación de contenidos, medios de comunicación, normalización de textos, recuperación de información, evaluación

Title
Development of an automatic classification of content system in Spaniard-Mexican mass media

Abstract
The objective of this research is to develop an automatic classification system for the contents retrieved through the Resync platform specializing in the investigation of sources of information media. This investigation is justified due to the lack of automated methods to organize the information gathered and the need to scrutinize the thematic categories addressed by the media by country. To resolve these problems, we transform the Eurovoc multilingual thesaurus in a pseudo-ontology vocabulary that is used as a qualifier for the documentary corpus. The test collection used has 400,000 contents from Mexican and Spaniard media published during the months of June-July 2011. Additionally, are designed and tested 5 automatic classification algorithms, accurate consultation and generic classification using the vocabulary above, for their harmonization with the collection of evidence. You get all the quantitative results of the experiment, concluding a progressive escalation in the percentage of classified content, given by the precision of the algorithm and its conditioning. Finally, the basis for qualitative evaluation of the classification made by the system, in order to perfect the process described herein.

Keywords
Automatic classification, ontology, thesauri, automation, content syndication, mass media, text normalization, information retrieval, evaluation

Introducción

El desarrollo de un sistema de clasificación automática forma parte de la presente investigación, ya iniciada sobre la plataforma de contenidos sindicados Resync en la que se presenta una prueba de recuperación de contenidos en medios de comunicación hispano-mexicanos (BLÁZQUEZ OCHANDO, M. and Serrano Mascaraque, E., 2011). El elevado volumen de datos recopilados, más de 5.000 canales de sindicación y 60.000 noticias en 24 horas, revelaron la necesidad de elaborar un sistema de clasificación para el control y explotación de la información. Si las pruebas sobre la plataforma continuaran con periodos de duración más largos, la carencia de un método de clasificación provocaría que miles de contenidos fueran difícilmente recuperables. Si se tiene en cuenta que uno de los objetivos perseguidos es conocer la tipología temática de la información publicada por los distintos medios de comunicación, se justifica más si cabe el desarrollo de un método que permita su organización.

Por ese motivo se plantean diversos interrogantes a los que se pretenden dar solución, destacando los siguientes; qué dificultades plantea el desarrollo de un sistema de clasificación automático, cómo diseñarlo, con qué metodología; qué factores influyen en los algoritmos de consulta dentro del proceso de clasificación; qué tipo de clasificación automática se puede llevar a cabo; hasta qué punto puede el corpus documental de la colección jugar un papel importante en todo el proceso; qué lenguaje documental puede ayudar a clasificar los contenidos; qué diferencias cuantitativas se obtienen al utilizar los distintos algoritmos de clasificación.

Una forma de dar respuesta a tales preguntas, viene dada por la investigación realizada por la comunidad científica en congresos y foros especializados en la materia, como por ejemplo el programa de clasificación y traducción multilingüe promovido por (NIST, 2011). De utilidad para desarrollar esta investigación son aquellos trabajos en los que se emplean modelos basados en ontologías, tanto para la organización de páginas web (PRABOWO, R. et al., 2002), como para los documentos recopilados por medio de minería de datos (GELERNTER, J., 2008). Las ontologías y las clasificaciones automáticas tienen por otra parte, un uso extensivo en el campo de la biomedicina, como demuestran múltiples investigaciones presentadas en las Text REtrieval Conferences, véanse (SUBRAMANIAM, V. et al., 2005) y (COHEN, A. M. and Hersh, W. R., 2006), trabajos en los que se emplea como elemento categorizador la lista de encabezamientos de materia del MeSH y como base de conocimiento los documentos recopilados en Medline (CAMOUS, F. et al., 2007). Su metodología determina claramente la definición de la base de conocimiento objetivo y un vocabulario controlado para la clasificación, formado por categorías y ontologías. Seguidamente se diseñan esquemas de equiparación de conjuntos, algoritmos y excepciones de los procesos de clasificación. Este tipo de desarrollos han inspirado la concreción de un modelo de clasificación que involucra plenamente los lenguajes documentales y la recuperación de información clásica como piedras angulares de la categorización de textos. Establecer, en consecuencia, la pertenencia de un documento a una categoría temática, se convierte en una tarea compleja debido a la característica secuencial del proceso automático, tal como afirma (SÁNCHEZ JIMÉNEZ, R., 2007), ya que no incorpora un razonamiento humano y por lo tanto cualitativo de los contenidos que son analizados por los programas. En todos los casos, se prueban modelos de clasificación, que procuran aproximarse tanto en la precisión, como en la exhaustividad, a la categorización más perfecta posible.

Metodología

La metodología aplicada en la investigación es de tipo experimental, ya que se trabaja desarrollando un programa adaptado a una plataforma de recuperación de contenidos sindicados, para la que no existe módulo o herramienta de clasificación automática disponible. El paradigma tecnológico empleado es Apache, MySQL y PHP, en correspondencia con la arquitectura utilizada en la plataforma Resync.

Las especificaciones de diseño para el sistema de clasificación automática son las de una aplicación de tipo no supervisada y autónoma en su ejecución y resolución. No paramétrica, por incorporar un vocabulario controlado, que actúa como patrón de recuperación y agrupación. De clasificación simple, evitando solapamiento de categorías o varias de ellas por documento, con el objetivo de asignar aquella que mayor relevancia obtenga. Su punto de vista de clasificación es centrado en el documento, ya que a partir de una categoría determinada es posible recuperar todos los documentos clasificados en la misma. Esta característica permitirá en sucesivas investigaciones usar de mecanismos de post-clasificación, basados en k-NN[1] (HART, P. and Cover, T., 1967).

Paralelo al desarrollo tecnológico, se define una metodología basada en las siguientes fases de trabajo: 1) Elección del vocabulario de clasificación, 2) Transformación y adaptación del vocabulario de clasificación, 3) Preparación del corpus documental, 4) Normalización e indexación, 5) Diseño, ejecución y evaluación de algoritmos de clasificación automática.

1) Elección del vocabulario de clasificación

Este apartado explica la importancia de elegir un vocabulario de clasificación adecuado, para su comparación y contraste con una colección de prueba. Ello viene determinado por el grado de heterogeneidad de la misma, siendo necesaria una correlación clara del espectro temático abordado. Dicho vocabulario puede constituirse de diversas formas:

- *Recopilación de las cadenas de consulta del usuario.* Las búsquedas efectuadas son almacenadas junto a la información de la experiencia de usuario. Esto es el número de resultados obtenidos con la consulta, qué resultados fueron seleccionados por el usuario para su análisis, por cuánto tiempo fueron visualizados o cuál es su nivel de enlazamiento.

- *Selección de textos especializados.* Utilizando textos especializados en un área de conocimiento, es posible extraer los términos más representativos y con mayor capacidad discriminatoria. De esta forma es posible entrenar sistemas de clasificación de forma extensiva.

- *Uso de vocabularios controlados.* La extracción de los términos de un tesauro o una ontología permiten elaborar conjuntos terminológicos aceptados.

