Biblios
ISSN 1562-4730 (online)

Minería de texto en la clasificación de material bibliográfico

Marcial Contreras Barrera

Universidad Nacional Autónoma de México – UNAM, México

ARTICLE

Resumen

Objetivo: Desarrollar un clasificador automatizado para la clasificación de material bibliográfico por medio de la minería de texto. Metodología: La minería de texto es empleada para el desarrollo del clasificador, basado en un método de tipo supervisado conformado por dos etapas; de aprendizaje y reconocimiento, en la etapa de aprendizaje, el clasificador aprende patrones a través del análisis de registros bibliográficos, de la clasificación Z, del área de bibliotecología, ciencias de la información y recursos de información recuperados de la base de datos LIBRUNAM , en esta etapa se obtiene el clasificador capaz de reconocer diferentes subclases (LC). En la etapa de reconocimiento el clasificador se valida y evalúa a través de pruebas de clasificación, para este fin se toman aleatoriamente registros bibliográficos de la clasificación Z, clasificados por un catalogador y procesados por el clasificador automatizado, con el fin de obtener la precisión del clasificador automatizado. Resultados: La utilización de la minería de texto permitió el desarrollo del clasificador automatizado, a través del método de clasificación de documentos de tipo supervisado. La precisión del clasificador fue calculada realizando la comparación entre los temas asignados de manera manual y automatizada obteniendo un grado de precisión del 75.70%. Conclusiones: La aplicación de la minería de texto facilito la creación del clasificador automatizado, permitiendo obtener tecnología útil para la clasificación de material bibliográfico con la finalidad de mejorar y agilizar el proceso de organización de información.

Palabras clave

Minería de texto ; Clasificación ; Clasificador automatizado ; Material bibliográfico

Text mining in the classification of digital documents

Abstract

Objective: Develop an automated classifier for the classification of bibliographic material by means of the text mining. Methodology: The text mining is used for the development of the classifier, based on a method of type supervised, conformed by two phases; learning and recognition, in the learning phase, the classifier learns patterns across the analysis of bibliographical records, of the classification Z, belonging to library science, information sciences and information resources, recovered from the database LIBRUNAM, in this phase is obtained the classifier capable of recognizing different subclasses (LC). In the recognition phase the classifier is validated and evaluates across classification tests, for this end bibliographical records of the classification Z are taken randomly, classified by a cataloguer and processed by the automated classifier, in order to obtain the precision of the automated classifier. Results: The application of the text mining achieved the development of the automated classifier, through the method classifying documents supervised type. The precision of the classifier was calculated doing the comparison among the assigned topics manually and automated obtaining 75.70% of precision. Conclusions: The application of text mining facilitated the creation of automated classifier, allowing to obtain useful technology for the classification of bibliographical material with the aim of improving and speed up the process of organizing digital documents.

Keywords

Text mining ; Classification ; Automated classifier ; Bibliographic material

1 Introducción

Para Zhang y Gu (2011), el 90% de la información disponible se encuentra de forma no estructurada y semi estructurada, almacenada en computadoras o sistemas de almacenamiento, dificultando su búsqueda y consulta a través de los sistemas de recuperación de información, por lo que es necesario el uso de la tecnología para facilitar su análisis, con la finalidad de agilizar su organización en beneficio de los usuarios de información.

Diversas investigaciones tienen como meta aportar soluciones a los problemas del procesamiento y organización de información digital, desarrollando métodos enfocados a facilitar y agilizar actividades tales como la indización de documentos, creación automatizada de tesauros, identificación de semejanza entre documentos, categorización, clasificación, generación de resúmenes automáticos, búsqueda de información en texto completo, extracción de información, identificación de relaciones y términos, por mencionar algunas. Las investigaciones proponen métodos y técnicas basados en; reconocimiento de patrones, aprendizaje de máquina, métodos estadísticos, técnicas de procesamiento de Lenguaje Natural, y métodos de redes neuronales artificiales.

Los métodos son desarrollados y empleados por áreas como la minería de datos, la minería de texto y la minería Web. Dichas áreas hacen referencia al hecho de que a partir de bases de datos, documentos de texto y páginas web se puede extraer información relevante a partir de un conjunto de datos. Según Swanson (1991) con la aplicación de la minería de texto se pueden identificar o descubrir patrones y nuevo conocimiento a partir de una colección de textos, es decir, la minería de texto es el proceso encargado del descubrimiento de conocimientos que no existía explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos. Por lo que la minería de texto es utilizada para el análisis de información de tipo no estructurado o semi estructurado para encontrar información relevante.

