II ENCUENTRO DE

JÓVENES INVESTIGADORES

"Consolidando espacios del quehacer científico en San Juan"

MINERÍA DE TEXTO EN LA DETERMINACIÓN AUTOMÁTICA DE CÓDIGO DEWEY (UNA PRIMER APROXIMACIÓN)

Autor:

Araya Jorge Matías

Institución: Facultad de Ciencias Exactas, Físicas y Naturales

Estructura General

1. Antecedentes y Fundamentación
2. Objetivos Generales
3. Objetivos Específicos
4. Desarrollo
5. Consideraciones Finales
6. Bibliografía

1. Antecedentes y Fundamentación

El presente trabajo se propone automatizar el proceso de determinación de la codificación Dewey asociada a todo material bibliográfico mediante la aplicación de algoritmos de minería de texto (Text Mining - TM).

TMes un área de investigación nueva y emocionante que trata de resolver el problema de sobrecarga de información textualmediante técnicas de minería de datos (Data Mining - DM), aprendizaje automático (Machine Learning - ML),procesamiento de lenguaje natural (PNL), recuperación de información (Information Recovery - IR), y gestión del conocimiento.La minería de textos implica el pre-procesamiento de colecciones dedocumentos (categorización de texto,extracción de información, extracción de términos), el almacenamiento de representaciones intermedias,estudio y aplicación de técnicaspara analizar las representaciones intermedias(tales como elanálisis de distribución, agrupamiento, análisis de tendencias, y de reglas de asociación), y visualización delos resultados.**[1]**

La numeración Dewey es un código que permite reconocer y ubicar en una biblioteca al material bibliográfico según su área de conocimiento y en éste momento, en el ámbito de la biblioteca de la Facultad de Ciencias Exactas Físicas y Naturales (FCEFN), es una tarea manual. En esta propuesta se pretende que, desde los títulos del material bibliográfico perteneciente a una Biblioteca, y mediante métricas asociadas a específicos algoritmos de TM, determinar la afinidad y pertinencia que ese material bibliográfico posee entre sí y con las distintas áreas de conocimiento.

La determinación del modelo y su validación se obtendrá desde un conjunto de títulos bibliográficos, con Dewey asignados, que servirán de conjunto de entrenamiento para encontrar el modelo que permita obtener los códigos a títulos bibliográficos que aún no poseen la asignación Dewey, sea porque es bibliografía comprada recientemente y carece del mismo o es material editado por la propia Universidad Nacional de San Juan (UNSJ).

Como caso de estudio se trabajará con material bibliográfico correspondiente a la Biblioteca Emiliano Pedro Aparicio de la FCEFN y editado en idioma español.

Actualmente la Biblioteca Emiliano Pedro Aparicio es una de las bibliotecas de la Universidad Nacional de San Juan, que cuenta con mayor nivel de informatización y aún así es posible proponer mejoras que hagan a un mejor funcionamiento.Diferentes trabajos, en el marco de anteriores proyectos de investigación referidos al área de la minería de datos, han permitido realizar distintos aportes en el ámbito de estabiblioteca, muchos de los cuales permitieron la redacción y defensa de diferentes trabajos de grado y posgrado**[2][3][4][5]**. Estos trabajos aportaron conocimiento a la Dirección bibliotecaria que ayuda en la toma de decisiones, y permite acelerar o mejorar tareas manuales inherentes a la biblioteca.La presente propuesta contenida en el marco del proyecto "Minería de Datos en la Determinación de Patrones de Uso y Perfiles de Usuarios" Cod. 21/E889, aprobado por CICITCA, es una contribución más en esa dirección.

MelvilDewey era bibliotecario en Amherst College en Massachusetts **[6]** cuando tuvo la idea de crear un sistema de clasificación que respondiera a las necesidades de la biblioteca del colegio.