En este caso, se ha tenido en cuenta el tesauro multilingüe Eurovoc, cuyo vocabulario controlado se utilizará, para clasificar el corpus documental de noticias de medios de comunicación hispano-mexicanos. La elección de este tesauro se debe a su amplia

[1] Dado que el sistema de clasificación agrupa los documentos más relevantes de la colección para cada categoría, éstos pueden ser empleados en posteriores investigaciones como grupos de entrenamiento para la reclasificación de contenidos basándose en el algoritmo de vecinos más próximos.

diversidad temática, mejor adaptada para clasificar la alta heterogeneidad de contenidos de la colección (DEXTRE CLARKE, S.G., 2010)[2]. Por otro lado existen diversos estudios (STEINBERGER, R., 2010) que vienen empleando el tesauro Eurovoc, como base de pruebas para la clasificación e indexación automática de publicaciones oficiales. Otra razón que justifica su empleo y experimentación es la existencia de otras iniciativas de clasificación automática (DAEDALUS, 2011), lo que indica su idoneidad para un uso experimental y académico.

2) Transformación y adaptación del vocabulario de clasificación

El tesauro Eurovoc, en primera instancia, no puede utilizarse para la clasificación intensiva de contenidos. En su presentación jerárquica original, los términos de cada categoría principal deben ser agrupados. Este proceso se lleva a cabo mediante expresiones regulares que eliminan todos los términos relacionados, manteniendo en todo caso los términos específicos de todos los niveles, véase *tabla1*. El resultado de este proceso se resume en un listado de términos separados por comas, que mantienen su correspondencia con la categoría principal.

	Función / Descripción	REGEXP	Reemplazo
Paso 1	Eliminación de términos relacionados y sus saltos de línea.	RT .*\r	
Paso 2	Eliminación de las etiquetas NT (narrow term) correspondientes a términos específicos con independencia del nivel.	NT[1-6]	
Paso 3	Sustitución de los saltos de línea por comas, lo que supone una lista de términos convenientemente separados.	\r	,

Tabla 1. Operaciones de adaptación del vocabulario de clasificación

Este esquema de términos es trasladado a una estructura de almacenamiento, denominada *matriz multidimensional*, en la que se guarda la relación de los términos específicos recopilados, con su categoría general inmediatamente superior, así como con su cabecera terminológica, véase *tabla2*.

Esquema de la matriz multidimensional	
Término cabecera	`$ontindex[] = array(cod => "código del término cabecera", title => "TC Término Cabecera propiamente dicho");`
Bolsa de términos	`$ontology[] = array(cod => "código del término cabecera", head => "Categoría principal", bag => "término1, término2, término3, término4,…"`
Ejemplo de matriz multidimensional con contenidos	

```
$ontindex[] = array(cod => "3", title => "VIDA ECONÓMICA");

$ontology[] = array(cod => "3", head => "Contabilidad nacional", bag => "agregado económico,
gasto nacional, producto interior bruto, producto nacional, producto nacional bruto, renta
nacional, contabilidad económica agrícola, contabilidad regional, producto regional bruto, renta
per cápita, renta por persona activa, renta, acumulación de rentas, ahorro, ahorro forzoso,
distribución de la renta, distribución de la riqueza, empobrecimiento, pobreza, mendicidad,
renta baja, riqueza, nivel de vida, poder adquisitivo, paridad de poder adquisitivo, presupuesto
familiar, redistribución de la renta, renta familiar, transferencias sociales, sistema de
contabilidad, sistema normalizado de contabilidad, Sistema Europeo de Contabilidad");
```

Tabla 2. Matriz de almacenamiento de categorías temáticas y vocabularios

[2] La aportación de Stella Dextre en la Conferencia Eurovoc 2010, dejó patente la interoperabilidad del Tesauro Eurovoc con otros vocabularios al cumplir el estándar ISO 25964, por el que el tesauro se convierte en apto para la recuperación de información y posibilitando la cobertura de otros tesauros a través de mapas de equivalencias jerárquicas y asociativas de los términos que lo componen.

Obsérvese que la *tabla2*, muestra un término cabecera "*Vida económica*" relacionado con una categoría principal "*Contabilidad nacional*" bajo la que se aglutina un vocabulario de términos "*agregado económico*, *gasto nacional…*" que serán la base para la construcción de todas las cadenas de consulta del sistema de clasificación. En dicho vocabulario, se presentan términos que originariamente ocupaban varios niveles de especificidad. Este proceso los iguala al incluirlos en una lista y permitirá a la postre combinarlos para obtener los mejores resultados posibles. Se tiene en consecuencia, la estructura principal del Tesauro Eurovoc cuya apariencia, más parece una ontología, que un tesauro, al eliminar su jerarquización. Se debe observar que en este estado, los términos constan de signos diacríticos, mayúsculas y caracteres especiales, que deben ser normalizados. Dicho proceso es obligado tanto para el vocabulario de clasificación como para el corpus documental, con la finalidad de facilitar la clasificación (BERRY, M.W. and Browne, M., 2005).

3) Preparación del corpus documental

La colección de documentos de prueba utilizada en esta investigación, fue generada con la plataforma Resync desde el 22 de Junio de 2011 hasta el 31 de Julio del mismo año. Está compuesta por más de 400.000 contenidos publicados por los medios de comunicación españoles y mexicanos, entre los que destacan los medios de prensa de forma más notable, véase *tabla3*.

	Prensa		Radio		Televisión		Nº total de contenidos
	Fuentes	Contenidos	Fuentes	Contenidos	Fuentes	Contenidos	
España	1.055	155.740	710	23.980	494	45.501	225.221
México	677	186.899	141	10.520	110	8.952	206.371
Total	1.732	342.639	851	34.500	604	54.453	431.592

Tabla 3. Características de la colección de prueba, objeto de clasificación

Para poder analizar el comportamiento del sistema de clasificación en el caso español y mexicano, se probará de forma diferenciada cada algoritmo con cada país, permitiendo determinar qué textos resultan más sencillos de clasificar, en qué cantidad y en su caso a qué categoría corresponden.

4) Normalización e indexación

Tanto el vocabulario de clasificación obtenido como el corpus documental, requieren un proceso de depuración y normalización, que se resume en las siguientes fases:

– *Supresión del código fuente.* Aplicable a los contenidos recuperados por la plataforma Resync y que constituyen el corpus documental. En muchos casos, estos contenidos tienen etiquetado HTML que dificulta el tratamiento del texto para su recuperación. Por este motivo debe transformarse el contenido hipertextual en texto plano. Para conseguirlo se utiliza la *función strip_tags*[3] para eliminar automáticamente las principales etiquetas que causarían dificultades en el proceso de indexación. Si bien este filtro elimina la mayor parte del etiquetado, no lo valida, haciendo inviable la supresión de aquellos códigos mal escritos o complejos. En tales casos se requiere de una batería de filtros que contemplen las principales excepciones por omisión de la función anterior, o bien por medio del empleo de

[3] strip_tags. PHP Ref. http://php.net/manual/es/function.strip-tags.php

expresiones regulares, utilizando la *función preg_replace*[4], con la que se busca un patrón coincidente para su reemplazo, por otra cadena de texto ó la cadena vacía. La dificultad de este estadio de la depuración ha sido en muchos casos puesta de manifiesto en el desarrollo (CUNNINGHAM, H. et al., 2012)[5] de programas especializados en la depuración de textos.