La minería de texto se encarga de aprovechar y desarrollar métodos automatizados para procesar documentos digitales no estructurados de forma rápida y eficiente, con la finalidad de organizarlos y analizarlos. Por lo que en este estudio se emplea la minería de texto en el procesamiento de documentos digitales para ser usada en la clasificación de documentos no estructurados partiendo de la siguiente hipótesis: La minería de texto facilita el procesamiento de documentos digitales por medio de los métodos de extracción de información y clasificación, permitiendo la organización de documentos.

La presente investigación tiene como objetivo, desarrollar un clasificador automatizado para la clasificación de material bibliográfico por medio de la minería de texto. La minería de texto es empleada para el desarrollo del clasificador, basado en un método de tipo supervisado, conformado por dos etapas; de aprendizaje y reconocimiento, en la etapa de aprendizaje, el clasificador aprende patrones a través del análisis de registros bibliográficos, pertenecientes a la clasificación Z, del área de bibliotecología, ciencias de la información y recursos de información, recuperados de la base de datos LIBRUNAM. En esta etapa de aprendizaje el clasificador es capaz de reconocer diferentes subclases. En la etapa de reconocimiento el clasificador se valida y evalúa a través de pruebas de clasificación, para este fin se toman aleatoriamente registros bibliográficos de la clasificación Z, clasificados por un catalogador y procesados por el clasificador automatizado, con el fin de obtener la precisión del clasificador automatizado. El trabajo se encuentra estructurado de la siguiente forma: introducción, minería de texto y clasificación de documentos, metodología de la minería de texto, aplicación de la minería de texto en la base de datos LIBRUNAM, resultados y conclusiones.

2 Minería de texto y clasificación de documentos

La clasificación de documentos es un proceso utilizado en la organización de la información con el fin de facilitar su búsqueda y recuperación. De acuerdo con (Lévano, 2011) la clasificación permite la agrupación de los documentos sobre un mismo tema a partir de características similares, donde un documento puede pertenecer a una sola clase o a varias clases. La clasificación se puede realizar de manera manual o automatizada; la clasificación automatizada es realizada por sistemas de cómputo y se divide en dos etapas, de entrenamiento y reconocimiento. En la etapa de entrenamiento se utiliza un conjunto de documentos para la construcción del modelo y en la etapa de reconocimiento se realiza la clasificación.

La clasificación en la gestión de documentos es de suma la importancia por lo que diversas investigaciones se realizan con la finalidad de obtener métodos computacionales que ayuden en dicha gestión. Actualmente existen diferentes métodos como; Support Vector Machine (SVM), aprendizaje de máquina, Naive Bayes, Regresión Logística, máxima entropía, entre otros.

Xiu-Li, Feng, & Jiang (2007) utilizan el método de máxima entropía (ME) con la finalidad de mejorar la precisión en la clasificación de documentos, obteniendo los resultados mostrados en la tabla 1.

Tabla 1 - Resultados obtenidos en el método de máxima entropía

Method	Precision
Baseline (ME)	66.50%
ME + MI	65.17%
ME + AMI	67.67%
ME + CE	69.28%

Nota. Fuente: (Xiu-Li, Feng, & Jiang, 2007). An improved document classificaction approach with maximum entropy and entropy feauture selection. 2007 International Conference on Machine Learning and Cybernetics. Hong Kong: IEEE.

Abdullah (2014) realiza un clasificador de documentos basado en el MeSH (Medical Subject Headings), obteniendo un porcentaje de precisión del 60%. Wang (2010) aplica el método de Kernel Logistic Regression para realizar la clasificación y utiliza el método Fisher discriminant analysis (LFDA), para extraer las características de los documentos obteniendo los resultados mostrados en la tabla 2.

Tabla 2 - Precisión del método "Logistic Regression"

Algorithm	Accuracy
KNN	68.1%
LSI-KNN	71.5%
LFDA-LR	78.2%
LFDA-KLR	83.7%

Nota. Fuente: (Wang, 2010). Document Classification Algorithm Based on Kernel Logistic Regression. Industrial and Information Systems (IIS), 2010 2nd International Conference on (Volume:1) (págs. 76 - 79). Dalian : IEEE.