La propuesta de Dewey consiste en que el número asignado no indica el emplazamiento de los libros en los estantes, sino que responde a la relación de las materias entre sí, y se basa en la numeración arábiga. Esto último tienen la ventaja de ser casi universal, a diferencia de las letras, dado que existen varios alfabetos así como otras formas de representación.

Dewey decidió que todas las materias deben de tener por lo menos tres decimales. Esto quiere decir que si se tiene una materia principal con un número básico de solo una o dos

cifras se añade un cero o dos para completar. El sistema es en principio jerárquico, a modo de ejemplo:

600 Tecnología (Ciencias aplicadas).
620 Técnica.
621 Física aplicada.
621.3 Electrotecnia.
621.38 Electrónica.
621.388 Televisión.
621.388 5 Sistema de comunicación.
621.388 57 Televisión por cable.

Asociado al contenido temático de determinado material bibliográfico surge su Código Decimal Dewey (DDC).

La minería de datos (Data Mining - DM) se refiere al proceso de extracción de conocimiento que es de interés para el usuario**[7][8]**. Se trata de un área de investigación y desarrollo interdisciplinariaque abarca diversos dominios, y lejos de estar saturada, se amplía con nuevas técnicas y orientaciones. En esta era de la exploración de datos multimedia, minería de datos ya no debe limitarse a la extracción de conocimiento a partir de grandes volúmenes de conjuntos de datos de alta dimensión en bases de datos tradicionales solamente. Así cuando los datos bajo análisis, son inherentes a la web surge el Web Mining-WM- **[9].** Cuando losdatos, para las tareas de minería, provienen de documentos de texto, forma en que se encuentra el 80% de la información existente, surge el TM. **[10]**

La presente propuesta es una continuación del trabajo de licenciatura en Sistemas de Información defendido en Diciembre de 2011.

En **[11]** la pertinencia del material bibliográfico bajo estudio, se determina desde una encuesta realizada a los docentes que conforman el grupo de profesores de las diferentes

unidades académicas de la Facultad y que hará las veces de validación del nuevo procedimiento automático.

En **[4] [13]** se hizo un análisis comparativo de diferentes métricas para la determinación de pertinencias bibliográficas determinándose que la medida similitud del coseno era la que mejor resultados entregaba, en comparación contra lo manifestado por los docentes "expertos".

Los sistemas IR toman un conjunto de documentos (colección) para procesar y luego poder responder consultas. Se puede clasificar los documentos en estructurados y no estructurados. Los primeros son aquellos en los que se pueden reconocer elementos estructurales con una semántica bien definida, mientras que los segundos corresponden a texto libre, sin formato**. [4]**

En el área de IR los documentos se representan como vectores en un espacio n-dimensional. Si un cierto valor t ocurre n veces en un documento d, entonces la t-ésima coordenada del documento d es simplemente n. Se puede seleccionar normalizar la longitud del documento a 1, usando normas L1, L2 o L∞ (1).

$$\| d_1 \| = \sum_t n(d,t) ; \|d_2\| = \sqrt{\sum_t n(d,t)^2} ; \quad (1)$$

$$\|d_\infty\| = max_t\, n(d,t)$$

Donde **n(d,t)** es el número de ocurrencias del término **t** en un documento **d**. Esta representación no rescata que algunos términos, llamados palabras claves, (ej: algoritmo) son más representativos que otros (ej.: El, la,…). Si **t**no ocurre en $\mathbf{n_t}$ documentos, de un total de **N**,$\mathbf{n_t/N}$, indica cuan "rara" es la aparición de **t** en los documentos. De aquí la importancia del término. La frecuencia inversa del documento (Inverse Document Frequency **IDF**) =$\mathbf{1+\log (n_t/N)}$ se usa para estirar las diferencias en los ejes del espacio vectorial. Igual concepto surge en términos positivos, si **t** ocurre en $\mathbf{m_t}$ documentos, de un total de **N**,**IDF**=$\mathbf{\log (N/m_t)}$ y requiere menor esfuerzo de cómputo.