– *Tokenización.* Se lleva a cabo cuando el texto se encuentra libre de toda codificación. Éste se descompone palabra a palabra utilizando como medio divisor los espacios que las separan. La función utilizada para este efecto es *explode*[6], que automáticamente almacena los términos separados uno a uno, dentro de una *matriz sencilla*.

– *Normalización de caracteres.* Los términos almacenados son transformados con la codificación hexadecimal, para hacer efectiva la sustitución de caracteres acentuados, el reemplazo de caracteres especiales y la eliminación de signos de puntuación.

– *Eliminación de palabras vacías.* Una lista de palabras vacías se va contrastando con cada término del texto previamente depurado, tokenizado y normalizado. Este proceso se ejecuta con la *función stristr*[7], que determina la coincidencia de cada término con cada palabra vacía, lo que permite generar una nueva matriz de términos aceptados, que será definitiva y convenientemente almacenada para su indexación posterior.

5) Diseño, ejecución y evaluación de algoritmos

El objetivo de diseño de los algoritmos es lograr métodos de clasificación lo más precisos posibles y por otro lado desarrollar algoritmos de clasificación con un alto grado de exhaustividad para conseguir un alto porcentaje de documentos categorizados. En la consecución de estos objetivos, se cuenta con las opciones y propiedades de consulta de la base de datos MySQL. De entre todas las disponibles, se harán uso de las siguientes:

– *Búsqueda en lenguaje natural*[8], para cadenas de consulta, confrontadas con el texto indexado de la colección. Se obtienen documentos en ordenados según rango decreciente de relevancia, siendo los últimos, aquellos con un valor cercano a 0, que indica una similaridad nula. La relevancia se calcula en base al número total de palabras en la colección y el número de documentos que contienen los términos de la consulta en particular, aplicando el cálculo de similaridad del coseno y la métrica del espacio vectorial (ARSLAN, A. and Yilmazel, O., 2010, p.367). El Cálculo del peso de los términos, se ciñe al modelo de ponderación TF-IDF clásico (ROBERTSON, S., 2004, pp.503-520), determinando que las palabras con mayor frecuencia de aparición y por ende menor valor semántico y discriminatorio, tengan un peso menor que aquellas menos frecuentes, con una frecuencia media o baja. De

[4] preg_replace. PHP Ref. http://php.net/manual/es/function.preg-replace.php
[5] El programa GATE de la Universidad de Sheffield, sirve como herramienta de simulación para el procesamiento de textos especializados y la resolución de sus posibles problemas. En este sentido incluye mecanismos para la eliminación del etiquetado de un texto procedente de la web en diversos formatos, entre ellos HTML, XML y PDF para su posterior recuperación utilizando métodos de modelado semántico.
[6] explode. PHP Ref. http://php.net/manual/es/function.explode.php
[7] stristr. PHP Ref. http://php.net/manual/es/function.stristr.php
[8] Natural Language Full-Text Searches. MySQL Ref. http://dev.mysql.com/doc/refman/5.1/en/fulltext-natural-language.html

esta forma se calcula la relevancia final de los términos en los documentos y con ello su peso específico.

– *Búsqueda en lenguaje natural con expansión de consulta*[9], proporciona un método de retroalimentación automática que permite concatenar la búsqueda original con los textos de los documentos más relevantes que hayan sido localizados en la primera consulta. Con esta información se genera una segunda búsqueda que logra recuperar el resto de documentos, hipotéticamente relevantes a la consulta planteada por el usuario. Este método tiende a generar bastante ruido, retornando documentos que probablemente no serían relevantes, a costa de un mayor porcentaje de documentos recuperados ó exhaustividad.

– *El umbral de términos aceptados*[10], determina la eliminación de aquellos cuya frecuencia de aparición supere el 50% de los documentos de la colección, considerándose en tal caso como una palabra vacía (SCHWARTZ, B. et al., 2008, pp.243-246). Esta medida sólo afecta a las consultas en lenguaje natural y beneficia el tratamiento de grandes colecciones, centrando la recuperación en aquellos documentos más relevantes en detrimento de una supuesta exhaustividad en la recuperación de muchos resultados potencialmente irrelevantes.

– *Búsqueda en modo booleano*[11], permite realizar búsquedas sobre el texto completo de la indexación de la colección, utilizando los operadores de sobre-ponderación, (SCHWARTZ, B. et al., 2008, pp.246-247). La relevancia se determina por el cumplimiento de los argumentos y condiciones de la consulta, generando como coeficiente, un número real positivo de no más de dos decimales de precisión.

A partir de estas especificaciones de consulta básicas, se han diseñado cinco algoritmos de clasificación, de los cuáles, tres han sido preparados para la categorización por precisión y los dos restantes para la categorización general de la colección de prueba, véase *tabla4*.

Tipo	Referencia	Algoritmo	Cadena de consulta	Condiciones
Clasificación temática de precisión	Prueba1	Consulta en lenguaje natural de los textos indexados en la colección de prueba, aplicando la intersección en modo booleano de los términos presentes en la cadena de búsqueda.	Bolsa de términos completa, separados por espacios para la cadena de consulta en lenguaje natural. Bolsa de términos completa con sobre-ponderación para la cadena de consulta en modo booleano.	– Términos con más de 7 caracteres para conformar la cadena de consulta en lenguaje natural. – Clasificación de documentos cuya similaridad sea un coeficiente igual o superior a 10.
	Prueba2	Consulta todos los términos de cada categoría temática en modo booleano sobre los textos indexados.	Bolsa de términos completa con sobre-ponderación para la cadena de consulta en modo booleano.	– Clasificación de documentos cuya similaridad sea un coeficiente igual o superior a 5.

[9] Full-Text Searches with Query Expansion. MySQL Ref. http://dev.mysql.com/doc/refman/5.1/en/fulltext-query-expansion.html
[10] Full-Text Search Functions. MySQL Ref. http://dev.mysql.com/doc/refman/5.1/en/fulltext-search.html
[11] Boolean Full-Text Searches. MySQL Ref. http://dev.mysql.com/doc/refman/5.1/en/fulltext-boolean.html

	Prueba3	Consulta en lenguaje natural con expansión de consulta de los textos indexados en la colección de prueba, aplicando la intersección en modo booleano de los términos presentes en la cadena de búsqueda.	Bolsa de términos completa, separados por espacios para la cadena de consulta en lenguaje natural con expansión de consulta. Bolsa de términos completa con sobre-ponderación para la cadena de consulta en modo booleano.	– Términos con más de 7 caracteres para conformar la cadena de consulta en lenguaje natural – Clasificación de documentos cuya similaridad sea un coeficiente igual o superior a 10.
Clasificación temática general	Prueba4	Consulta en modo booleano de los textos indexados en la colección de prueba.	Cadenas de consulta formadas a partir de combinaciones de 5 términos hasta agotar posibilidades del vocabulario de la categoría temática.	– Recuperación de todos los contenidos sin condiciones.
	Prueba5	Consulta en lenguaje natural de los textos indexados en la colección de prueba.	Cadenas de consulta formadas a partir de combinaciones de 5 términos hasta agotar posibilidades del vocabulario de la categoría temática.	– Recuperación de todos los contenidos sin condiciones.