El desarrollo de diferentes métodos en la clasificación de documentos y su aplicación, tiene como una de sus principales metas la clasificación de documentos con la mayor precisión posible.

3 Metodología de la minería de texto

La metodología empleada para realizar la minería de texto puede ser general o específica, una metodología general como la propuesta por Verma, Ranjan & Mishra (2015) se define en dos fases; la fase de refinación del texto donde los documentos son transformados y representados en estructuras de datos; y la fase llamada destilación del conocimiento, donde se identifican patrones o conocimiento a partir de las estructuras de datos.

La metodología más específica propuesta por M.Sukanya (2012), establece una serie de etapas o pasos descritos a continuación: En el primer paso se determina el propósito de estudio de la minería de texto, por ejemplo, en el caso de textos de biología, se pueden identificar y etiquetar entidades biológicas, extraer y normalizar sinónimos, homónimos y abreviaturas, generar hipótesis, etc.

En el segundo paso se recolecta, identifica y valida información. En esta fase se realiza la recuperación de información (IR), en la que se buscan e identifican las fuentes más relevantes, para el objetivo de estudio de la minería de texto. A continuación se recopilan los documentos detectados en el mejor formato, se seleccionan, se evalúa su relevancia, y se realizan las anotaciones necesarias. En esta etapa, se cuenta con el conjunto de documentos necesarios para la realización de la minería de texto.

Durante el tercer paso, se realiza el procesamiento de texto eliminando datos que no ayudan al propósito de la minería de texto, realizándose algunas de las siguientes acciones: análisis léxico, tratamiento y separación de palabras vacías,(artículos, preposiciones, conjunciones), tratamiento de términos flexionados (términos relacionados morfológicamente, variaciones de género, número o tiempo verbal), tratamiento de palabras compuestas, normalización de palabras, obtención de las raíces de las palabras y etiquetado de palabras, además de corregir algunos problemas que presenten los documentos como: los problemas de formato, la polisemia, homonimia, sinonimia. Este paso exploratorio se basa principalmente en lingüística computacional

(análisis morfológico y sintáctico), además de algoritmos informáticos (Abbott, 2013). Su objetivo es facilitar la selección de características deseadas, para identificar palabras clave, identificación de entidades, individuos, organizaciones, lugares, oraciones, conceptos, etc.

En el cuarto paso se realiza la extracción y análisis de clases, relaciones, asociaciones o secuencias, con el fin de encontrar evidencias de conceptos y de estructuras existentes. En esta etapa los documentos se pueden representar a través del modelo del espacio vectorial (Salton, 1989). En donde cada documento es modelado como un vector de dimensión **n** y es representado de la forma. $D_i = (d_{i1}, d_{i2}, \ldots, d_{it})$, donde cada d_{ij} representa el número de repeticiones de la palabra en el documento. Los datos obtenidos en esta etapa son representados en alguna estructura informática que facilita su análisis, las estructuras representan las relaciones entre las entidades de un mismo tipo de datos, palabras o conceptos clave, documento-términos, términos-autores, etc.

En el último paso se presentan los resultados, a través de resúmenes, marcados de texto, relaciones, taxonomías y visualización, para su interpretación. También se puede almacenar la información procesada en bases de datos para su recuperación posterior. De manera resumida la figura 1 muestra la metodología de la minería de texto.

Figura 1. Metodología de la minería de texto.

3.1 Aplicaciones de minería de texto en bibliotecas

La minería de texto puede ser aplicada en el análisis de las bitácoras de los catálogos en línea OPAC (*Online public access catalog*), para identificar los índices de búsqueda más utilizados en el catálogo (título, autor, temas, otros), con el propósito de mejorar la presentación de los catálogos.

También el análisis de las bitácoras es útil para identificar el uso de las temáticas más frecuentes, permitiendo mejorar el desarrollo de colecciones de las bibliotecas.

Así mismo se pueden analizar las bases de datos de circulación las cuales registran el préstamo de libros en las bibliotecas, con la determinación de estudiar el comportamiento de uso de las colecciones bibliográficas en relación al usuario, y así identificar las necesidades de información de los usuarios.