Así, el valor $\mathbf{(n(d, t)/ \| d_1 \|)\times IDF\ (t)}$representala t-ésima coordenada del documento **d** en el modelo de espacio vectorial pesado, y puede tomar cualquier valor numérico a diferencia de la representación booleana donde la información vectorial mediante {0, 1} solo representa su ausencia o presencia. A pesar de ser extremadamente duro y no

capturar nada de la semántica del lenguaje, este modelo trabaja bien en definidos contextos. **[3]**

Diversas formas de medida se proponen para contrastar documentos. Una de las más conocidas es similitud del coseno, que no es otra cosa que el coseno del ángulo que forman un vector consulta **q** (un título bibliográfico) y un vector documento **dj** (planes de estudios). **[3]**

2. Objetivos Generales

Determinar,en forma automática, la codificación Dewey de material bibliográfico por catalogar, mediante estrategias de Minería de Texto (TM).

3. Objetivos Específicos

- Evaluar el estado del arte desde el acceso a cursos on-line y lectura del material ofrecido en el curso: AnIntroductiontoInformationRetrieval. Cambridge UniversityPress **[14]**.
- Estudiar y aplicar la herramienta de software libre RapidMiner (RM) y su módulo de minería de texto.
- Ampliar la información disponible de los documentos de texto propuestos como consulta, a los índices de contenido del material bibliográfico en cuestión.

4. Desarrollo

Inicialmentese estudió, analizó y evaluó el material del curso **[14]**. Del mencionado curso se cuenta con material bibliográfico correspondiente a las clases teóricas y de resolución de ejercicios. Con esto, se busca tomar conocimiento y/o reforzar el vocabulario referido al área de minería de texto y recuperación de información.

En este caso se está trabajando con una herramienta de software libre de gran divulgación y aceptación en la comunidad científica dedicada a tareas de Procesamiento de Lenguaje Natural y Recuperación de Información,RapidMiner, que al momento de esta presentación está en su versión 5.3.013, es un entorno de aplicación de algoritmos de aprendizaje de máquina y minería de datos. Allí se pueden aplicar todos los pasos involucrados en la minería de datos, desde el pre procesamiento hasta la visualización de resultados al evaluar diferentes estrategias de segmentación, de clasificación y de reglas de asociación mediante una interfaz amigable **[12]** y que se ofrece bajo una licencia AGPL versión 3.0 (software libre). Las capacidades de la herramienta citada se potencian con el agregado de un entorno de TM TextPlugin 4.2 sobre el que se pueden implementar los diferentes pasos involucrados en la minería de texto.

La Figura 1 propone el esquema modular que permite, en la herramienta RM, encontrar los grados de similitud entre material bibliográfico y áreas de conocimiento de los diferentes departamentos de la FCEFN.