Tabla 4. Cuadro de algoritmos diseñados para la prueba de clasificación automática

El algoritmo *prueba1*, clasifica todos los documentos cuyo valor de relevancia sea superior o igual a 10, equiparando mediante consulta booleana los términos de la categoría temática presentes en el título del documento y mediante consulta en lenguaje natural con respecto al campo de indexación principal de la colección. Las cadenas de consulta empleadas para el modo booleano, se conforman con todos los términos de la categoría temática, a los que se les asigna un operador de sobre-ponderación, para mejorar la precisión de los resultados, evitando un operador de intersección con el que se obtendrían escasos resultados, limitando excesivamente la relación entre exhaustividad y precisión. En cambio la cadena de búsqueda utilizada para la consulta en lenguaje natural se efectúa con todos los términos de la categoría temática cuya extensión supere los 7 caracteres, lo cual reduce significativamente el tamaño de la consulta. Esta modificación consigue una mayor precisión, dado que las palabras con mayor representatividad corresponden en muchos casos a las de mayor longitud de caracteres.

En el caso del algoritmo de *prueba2*, se clasifican todos los documentos cuyo coeficiente de relevancia sea superior a 5. Esto significa que de entre todas las condiciones que se establezcan en la consulta, al menos deben cumplirse cinco de ellas. El modo de consulta es booleano y se emplean todos los términos de cada categoría temática, con operador de sobre-ponderación. Este método de clasificación es el más exigente, pero no necesariamente el más preciso.

El algoritmo de *prueba3*, utiliza la misma metodología que en el algoritmo de *prueba1*, con la salvedad de que la búsqueda en lenguaje natural se amplía con expansión de consulta. Esta modificación ayudará a comprender, hasta qué punto la retroalimentación de resultados precisos, mejora la exhaustividad con resultados igualmente relevantes.

En relación al algoritmo de *prueba4*, se emplea el modelo booleano de consulta sobre el campo de indexación de la colección de prueba. La cadena de búsqueda utilizada se conforma mediante un proceso de combinaciones de términos para cada categoría

temática[12]. Dicho de otra forma, se combina el vocabulario del tesauro en grupos de 5 términos por categoría temática principal, generando millares de cadenas de consulta que se procesan en el sistema. Al contrario que en los casos anteriores, no se aplica ningún tipo de limitación en los coeficientes de relevancia que se obtengan, puesto que el objetivo que se persigue, tanto en el algoritmo 4 y 5, es la clasificación general de todos los contenidos de la colección.

El algoritmo de *prueba5*, utiliza el mismo método para generar las cadenas de consulta para el efecto categorizador de la colección. Su única diferencia es el modo de recuperación en lenguaje natural, que le confiere una mayor exhaustividad que el modelo booleano y una mayor precisión en los coeficientes de relevancia que se obtienen en los resultados.

La ejecución del proceso de clasificación para cada algoritmo, tiene como elemento común, que no se guardan todas las posibles clasificaciones aceptadas, sino aquella que mejores coeficientes de relevancia genera. El programa de clasificación prepara las consultas, practicando todas las combinaciones posibles de clasificación para cada algoritmo. De tal manera, se asegura que la categorización otorgada a cada documento es la mejor posible dentro de las posibilidades combinatorias del algoritmo.

La última fase de la investigación, se debe centrar en la evaluación de los algoritmos a través de los resultados obtenidos. El programa de clasificación permite generar diversos informes tabulados con los datos cuantitativos de los contenidos clasificados en cada categoría temática, tal como se muestra en la *tabla6*.

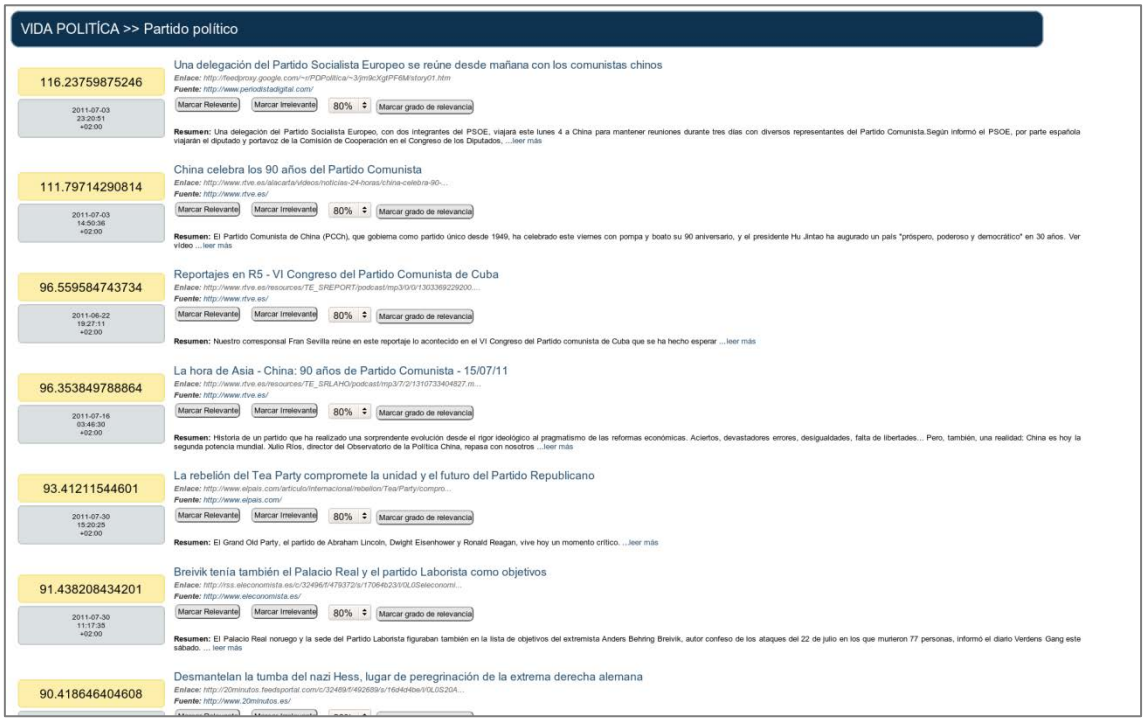

Figura 1. Formulario de evaluación de resultados

[12] El cálculo se efectúa con la fórmula de combinación básica $\left(C_{n,m} = \frac{n!}{m!(n-m)}\right)$

No obstante el estudio cuantitativo, también deberá ser acompañado de un análisis cualitativo. Dada la complejidad y extensión del trabajo que aquí se presenta, no se aborda un análisis cualitativo, quedando pendiente para la siguiente investigación. A pesar de ello, sí se han creado los mecanismos necesarios para que los resultados puedan ser evaluados por especialistas y usuarios, a modo de formulario, véase *figura1*. En este formulario se contempla el coeficiente de relevancia, la fecha de registro, el título, resumen, texto completo, enlace de la noticia y fuente de procedencia de cada documento clasificado según categoría temática. A cada resultado se le añade un mecanismo de evaluación para marcar un documento como relevante, irrelevante o con cierto grado de relevancia. De esta forma, cuando un usuario evalúa la corrección con que el sistema ha clasificado el documento, su puntuación, identificación y algoritmo son enviados y almacenados automáticamente en la plataforma Resync, para su posterior análisis. Esta información resulta importante para terminar de esclarecer cuál es el mejor algoritmo de clasificación y cuáles son los errores más frecuentes que serán tenidos en cuenta para su refinamiento.