4 Aplicación de la minería de texto en la base de datos LIBRUNAM

Desde el punto de vista de la minería de texto la base de datos de LIBRUNAM representa una base de datos de conocimiento y por lo tanto una fuente de conocimiento debido a que en el tiempo de existencia acumula el conocimiento del personal, manifestado en los procesos de clasificación y catalogación bibliográfica. Para el actual estudio se extrae ese conocimiento y se aplica en la realización del clasificador automatizado.

En el presente apartado se aplica la minería de texto y el análisis de los registros bibliográficos de la base de datos LIBRUNAM, en el desarrollo del clasificador automatizado, basado en la clasificación Library of Congress (LC), y su posterior aplicación en la asignación temática en documentos digitales. En el desarrollo del clasificador solo se toma en consideración la clasificación Z y algunas de sus subclases. A continuación se describen las etapas de la minería de texto.

4.1 Etapa 1.- Propósito de la minería de texto

Aplicar la minería de texto en el desarrollo de un clasificador automatizado para realizar la clasificación de documentos pertenecientes al área de bibliotecología, ciencias de la información y recursos de información (clasificación Z), como se muestra en la figura 2.

FIGURA 2. CLASIFICACIÓN AUTOMATIZADA DE DOCUMENTOS.

4.2 Etapa 2.- Recuperación de la Información

La recuperación de la información inicia con la identificación de las fuentes de información, en este caso es la base de datos LIBRUNAM en la cual se buscan, recuperan, y seleccionan los registros bibliográficos de la clasificación **Z,** correspondiente a bibliotecología, ciencias de la información y recursos de información. El número total de registros[i] en esta clasificación es de 23,630 y un total de 280 subclases. Un segmento de la frecuencia de las clasificaciones es mostrado en la gráfica 1

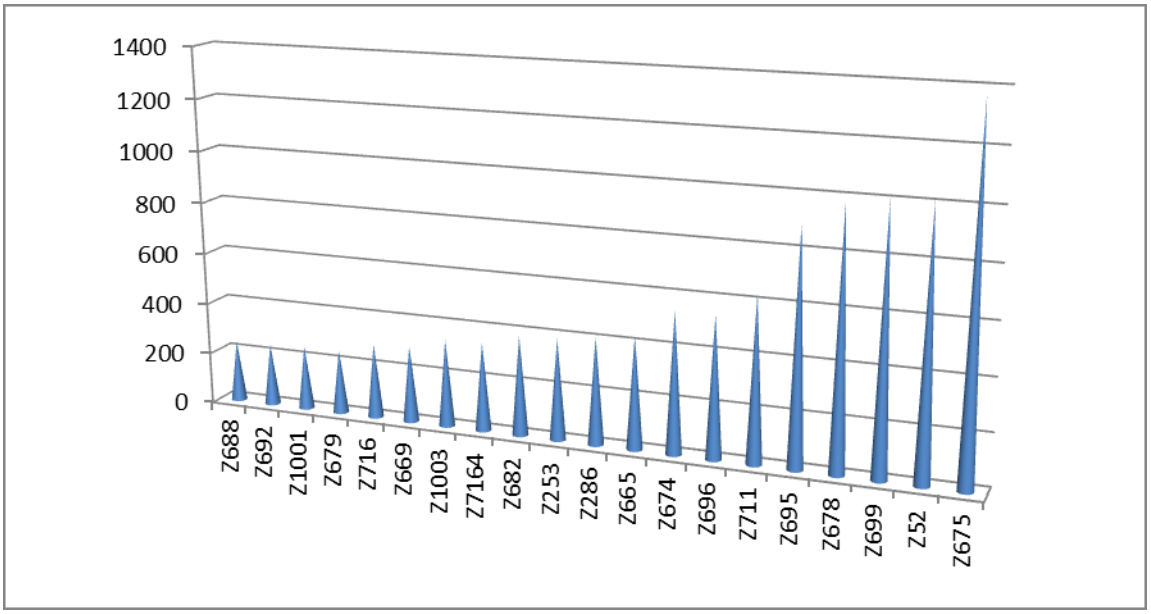

Gráfico 1 - Frecuencia de las subclases en los registros bibliográficos.

4.3 Etapa 3.- Extracción de la información

En esta etapa se generan archivos con cada una de las subclases de la clasificación LC correspondiente a bibliotecología, para automatizar esta tarea se desarrolló un programa de computadora para separar los archivos en subclases. La formación de subclases del clasificador tuvo como criterio considerar los registros con frecuencia mayor a 50 pertenecientes a una misma subclase, reduciendo el número de clases en estudio de 280 a 69.