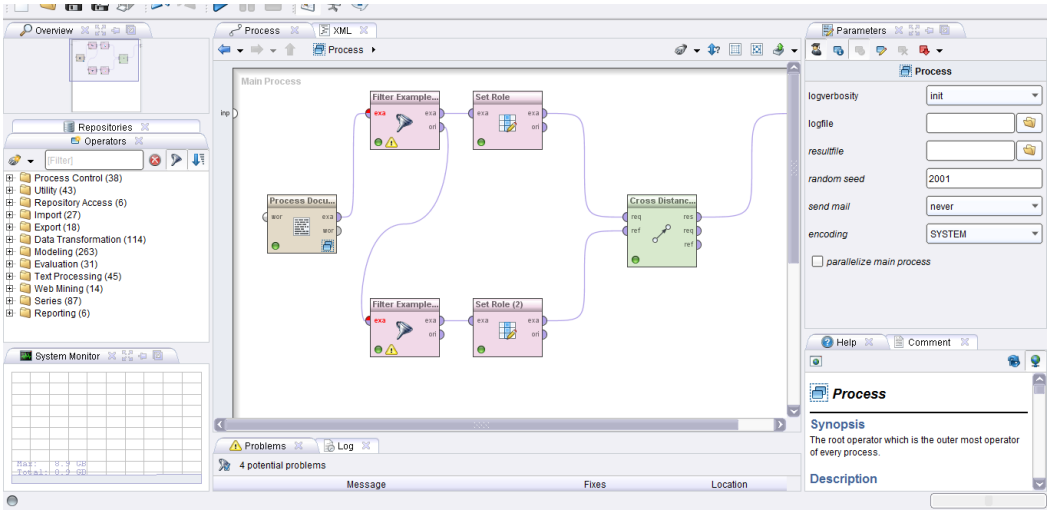

Figura1: Esquema modular de la aplicación en el entorno de software RapidMiner.

En la Figura 1 los documentos, Títulos Bibliográficos (Consulta o request **req**) y planes de estudio de carreras de los diferentes departamentos de la FCEFN (Base de Datos de Referencia **ref**) son reprocesados por un módulo de RM. En esta instancia de preprocesamiento, para cada documento y mediante la secuencia de cinco pasos, se realizan sucesivamente la separación en palabras (tokenize), la eliminación de palabras

carentes de significado (Filter Stopwords), el filtrado de palabras de cierta cantidad de caracteres (Filter Token), se reducen los términos a una forma base o raíz (stem), y por último se regeneran los documentos con cadenas de hasta una cierta cantidad de palabras (Generate n-Grams).
La conversión a mayúsculas de todos los documentos, la eliminación de acentos y la letra Ñ, entre otras, si bien puede ser plasmada con módulos de RM, y que también forma parte del preprocesamiento, en este caso, se realizó fuera de línea.
Tras la instancia de preprocesamiento los documentos se separan en títulos bibliográficos por un lado (Consulta o Request, req) y Planes de estudio (Referencia, ref) por otro. Esta separación o filtrado permite, desde el módulo (Cross Distances), la aplicación de diferentes métricas de similitud entre documentos.

Se pretende de esta manera validar la hipótesis de que el agregado al título bibliográfico de su índice de contenido permitirá encontrar mejores aproximaciones de similitud para con un área de conocimiento, respecto de cuando solamente se utilizó como comparación el título de aquel material bibliográfico. Como caso de estudio y en el marco de una primera aproximación a la determinación de pertinencias o numeración Dewey, se trabajará con material bibliográfico perteneciente a la Biblioteca Emiliano Pedro Aparicio de la FCEFN.

En este caso el procesamiento de veinte títulos bibliográficos y sus índices de contenido nos han permitido alcanzar resultados promisorios que parecen, pese al aumento del tamaño del espació de búsqueda, que se está en la dirección correcta.Desde estudios anteriores el grupo de trabajo del que formo parte y la bibliografía en general, sustenta que la métrica de similitud del coseno es una de las que mejor desempeño presenta en aplicaciones de TM e IR cuando se trata al texto bajo la representación de un modelo vectorial. De esta manera hay mayor afinidad entre documentos cuyos vectores de representación tienen menor ángulo (en este caso el coseno aumenta). Así documentos similares presentan una métrica de similitud del coseno próximo a uno, entanto que documentos diferentes presentan una métrica de la similitud del coseno próxima a cero.

La métrica de similitud del coseno varía en forma continua desde 0 para documentos diferentes, hasta 1 cuando dos documentos son iguales.

A modo de ejemplo se observa que el librocuyo título es: “Como programar en Java” y cuando sólo se considera el título como consulta brinda la métrica de similitud **0.004** para con el área de conocimiento de la biologíacuando desde la evaluación de expertos docentes, nuestra alternativa de validación, el libro tiene una mayor afinidad a los contenidos del área de conocimiento informática. Este valor cambia rotundamente cuando se considera en la consulta, además del título bibliográfico, el índice de contenidos del material bibliográfico en cuestión. En este caso aparece la métrica de similitud más alta **0.050**con el área de conocimiento informáticaremarcado en rojo en la Figura 2.