Funcionamiento

El programa Resync, articula el mecanismo de clasificación automática en dos fases diferenciadas. La primera fase, denominada *de normalización*, véase *figura2*, carga el vocabulario de clasificación depurándolo con la eliminación y reemplazo de caracteres especiales, la normalización de textos y la eliminación de palabras vacías, hasta su almacenamiento final en base de datos. También se efectúa el proceso de depuración e indexación del corpus documental, suprimiendo etiquetados y rutinas de programación, distinguiendo el texto y normalizándolo. Al igual que el vocabulario de clasificación, se almacena en el campo de indexación previsto en el sistema de cara a su posterior recuperación.

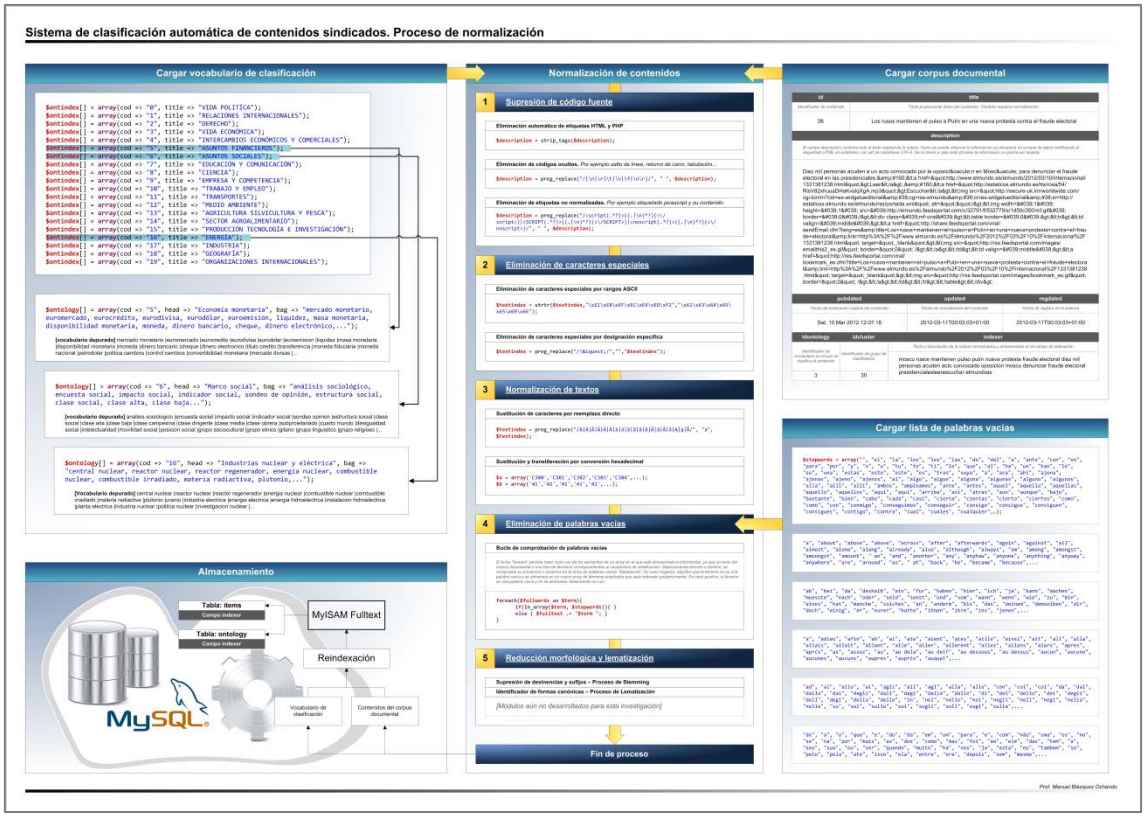

Figura 2. Primera fase de la clasificación: Normalización.
Disponible en: http://www.mblazquez.es/documents/ diagramas-mexico-2012_01.png

La segunda fase, denominada *de ejecución de consultas*, véase *figura3*, procesa todos los términos de cada categoría temática generando las consultas necesarias, tal como se especifica para cada algoritmo de clasificación. Este proceso supone un bucle de clasificación y reclasificación de cada combinación de términos de consulta, para todos los documentos de la colección. De esta forma se asigna una categoría por documento, que será reemplazada cuando exista otra, cuyo coeficiente de relevancia y por lo tanto de similaridad, sea mejor que la anterior. La información de la nueva clasificación, se almacena en la tabla del propio documento, asignando un identificador de la categoría con que fue clasificada, la referencia de la cadena de consulta utilizada en su caso y la puntuación obtenida.

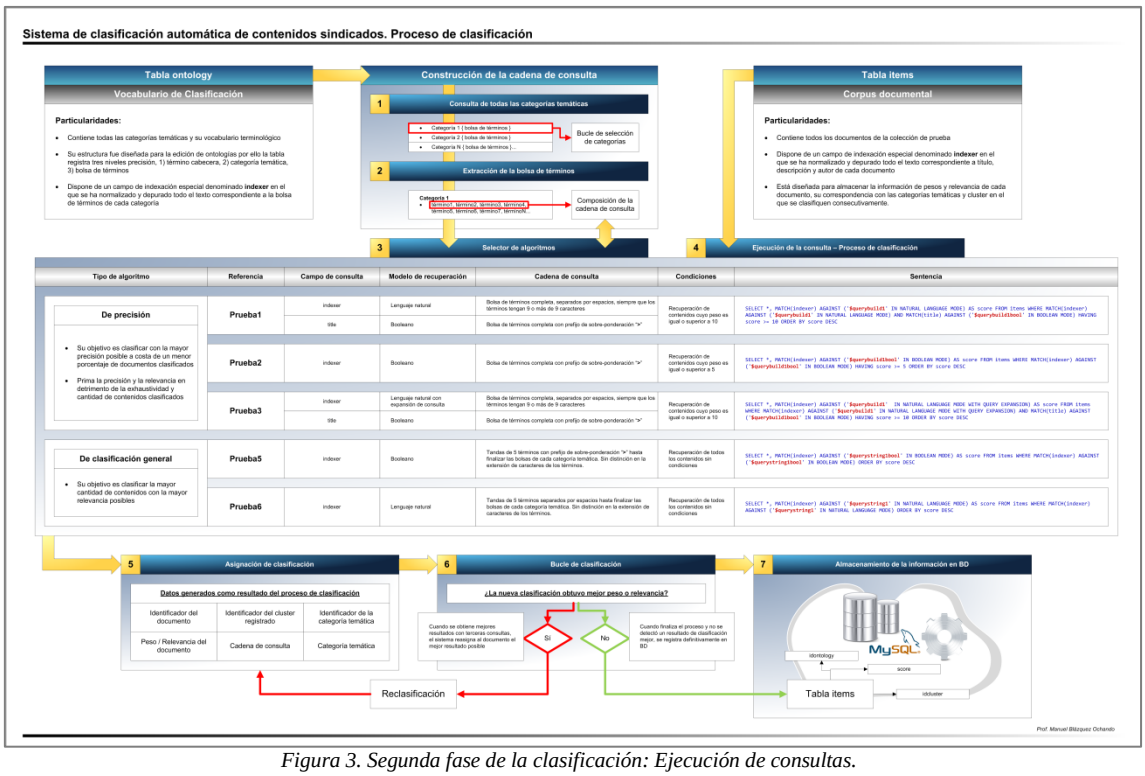

Figura 3. Segunda fase de la clasificación: Ejecución de consultas.
Disponible en: http://www.mblazquez.es/documents/ diagramas-mexico-2012_02.png