4.4 Etapa 4.- Procesamiento de texto

En esta etapa se realiza la limpieza de datos y la generación de archivos formados por las subdivisiones de la clase Z. Además los registros MARC son transformados a un formato sencillo de manipular por la computadora. El registro bibliográfico compuesto por las etiquetas MARC, número de sistema, 050(clasificación LC), 245(título), 650(tema), son transformados a su representación del espacio vectorial, además de realizar el cálculo de frecuencias y la creación de la matriz de adyacencia con los lexemas más representativos de los temas de los registros bibliográficos.

4.5 Etapa 5.- Aplicación del método de minería de texto

El método de minería de texto empleado para realizar la clasificación, se divide en dos fases, de aprendizaje y de reconocimiento. En la fase de aprendizaje el método identifica patrones a través de un conjunto de documentos y en la fase de reconocimiento se realizan pruebas de clasificación, para validar la precisión.

El clasificador fue desarrollado extrayendo y analizando las diferentes clasificaciones LC existentes en la base de datos LIBRUNAM (etiqueta MARC 50), y a partir de la etiqueta, se identificaron las subclases, cada subclase quedo representada por archivos conteniendo la etiqueta MARC 650 de temas.

Realizando la subdivisión de clases se calcularon 280 subdivisiones pero por cuestiones de representación y de procesamiento, se consideraron solo aquellas en las que el número de títulos fuera mayor a 50, dando como resultado solo 69 subclases. Finalmente el clasificador automatizado queda representado por la matriz de las subclases y el conjunto de temas que describen esa clase por medio de la matriz de la tabla 3. La subclase con más registros en la base es la Z675 correspondiente a tipo de bibliotecas, seguida de procesador de palabras perteneciente a la Z52 y la Z699 relacionada a recuperación de la información.

Row No.	label	metadata_file	Administrac...	América	Archivos	Arte	Automatizac...	Automatizac...	Biblioteca	Bibliotecarios	Bibliotecolo...
3	Bibliotecología	Z1006	0	0	0	0	0	0	0	0	0.013
4	Bibliotecología	Z1008	0	0	0	0	0	0	0.081	0	0
5	Bibliotecología	Z1035	0	0.008	0	0	0	0	0	0	0
6	Bibliotecología	Z1037	0	0	0	0	0	0	0	0	0
7	Bibliotecología	Z116	0	0	0	0	0	0	0	0	0
8	Bibliotecología	Z1209	0	0.974	0	0.150	0	0	0	0	0
9	Bibliotecología	Z1426	0	0	0.707	0.707	0	0	0	0	0
10	Bibliotecología	Z1427	0	0	0.894	0	0	0	0	0	0
11	Bibliotecología	Z1601	0	0.630	0	0	0	0	0.126	0	0
12	Bibliotecología	Z244	0	0	0	0	0	0	0	0	0
13	Bibliotecología	Z246	0	0	0	0.013	0	0	0	0	0
14	Bibliotecología	Z250	0	0	0	0.051	0	0	0	0	0
15	Bibliotecología	Z253	0	0	0	0	0	0	0	0	0
16	Bibliotecología	Z271	0	0	0	0	0	0	0	0	0
17	Bibliotecología	Z278	0	0	0	0	0	0	0	0	0
18	Bibliotecología	Z286	0.004	0.004	0	0.004	0	0	0	0	0
19	Bibliotecología	Z4	0	0	0	0	0	0	0	0	0
20	Bibliotecología	Z43	0	0	0	0	0	0	0	0	0
21	Bibliotecología	Z5055	0	0	0	0	0	0	0	0	0
22	Bibliotecología	Z52	0	0	0	0	0	0	0	0	0
23	Bibliotecología	Z5814	0	0.040	0	0	0	0	0	0	0
24	Bibliotecología	Z6621	0	0.224	0	0	0	0	0.447	0	0
25	Bibliotecología	Z665	0	0.159	0.096	0.032	0	0	0	0.064	0.638
26	Bibliotecología	Z666	0	0	0.091	0	0	0	0	0	0.730
27	Bibliotecología	Z668	0.020	0	0	0	0	0	0	0.054	0.013

Tabla 3. Matriz con las palabras más representativas de cada subclase

La asignación de los temas de las subclases se realizó por medio del método de minería de texto RAKE[ii]. El método RAKE facilita la identificación del tema a través del cálculo de frecuencia de los términos más relevantes ocurridos en las etiquetas 650 dando como resultado las temáticas mostradas (columna de temas) en la tabla 4.