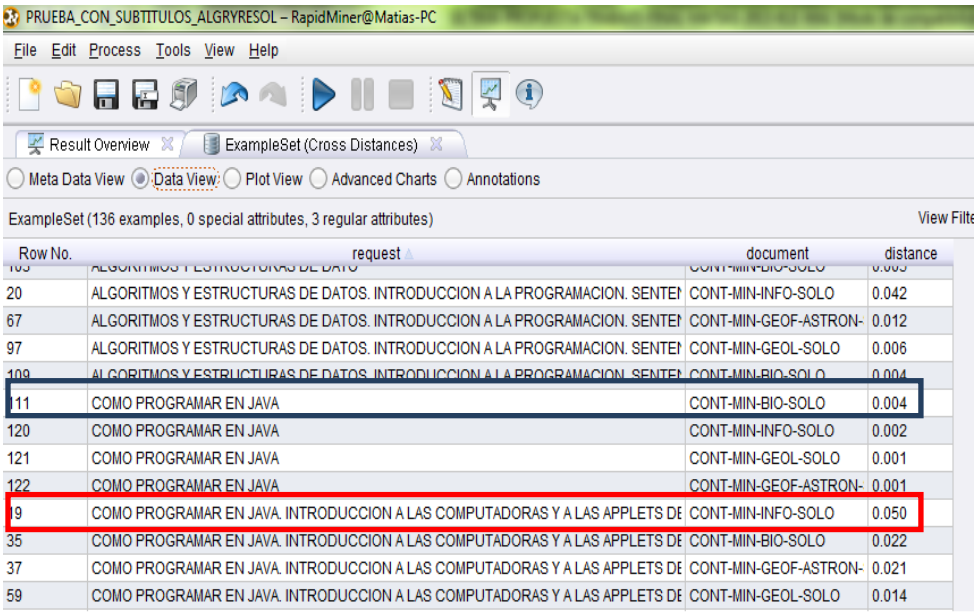

Figura 2: Métricas de similitud de material bibliográfico respecto de diferentes áreas de conocimiento.

Se pretende además extender parte de código de RM, programando la métrica Okapi que la bibliografía consultada**[17]** valora como de excelente performance. Para ello se está en plena investigación del proyecto VEGA, sobre el que se desarrolla RM, trabajando desde un entorno de programación libre como Eclipse Helios Versión Helios ServiceRelease 1.

5. Consideraciones Finales

Si bien los Títulos bibliográficos se pueden bajar del catálogo de la biblioteca de la FCEFN nos encontramos momentáneamente con el inconveniente de la digitalización y almacenamiento de los índices del material bibliográfico. Esta tarea si bien se está llevando adelante mediante la utilización de escáner y la aplicación de herramientas de software libre de reconocimiento óptico de caracteres, los mismos no tienen un adecuado desempeño lo que nos ha demorado en la aplicación dado que el OCR es pobre y en muchos casos se debe recurrir a la transcripción manual de los índices.

Consideramos que tras la carga de los índices y dado que el conjunto de material bibliográfico de entrenamiento posee su numeración Dewey la asignación del DDC para nuevo material bibliográfico debería ser exitosa.

6. Bibliografía – Referenciada y/o Consultada

[1] Feldman, R; Sanger, J. THE TEXT MINING HANDBOOK. Advanced Approaches in Analyzing Unstructured Data. Cambridge. UniversityPress 2007. 423 pág.

[2]Beguerí, Graciela, Olguín, Luis. "*Estudio sobre la Percepción del Usuario en una Biblioteca Universitaria. "Normas ISO 11620, IRAM –ISO11620"*. 2006. Publicado en: http://www.uniram.com.ar/jornadas/XXV/TC-14.pdf.