Resultados

Los resultados obtenidos demuestran que los algoritmos de precisión 1, 2 y 3, logran clasificar entre el 1% y el 25% del total de la colección, véase *tabla5*. Ello significa que el 75% restante no cumple los requerimientos mínimos que éstos plantean, restringiendo la clasificación a documentos que claramente resultan relevantes para las consultas planteadas. El resultado, también se debe a la manera en que se conforman las cadenas de consulta, con la totalidad de términos de cada categoría temática.

Referencia	**España**225.221		**México** 206.371	
	Contenidos	**Porcentaje**	**Contenidos**	**Porcentaje**
Prueba1	16.660	7,40%	22.264	10,79%
Prueba2	4.068	1,81%	2.605	1,26%
Prueba3	53.001	23,53%	51.983	25,19%
Prueba4	157.258	69,82%	158.575	76,84%
Prueba5	199.006	88,36%	204.644	99,16%

Tabla 5. Número total de contenidos clasificados

Se percibe una notable diferencia, cuando se clasifican contenidos en lenguaje natural, mediante el empleo del método de expansión de consulta. Éste incremento se estima en torno a 28.000 documentos más, tanto en el caso español como en el mexicano, lo que significa más del doble de lo recuperado, en la primera consulta del algoritmo de prueba1.

En cuanto a los algoritmos de clasificación general, se comprueba que el más completo es prueba5, con porcentajes cercanos y superiores al 90% para los contenidos producidos por medios de comunicación mexicanos y españoles. Este resultado, muy completo en términos cuantitativos, merece ser revisado cualitativamente, ya que el algoritmo se diseñó para clasificar sin limitaciones, lo cual explicaría el resultado obtenido. En cambio el algoritmo de prueba4, basado en la recuperación booleana, obtiene un porcentaje cercano al 70% mucho más realista si se tiene en cuenta que los resultados obtenidos, al menos deben contener uno de los 5 términos especificados, en alguna de sus múltiples combinaciones de consulta.

También resulta muy destacable que los contenidos publicados en medios mexicanos, se han clasificado mejor que los producidos por los medios españoles, con una diferencia media del 4,46%. Este fenómeno, puede ser debido a muchos factores, entre ellos destacan la variable idiomática con la que hipotéticamente se expresaría mejor el idioma, la sectorización de las noticias, mejor identificadas y descritas ó la extensión de los textos, marginalmente mayor que en el caso español.

En otro nivel del análisis, se encuentra la distribución de contenidos por categorías, según algoritmos y países, que puede consultarse en la siguiente *tabla6*. Se aprecia una correlación clara en la mayoría de contenidos clasificados, entre España y México.

Categoría temática	Prueba1		Prueba2		Prueba3		Prueba4		Prueba5	
	España	México	España	México	España	México	España	México	España	México
Vida Política										
Marco político	300	238	416	304	845	1894	1538	2014	951	710
Partido político	64	15	2	0	63	14	92	305	3908	4448
Procedimiento electoral	1146	827	144	32	1281	1032	1971	1339	5218	5087
Parlamento	654	202	86	3	795	223	610	341	2091	1389
Trabajos parlamentarios	12	35	0	0	33	38	117	65	826	629
Vida política y seguridad pública	344	274	319	159	415	538	2007	1547	5161	4599
Poder ejecutivo y administración pública	3246	1431	556	100	4252	1548	3704	3985	2910	3445
Relaciones Internacionales										
Política internacional	68	76	0	0	88	102	95	26	4008	4033
Política de cooperación	49	18	0	0	50	20	9	9	1430	1055
Equilibrio internacional	29	540	5	0	664	519	289	364	2677	1920
Defensa	51	105	2	2	70	226	137	611	2126	3600
Derecho										
Fuentes y ramas del Derecho	20	15	4	0	96	95	745	488	1113	1181
Derecho civil	284	359	41	5	1462	1115	1899	1190	1359	1628
Derecho penal	100	244	7	2	461	891	760	1103	2036	3042
Justicia	188	189	44	7	848	502	1017	672	1532	1554
Organización de la justicia	12	101	7	1	803	377	1271	659	684	698
Derecho internacional	287	387	1	0	335	564	377	480	1333	1556
Derechos y libertades	96	439	4	2	260	460	71	83	1822	2698
Vida Económica										
Política económica	10	425	0	0	439	404	245	250	1436	1358
Crecimiento económico	0	0	3	0	0	0	125	68	1271	876
Región y política regional	22	8	0	0	19	8	20	4	1358	1469
Estructura económica	15	2	0	0	16	2	552	221	3245	2120
Contabilidad nacional	7	65	0	0	345	294	257	172	1545	1934
Análisis económico	2	6	1	0	18	35	231	230	206	125
Intercambios económicos y comerciales										
Política comercial	114	610	7	0	1013	741	675	582	1541	1371
Política arancelaria	48	46	0	0	196	382	361	605	548	598
Intercambios económicos	208	499	7	0	749	500	867	524	2557	2307
Comercio internacional	13	18	0	0	14	18	0	3	1417	2037
Consumo	83	101	0	0	123	231	42	306	68	134
Comercialización	0	12	0	0	1	14	86	10	534	660
Distribución y venta	415	540	85	47	743	695	1191	1270	1570	1838
Asuntos financieros										