Frecuencia	Clase	Tema
385	Z682	Bibliotecarios, Salarios, Profesión, Oportunidades
391	Z253	Pagemaker, Newpaper, Diseño de revistas
412	Z286	Electronic Publisher
435	Z665	Information science, Bibliotecología, Biblioteconomía
543	Z674	Servicios de información
546	Z696	Clasification
634	Z711	Reference service, Information services
902	Z695	Tesaurus
987	Z678	Bibliotecas, Administración de bibliotecas, Library management
1008	Z699	Information retrieval science
1020	Z52	Word perfect, Windows, Academic libraries
1376	Z675	Special libraries, University libraries

Tabla 4. Asignación de temas a las subclases.

En la etapa de reconocimiento, se realizan las pruebas necesarias con la finalidad de validar la precisión del clasificador propuesto. Un documento es asignado a una subclase si el resultado de la fórmula del coseno o sea la similitud entre documentos cumple el siguiente criterio, si el coseno se aproxima a uno significa que los documentos tienen mucha similitud por lo tanto el documento debe de ser asignado a esa subclase, por el contrario si el coseno tiende a cero significa que los documentos no tienen similitud, y el documento no debe de asignarse a esa subclase.

La primera prueba de validación se realizó tomando un documento del cual se conocía su subclase (ZA4080, biblioteca digital), y realizando el cálculo de similitud entre cada una de las subclases y el vector del documento; el documento es asignado a la clasificación ZA4080, y las temáticas sugeridas son; digital libraries, information service, digital library development, y biblioteca digital, como se muestra en el siguiente resultado.

Similitud = 0.0 65
Similitud = 0.006057031660011445 66
Similitud = 1.0 67
Esta es la clasificación asignada ZA4080
Temas sugeridos: digital libraries 27
Temas sugeridos: digital library 6
Temas sugeridos: informacion 5
Temas sugeridos: information services 4
Temas sugeridos: Digital library development 3
Temas sugeridos: biblioteca digital 3
BUILD SUCCESSFUL (total time: 0 seconds)

En el resultado mostrado el sistema asigna correctamente la clasificación ZA4080, y la temática correspondiente a “biblioteca digital”.

4.6 Etapa 6.- Análisis de Resultados e Interpretación

La validación del clasificador automatizado, se realizó a través de pruebas de clasificación, para este fin se tomaron 108 registros bibliográficos del área de bibliotecología, a los cuales se asignó una clasificación por un catalogador.

Los registros bibliográficos fueron procesados y transformados en su representación del espacio vectorial y posteriormente por medio del clasificador automatizado se obtuvo su clasificación y tema de acuerdo al contenido léxico del título. Para la evaluación del clasificador se realizó la comparación entre los temas asignados por el sistema de cómputo y los asignados por el catalogador, un segmento de los resultados es mostrado en la tabla 5.

Clasificación	Tema	Acierto	Error
Z1003	Promoción de la lectura	1	
Z6954	Publicaciones periódicas eruditas	1	
Z666	Recursos electrónicos	1	
Z1001	Desarrollo de bibliotecas	1	
Z1001	Servicios de consulta (Bibliotecas)	1	
Z681	Servicios de consulta electrónicos		1
Z694	Servicios de información		1
Z710	Servicios de información	1	
Z711	Servicios de información	1	
Z711	Servicios de información	1	
Z711	Servicios de información	1	
Z739	Servicios de información	1	
Z669	Sistemas de información		1
Z699	Sistemas de información en salud	1	
Z666	Sociedad de la información	1	
Z716	Tecnología de la información	1	
Z8	Tecnología de la información		1
ZA3075	Teoría del conocimiento		1
Z1001	Universidades estatales	1	
ZA3075	Necesidades de información	1	
		82	26
		75.70%	24.3%

Tabla 5. Porcentaje de precisión en la clasificación

Los resultados obtenidos indican que el 75.70 % de los temas fueron asignados por el sistema de cómputo y el catalogador, mientras que las diferencias en la asignación de temas fue del 24.3 %.