[3]Beguerí, Graciela. "*Logística como garantía de satisfacción del usuario"*.Tésis de Maestría Universidad Nacional de Cuyo. Diciembre 2007.

[4]Klenzi, R; Gutierrez, L; Villafañe, V. "*Técnicas De Recuperación De Información En La Determinación De Pertinencias Bibliográficas*" WICC-2012. Abril 2012. Universidad Nacional deMisiones. Posadas-Misiones-Argentina.

[5]Malberti, María Alejandra "*Aplicación de minería de reglas de asociación en una biblioteca universitaria"* Caso de estudio: Biblioteca Universitaria de la Facultad de Ciencias Exactas, Físicas y Naturales- Universidad Nacional de San Juan (FCEFN-UNSJ) Tesis de Maestría. Universidad Nacional de la Matanza. 2008.

[6]Benito, Miguel. *"El sistema de clasificación decimal Dewey"*. [Online] [Consulta: 3 de diciembre de 2004] http://www.adm.hb.se/personal/mb/cdu/Dewey.htm, replicado en: http://www.buenastareas.com/ensayos/Sistema-De-Clasificacion-Dewey/1211783.html visita 23 de Mayo de 2012.

[7]Larose, Daniel T. "*Data mining methods and models".*Department of Mathematical Sciences.Central Connecticut State University.John Wiley & Sons, Inc Publication. 2006.

[8]Larose, Daniel T.*"Discovering Knowledge In Data -An Introduction to Data Mining***"**. Central Connecticut State University.John Wiley & Sons, Inc Publication. 2005.

[9] Markov, Zdravko; Larose, Daniel T. "*Data mining in the webs."*Uncovering Patterns in Web Content, Structure, and Usage 2007. 218 Seiten, Hardcover - Praktikerbuch-ISBN-10: 0-471-66655-6. ISBN-13: 978-0-471-66655-4 - John Wiley & SonsInc Publication.

[10] Feldman, R; Sanger, J. THE TEXT MINING HANDBOOK. Advanced Approaches in Analyzing Unstructured Data. Cambridge. UniversityPress 2007. 423 pag.

[11] Kao, S. C., Chang, H. C., & Lin, C. H.*"Decision support for the academic library acquisition budget allocation via circulation data base mining"*. Information Processing and Management, Vol. 39, Num 1, pag 133–147. 2003.

[12] Klenzi R. *"Aplicación de minería de datos a la gestión bibliotecaria"*Un caso de estudio Biblioteca Emiliano Pedro Aparicio de la Facultad de Ciencias Exactas Físicas y Naturales de la Universidad Nacional de San Juan (FCEFN-UNSJ). Tesis de Maestría. Universidad Nacional de la Matanza. 2008.

[13] Villafañe, Viviana "*Determinación de Pertinencias Bibliográficas Mediante Técnicas de Minería de Texto***"** Tesis final de graduación. Licenciatura en Sistemas de Información.Dic. 2011.

[14]Christopher D. Manning, PrabhakarRaghavan, HinrichSchütze;
AnIntroductiontoInformationRetrieval Link:(http://www.stanford.edu/class/cs276/)

[15] Curso: Generaci_on de Lenguaje Natural y Aplicaciones Escuela de Lingüística. Computacional 201026-31 de Julio 2010, Buenos Aires, Argentina

[16]MathiasLösch, UlliWaltinger, WolframHorstmann, and Alexander Mehler: *"Building a DDC-annotated Corpus from OAIMetadata"* - Faculty of TechnologyBielefeldUniversity, BielefeldUniversity Library, DepartmentforComputerScience and MathematicsGoethe University Frankfurt am Main.

[17] Liu Bing. "*Web Data Mining. Exploring Hyperlinks,Contents, and Usage Data"*Springer-Verlag 2007.