Relaciones monetarias	119	210	3	0	611	184	160	13	1822	1623
Economía monetaria	72	56	7	2	90	72	475	208	1510	1291
Instituciones financieras y de crédito	122	212	28	17	785	271	994	827	3165	1779
Libre circulación de capitales	53	2	0	0	24	8	514	103	1657	1304
Financiación e inversión	2	0	21	0	2	0	590	237	1719	991
Seguros	7	1	0	0	10	3	19	18	668	693
Hacienda pública	162	179	0	2	308	209	347	77	2570	1826
Presupuesto	45	12	0	0	45	12	33	177	1137	527
Fiscalidad	153	102	7	0	159	103	190	150	1970	1367
Precios	23	182	0	0	286	179	315	192	2969	1846
Asuntos sociales										
Familia	144	1049	1	2	656	1151	834	1261	2844	3799
Movimientos migratorios	0	93	0	0	0	99	120	161	259	532
Demografía y población	113	95	9	7	1373	1663	2415	2866	879	908
Marco social	32	38	0	0	87	95	229	134	1025	937
Vida social	653	1859	50	56	2158	3072	3537	4358	5270	6668
Cultura y religión	435	722	271	241	1558	1069	4446	2503	2503	1947
Protección social	171	328	0	0	656	331	726	957	2330	2298
Sanidad	355	493	78	63	484	675	1237	1275	3050	2919
Urbanismo y construcción	546	1030	19	0	1158	987	1834	2532	3970	4862
Educación y comunicación										
Educación	7	1	0	0	1	2	10	18	236	316
Enseñanza	85	27	0	0	188	148	352	63	1790	2659
Organización de la enseñanza	35	291	3	0	195	359	255	875	553	897
Documentación	256	328	22	8	903	631	2715	3109	3674	4571
Comunicación	547	601	69	46	1423	1304	4792	6254	5318	5553
Tratamiento de la información	8	7	0	0	28	61	1218	1571	1610	5661
Informática y tratamiento de datos	96	48	19	1	106	94	257	383	1031	1369
Ciencia										
Ciencias naturales y aplicadas	13	1	7	6	14	3	634	234	784	561
Humanidades	723	520	12	4	751	563	2096	1207	3470	2466
Empresa y competencia										
Organización de la empresa	114	58	7	7	317	210	1348	1085	1725	1571
Tipos de empresa	19	242	0	0	451	238	923	859	2429	1854
Forma jurídica de la sociedad	251	178	0	0	269	187	1071	555	1682	1203
Gestión administrativa	6	7	0	0	50	134	1273	629	901	861
Gestión contable	357	137	59	9	1114	633	4053	2333	1277	548
Competencia	137	98	1	3	145	133	470	568	973	1299
Trabajo y empleo										
Empleo	231	390	201	32	1493	1104	7354	4639	2926	2210
Mercado laboral	63	81	45	10	1473	1422	2594	1967	1879	2051
Condiciones y organización del trabajo	95	208	213	103	1639	2300	18356	11086	1991	1360
Administración y remuneración	240	213	28	16	522	225	2027	4150	929	890
Relaciones laborales	120	160	11	4	534	242	2464	1480	1305	1255
Transportes										
Política de transportes	7	24	0	0	131	236	441	422	1386	2145
Organización de los transportes	36	164	4	2	490	265	1119	1647	2886	3200
Transporte terrestre	12	25	0	0	111	175	1750	958	606	429
Transporte marítimo y fluvial	19	54	0	0	65	101	225	258	1431	903
Transporte aéreo y espacial	28	27	0	0	115	176	913	610	644	1070
Medio Ambiente										
Política del medio ambiente	70	22	0	0	75	23	337	130	3480	5250
Medio natural	67	736	25	59	687	985	2192	2068	1389	1296
Deterioro del medio ambiente	276	81	1	0	704	258	1214	892	943	1148
Agricultura, Silvicultura y Pesca										
Política agraria	0	12	0	0	0	12	14	35	751	581
Producción y estructuras agrarias	0	0	0	0	0	0	6	1	706	467
Sistema de explotación agraria	11	11	0	0	17	12	85	36	841	687
Explotación agrícola de la tierra	0	1	0	0	7	7	85	144	948	1200
Medio de producción agrícola	1	20	0	0	161	202	1031	440	240	236
Actividad agropecuaria	1	10	1	0	13	23	99	132	1028	946
Monte	5	30	21	8	84	29	494	395	943	745
Pesca	26	221	4	0	95	237	261	358	1116	1062
Sector Agroalimentario										
Productos de origen vegetal	2	30	10	9	86	114	813	589	626	399
Productos de origen animal	0	149	0	0	152	191	451	261	601	416
Productos agrarios transformados	4	49	3	0	129	65	249	122	816	591
Bebidas y azúcares	20	111	10	9	299	217	822	684	712	611
Productos alimenticios	15	0	0	0	6	28	40	44	582	428
Industria agroalimentaria	0	2	0	0	0	2	14	7	65	64
Tecnología alimentaria	1	0	0	0	1	0	104	49	251	374
Producción, Tecnología e Investigación										
Producción	5	2	2	2	21	22	832	844	782	1086
Tecnología y reglamentación técnica	22	58	0	1	228	421	2537	2069	1890	2391
Investigación y propiedad intelectual	69	126	17	7	478	312	4767	3203	2388	2164
Energía										
Política energética	25	63	0	0	24	68	218	216	787	551
Industrias carbonera y minera	0	8	0	0	16	9	119	62	164	254
Industria petrolera	78	99	2	2	161	114	483	420	345	378
Industrias nuclear y eléctrica	81	27	1	2	82	32	278	198	505	431
Energía blanda	1	0	0	1	1	0	48	43	624	615
Industria										
Política y estructura industriales	11	8	0	0	13	19	215	60	1374	1570
Química	10	61	0	0	54	77	449	328	404	548
Metalurgia y siderurgia	7	37	4	2	261	194	953	409	488	305
Industria mecánica	5	21	0	0	147	144	1542	883	2200	2028
Electrónica y electrotécnica	0	0	0	0	0	0	8	11	411	504

Construcción y obras públicas	94	58	0	0	178	177	1341	1081	824	1129
Industria de la madera	7	3	0	0	182	50	1135	645	374	535
Industria del cuero e industria textil	0	12	0	0	23	37	256	194	209	152
Industrias diversas	0	0	0	0	12	63	163	226	1352	637
Geografía										
Europa	224	134	590	136	5008	3329	13765	5574	1766	1146
América	580	902	319	966	1818	5901	8737	38893	2082	3685
África	85	119	76	50	687	807	3414	2312	1171	1122
Asia	60	91	35	43	665	1513	3918	4579	565	807
Oceanía	59	121	1	1	108	249	687	887	515	597
Oriente medio	2	20	9	2	427	667	1315	1633	251	282
Organizaciones Internacionales										
Geografía económica	0	0	0	0	0	0	414	144	1968	1922
Geografía política	0	0	0	0	0	0	0	0	1169	954
Naciones Unidas	49	99	0	0	98	120	965	834	7066	8588
Organizaciones europeas	5	0	0	0	12	0	200	94	1542	989
Organizaciones extraeuropeas	12	44	0	0	13	53	59	266	3868	4068
Organizaciones mundiales	17	195	1	0	239	451	358	1051	2048	2826
Organizaciones no gubernamentales	54	91	0	0	54	138	685	494	602	761

Tabla 6. Resultados cuantitativos detallados de la clasificación temática

Resultan coincidentes las categorías con más contenidos clasificados, tanto en el caso español como en el mexicano, a lo largo de todos los algoritmos de clasificación. Pueden destacarse los contenidos relativos al poder ejecutivo y la administración pública, el procedimiento electoral y el sistema de votación, el marco político y las noticias relativas a condiciones y organización del trabajo. Se detecta además una clara diferenciación de noticias según regiones geográficas, por un lado las noticias sobre países europeos son más numerosas en el caso español que en el mexicano y a la inversa, las noticias sobre países americanos son más cuantiosas en los medios mexicanos que en los españoles. También se demuestra que el algoritmo de prueba2 es el más restrictivo al acumular 70 categorías huérfanas en el caso mexicano y 57 en el caso español, de 124 disponibles, obtenidas originalmente en el tesauro europeo.