Existen diferentes razones en el porcentaje de error entre la asignación de temas, una de ellas es el siguiente ejemplo; el título “Recuperación de información sobre cáncer de mama”, el tema asignado por el bibliotecario es

el de mama, mientras que el tema trata de recuperación de información e información de cáncer de mama, en este caso la asignación temática es equivocada.

Otro ejemplo de asignación temática incorrecta es en el título de “Bibliodrome : diseño, creación y desarrollo de una biblioteca robótica ”, a la cual se le asignó el tema Robótica, debido a que es un título de bibliotecología y de diseño de bibliotecas debería de ser asignada a la subclase z679, en arquitectura de bibliotecas

En el título “Necesidades y comportamiento informativo de los estudiantes de maestría en el área de veterinaria”, el catalogador asigna el tema Veterinaria, como el título lo refleja se trata de necesidades de información y comportamiento informativo. Finalmente se asigna el tema derechos humanos y se habla de publicaciones digitales.

Los ejemplos anteriores muestran que la capacidad del catalogador en la asignación de los temas, dependerá de sus habilidades y conocimientos en las diferentes áreas del conocimiento. Implicando que existirá un porcentaje de temáticas asignadas correctamente y otro porcentaje de temáticas asignada incorrectamente.

5 Resultados

La utilización de la minería de texto permitió el desarrollo del clasificador automatizado, a través del método de clasificación de documentos de tipo supervisado. La precisión del clasificador fue calculada realizando la comparación entre los temas asignados de manera manual y automatizada obteniendo un grado de precisión del 75.70%, mientras que las diferencias en la asignación de temas fue del 24.3 %. Realizando la comparación entre la precisión del método de máxima entropía del 69.28%, contra la precisión del método de este trabajo que es del 75.70% la precisión es mayor, pero no así con la precisión obtenida con el método de regresión lineal el cual reporta una precisión máxima del 83.7%. Por lo que se puede determinar que el grado de precisión del clasificador automatizado en la asignación de temas es bueno.

6 Conclusiones.

La minería de texto como un área de estudio del procesamiento de los documentos digitales textuales, se encarga de desarrollar y utilizar métodos para facilitar y agilizar su organización. El desarrollo de los métodos de minería de texto no es una tarea fácil debido a que se deben de integrar diversos conocimientos que van desde el comportamiento mismo de la información, conocimientos estadísticos, procesamiento del lenguaje natural y el uso las tecnologías de minería de texto y de computación. El objetivo del presente estudio se cumplió al desarrollar y aplicar el clasificador automatizado en la clasificación de material bibliográfico, obteniendo un buen grado de precisión en la asignación de temas. El presente estudio tuvo la limitante de solo considerar el desarrollo del clasificador para la clasificación Z correspondiente a la Library of Congress, pero se puede seguir desarrollando el sistema de cómputo para integrar las demás clasificaciones y con ello tener la tecnología adecuada para la organización de los documentos digitales, además de buscar mejoras para mejorar la precisión. Finalmente se concluye que la aplicación de la minería de texto facilito la creación del clasificador automatizado, permitiendo obtener tecnología útil para la clasificación de material bibliográfico con la finalidad de mejorar y agilizar el proceso de organización de información.

Bibliografía

Abbott, D. (10 de Julio de 2013). Introduction to Text Mining. Recuperado el 17 de 6 de 2014, de http://www.vscse.org/summerschool/2013/Abbott.pdf

Abdullah Muhammad, A. (2014). Medical Document Classification Based on MeSH. 2014 47th Hawaii International Conference on System Sciences (págs. 2571 - 2575). Waikoloa, HI: I EEE.

Ananiadou, S., Kell, D. B., & Tsujiii, J.-i. (October de 2006). Text mining and its potential applications in systems biology. (ELSEVIER, Ed.) Trends in Biotechnology, 24(12), 9.

Arkaitz Zubiaga, V. F. (2009). Comparativa de aproximaciones a SVM semisupervisado multiclase para clasificación de páginas Web. Recuperado el 16 de 10 de 2015, de Dialnet: http://dialnet.unirioja.es/servlet/articulo?codigo=2973575

Dey, L., Rastogi, A. C., & Kumar, S. (2006). Generating Concept Ontologies Through Text Mining. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (págs. 23 - 32). Hong Kong : IEEE.

Katerina Frantzi, S. A. (August de 2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. (S. Link, Ed.) International Journal on Digital Libraries, 3(2), 115-130.

LAN, Q. (2010). Extraction of News Content for Text Mining Based on Edit Distance. Journal of Computational Information Systems, (págs. 3761-3777).

Lee, S., Baker, J., Song, J., & Wetherbe, J. C. (2010). An Empirical Comparison of Four Text Mining Methods . Proceedings of the 43rd Hawaii International Conference on System Sciences - 2010 (págs. 1-10). Hawaii : IEEE.

Lévano, G. L. (12 de 06 de 2011). Clasificación de colecciones. Recuperado el 12 de 08 de 2013, de http://www.ugel05.edu.pe/

M.Sukanya, S. (2012). Techniques on Text Mining. 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), (págs. 269-271). Ramanathapuram .

Maggini, M., Rigutini, L., & Turchi, M. (2004). Pseudo-Supervised Clustering for Text Documents. Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on (págs. 363 - 369). IEEE .

Mahdi Shafiei, S. W. (2007). Document Representation and Dimension Reduction for Text Clustering. Workshop on Text Data Mining and Management (TDMM) in conjuction with 23rd IEEE conference (págs. 770-778). Turquia: IEEE.

Maowen, W., Caidong, Z., Weiyao, L., & QingQiang, W. (2012). Text Topic Mining Based on LDA and Co-occurrence Theory. Computer Science & Education (ICCSE), 2012 7th International Conference on (págs. 525 - 528). Melbourne, VIC : IEEE .

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. En J. K. Michael W. Berry, Text mining : applications and theory. New Jersey: Mic hael W. Berry and Jacob Kogan.

Salton, G. (1989). Automatic text processing : The transformation, analysis, and retrieval of information by computer. E.U.A: Eddison Wesley.

Salton, G., & Mcgill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Swanson, D. R. (1991). Complementary structures in disjoint science literatures. In Proceedings of the 14th Annual International ACM/SIGIR Conference, 280-289.

Swanson, D., & Smalhaiser, N. (1994). Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. Neuroscience research communications, 15, 1-9.

Verma, V. K., Ranjan, M., & Mishra, P. (2015). Text mining and information professionals: Role, issues and challenges . Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on (págs. 133 - 137). Noida : IEEE .

Wang, Z. (2010). Document Classification Algorithm Based on Kernel Logistic Regression. Industrial and Information Systems (IIS), 2010 2nd International Conference on (Volume:1) (págs. 76 - 79). Dalian : IEEE.

Wei, W., & Barnaghi, P. M. (23 de sep de 2013). University of Surrey. Recuperado el 15 de octubre de 2015, de http://epubs.surrey.ac.uk/533646/

Xiu-Li, P., Feng, Y.-Q., & Jiang, W. (2007). An improved document classificaction approach with maximum entropy and entropy feauture selection. 2007 International Conference on Machine Learning and Cybernetics (págs. 3911-3915). Hong Kong: IEEE.

Zhang, Y., & Gu, H. (2011). Text Mining with Application to Academic Libraries. En Computer Science for Environmental Engineering and EcoInformatics (págs. 200-205). Springer Berlin Heidelberg.

Notas

[i] Fuente base de datos LIBRUNAM, 26 de agosto 2015.

[ii] Rapid Automatic Keyword Extraction (RAKE) (Rose, Engel, Cramer, & Cowley, 2010), es un algoritmo utilizado para la extracción de palabras clave, compuestas por una o más palabras basado en las estadísticas de las palabras y de las coocurrencias de las mismas.

iii Fuente base de datos LIBRUNAM, 26 de agosto 2015.

iv Rapid Automatic Keyword Extraction (RAKE) (Rose, Engel, Cramer, & Cowley, 2010), es un algoritmo utilizado para la extracción de palabras clave, compuestas por una o más palabras basado en las estadísticas de las palabras y de las coocurrencias de las mismas.

Datos del autor

Marcial Contreras Barrera
Técnico Académico, Subdirección de Informática, Departamento de Producción, Dirección General de Bibliotecas, Universidad Nacional Autónoma de México – UNAM, México.
marcial@dgb.unam.mx

Recibido – Received : 2016-03-29
Aceptado - Accepted: 2016-09-14