Conclusiones

1. El sistema de clasificación automática diseñado para la plataforma Resync hace posible la organización de los contenidos recuperados. También queda patente que el porcentaje de contenidos susceptibles de ser clasificados por el sistema, varía según las especificaciones del diseño de los algoritmos. Esto es, la delimitación entre el factor de precisión "*algoritmos de prueba 1, 2 y 3*" y el de cobertura "*algoritmos de prueba 4 y 5*", con el que se pretenda clasificar en cada caso.

2. Tiene sentido la obtención de métodos de clasificación más exigentes en los que prime la categorización más precisa posible, con la intencionalidad de obtener contenidos con un alto grado de relevancia y representatividad, como en las pruebas 1, 2 y 3 cuyos resultados de clasificación, oscilan entre un 1% y un 25%. Estos documentos, pueden ser los primeros que el usuario visualice, antes de acceder a contenidos cuya clasificación sea más imprecisa, aumentando de tal manera las posibilidades de relevancia y satisfacción de la consulta de cara al usuario.

3. Por otro lado, también parece necesario, el desarrollo de algoritmos de clasificación general, con un alto porcentaje de categorización de documentos, situado en torno al 70 y el 90% de la colección, en las pruebas 4 y 5. Aunque sus resultados no sean comparables al método de precisión, son necesarios para posibilitar la asignación de nuevos puntos de acceso a los contenidos de la colección. Su principal ventaja es el aporte de exhaustividad, complemento necesario de los resultados más precisos, cuando el usuario no logra satisfacer su demanda informativa.

4. Los contenidos publicados por los medios de comunicación mexicanos se clasifican mejor que los españoles, con una diferencia media del 4,46%. Este resultado se obtiene con una producción de contenidos muy similar.

5. Se refrenda que los lenguajes documentales tienen un papel fundamental en el proceso de clasificación automática, incluyendo los tesauros, ya que pueden ser transformados fácilmente en pseudo-ontologías, sobre las que construir combinaciones de patrones de consulta.

6. Los contenidos publicados por los medios de comunicación mexicanos y españoles resultan parejos en su categorización temática con los algoritmos de prueba 1, 2 y 3. Ello se desvela cuando se comprueba que entre las siete primeras temáticas con más noticias clasificadas, varías de ellas coinciden recurrentemente, a saber: noticias relativas a la administración pública, el poder ejecutivo, el procedimiento electoral, la vida social y política, el marco político y la organización del trabajo.

Bibliografía

ARSLAN, A. and O. YILMAZEL. 2010. Quality Benchmarking Relational Databases and Lucene in the TREC4 Adhoc Task Environment. *En: Proceedings of the International Multiconference on Computer Science and Information Technology*. Wisla: IEEE, pp.365-372. Disponible en: http://www.proceedings2010.imcsit.org/pliks/139.pdf

BERRY, M.W. and M. BROWNE. 2005. Document file preparation. *En: Understanding search engines: mathematical modeling and text retrieval*, Philadelphia: Siam, pp.11-27.

BLÁZQUEZ OCHANDO, M. and E. SERRANO MASCARAQUE. 2011. Plataforma para la investigación de contenidos sindicados: desarrollo del sistema ReSync y aplicación a los medios de comunicación hispano-mexicanos. *En: VIII Seminario Hispano Mexicano de Biblioteconomía y Documentación*. Madrid: Universidad Complutense de Madrid.

CAMOUS, F., S. BLOTT, and A.F. SMEATON. 2007. Ontology-based MEDLINE document classification. *En: BIRD07 Proceedings of the 1st international conference on Bioinformatics*. Heidelberg: Springer, pp.439-452. Disponible en: http://doras.dcu.ie/258/1/lncs_4414.pdf

COHEN, A. M. and W. R. HERSH. 2006. The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *Journal of Biomedical Discovery and Collaboration*. **1**(4). Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/16722582

CUNNINGHAM, H., K. BONTCHEVA, V. TABLAN et al. 2012. *GATE: General Architecture for Text Engineering*. [online]. [Consultado 1 Abril 2012]. Disponible en: http://gate.ac.uk/

DAEDALUS. 2011. *Clasificación Automática Eurovoc*. [online]. [Consultado 15 Feb 2012]. Disponible en: http://showroom.daedalus.es/es/tecnologias-de-la-lengua/eurovoc/

DEXTRE CLARKE, S.G. 2010. ISO 25964 - the new standard for thesauri and interoperability with other vocabularies. *En: EuroVoc Conference: Mind the lexical gap*. Luxembourg. Disponible en: http://eurovoc.europa.eu/drupal/sites/all/files/conference2010/EuroVocConference_ISO25964preview.ppt

GELERNTER, J. 2008. Data Mining of Maps and their Automatic Region - Time - Theme Classification. *En:* SigSpatial SPECIAL, (ed). *International Conference on Advances in Geographic Information Systems*. Irvine. Disponible en: http://www.cs.cmu.edu/~gelernter/sigspatial.pdf

HART, P. and T. COVER. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. **13**(1), pp.21-27. Disponible en: http://www.stanford.edu/~montanar/TEACHING/Stat319/papers/cover_nn.pdf

NIST. 2011. *Multilingual Automatic Document Classification and Translation Evaluation Program*. [online]. [Consultado 1 Abril 2012]. Disponible en: http://www.nist.gov/itl/iad/mig/madcat.cfm

PRABOWO, R., M. JACKSON, P. BURDEN, and H. KNOELL. 2002. Ontology-Based Automatic Classification for the Web Pages: Design, Implementation and Evaluation. *En: Proceedings of the 3rd International Conference on Web Information Systems Engineering WISE02*. Washington: IEEE Computer Society. Disponible en: http://maya.cs.depaul.edu/~mobasher/research/bib/bib-papers/PJBK02.pdf

ROBERTSON, S. 2004. Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation*. **60**(5), pp.503-520. Disponible en: http://www.soi.city.ac.uk/~ser/idfpapers/Robertson_idf_JDoc.pdf

SÁNCHEZ JIMÉNEZ, R. 2007. La documentación en el proceso de evaluación de sistemas de clasificación automática. *Documentación de las ciencias de la información*., pp.25-44. Disponible en: http://revistas.ucm.es/index.php/DCIN/article/view/DCIN0707110025A/18959

SCHWARTZ, B., P. ZAITSEV, V. TKACHENKO et al. 2008. High Performance MySQL. *En: Natural-Language Full-Text Searches*, Sebastopol: OReilly, pp.244-256.
STEINBERGER, R. 2010. Automatic Eurovoc indexing of parliamentary texts. *En: EuroVoc Conference: Mind the lexical gap*. Luxembourg.

SUBRAMANIAM, V., D. PUNJANI, and S. MUKHERJEA. 2005. Biomedical Document Triage: Automatic Classification Exploiting Category Specific Knowledge. *En: TREC Conference Proceedings*. Gaithersburg. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.946&rep=rep1&type=pdf