Minería de Texto en la Determinación Automática de Código Dewey (Una Primer Aproximación)

Mag. Raúl Klenzi, Alumno: Jorge Matias Araya,
Instituto de Informática (IdeI) / Departamento Informática (DI) / Facultad de Ciencias Exactas Físicas y Naturales (FCEFN) / Universidad Nacional de San Juan (UNSJ)
Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", San Juan
{rauloscarklenzi; jorgemaraya}@gmail.com

Resumen

Este trabajo propone una primer aproximación automática del proceso de determinación de codificación Dewey asociado a todo material bibliográfico mediante técnicas de Aprendizaje de máquina y Minería de texto.
El Sistema de Clasificación Decimal Dewey (CDD) en el ámbito de la biblioteca Emiliano Pedro Aparicio de la Facultad de Ciencias Exactas, Físicas y Naturales de la Universidad Nacional de San Juan (FCEFN-UNSJ), es una tarea que se realiza en forma manual. Es propósito del presente trabajo, poner a consideración una primera instancia de automatización del proceso, mediante tareas de segmentación y medidas de similitud sintáctica, y permitir de esta manera, asignar un código adecuado a material bibliográfico recientemente adquirido por la biblioteca.
La aplicación se lleva adelante utilizando la herramienta de software libre RapidMiner (RM) 5.3.015 bajo licencia AGPL versión 3.0.

Palabras Claves: Minería de Texto, Similitudes Sintácticas, Codificación Decimal Dewey.

Contexto

La presente propuesta está contenida en el marco del proyecto trianual "Minería de Datos en la Determinación de Patrones de Uso y Perfiles de Usuarios" Cod. 21/E889, aprobado por el concejo de ciencia y técnica de la UNSJ, sujeto a evaluación externa y cuya expiración se dio el 31 de diciembre pasado. En este contexto se realizaron diferentes trabajos de aplicación de minería de datos (MD) y Minería de Texto (MT) en el ámbito de la biblioteca de la FCEFN-UNSJ, que han permitido realizar la transición desde la determinación automática de pertinencias bibliográficas a una primer propuesta de automatización del CDD. Esto ha dado sustento a la realización de trabajos finales de grado en las carreras Licenciatura en Ciencias de la Computación (LCC), Licenciatura en Sistemas de Información (LSI) y defensa de tesis de maestrias **[1][2][3][4][9]**.
Actualmente la Biblioteca Emiliano Pedro Aparicio es una de las bibliotecas de la UNSJ, que cuenta con mayor nivel de informatización y aún así es posible proponer mejoras que hagan a un mejor funcionamiento.

Introducción

Melvil Dewey era bibliotecario en Amherst College en Massachusetts **[10]** 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.

Dewey decidió que todas las materias deben de tener por lo menos tres decimales. 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.

En la UNSJ, y en particular en la biblioteca de la FCEFN, ésta signatura de clase es asignada manualmente durante el proceso de catalogación de cada ejemplar y lo aprecia un usuario al acceder al catálogo digital. La tarea de determinación de código Dewey, la realiza un experimentado integrante de la Biblioteca, que tras una lectura del material bibliográfico sometido a catalogación, correlaciona su contenido con las diferentes áreas de conocimiento establecidas en sistemas de numeración Dewey. La tarea ante mencionada, la lleva adelante el integrante de la biblioteca y para diferentes idiomas, esta aproximación automática que se presenta estará orientada esencialmente a publicaciones en idioma español.
El proceso de automatización, y en el que se usa MD y MT, toma como instancia de entrenamiento un grupo de títulos bibliográficos que posee la biblioteca Emiliano Pedro Aparicio, en idioma español y que en su totalidad han sido catalogados con sus correspondientes CDD, y mediante métricas de similitud sintáctica respecto del título por catalogar se asocia este último al CDD del material catalogado de mayor similitud sintáctica.

MD se refiere al proceso de extracción de conocimiento que es de interés para el usuario **[5][6]**. Así cuando los datos bajo análisis, son inherentes a la web surge la Minería Web (MW) **[7]**. Cuando los datos provienen de documentos de texto, forma en que se encuentra el 80% de la información existente, surge la Minería de Texto (MT).**[8]**

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**. [8] [10]**
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–log (N/m_t)** y requiere menor esfuerzo de cómputo.

Así, el valor **$(n(d, t)/\|d_1\|)\times IDF(t)$** representa la 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][8][9]**

Líneas de investigación y desarrollo

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 u otros títulos bibliográficos).

La herramienta de software utilizada RM versión 5.3.015 permite 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 **[3]** y se ofreció, hasta diciembre pasado, bajo una licencia AGPL versión 3.0 (actualmente la versión licenciada RM 6 posee un módulo libre RM starter con excelentes características también). 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.

Para poder catalogar un nuevo Título se realizan una series de pasos que van desde un pre procesamiento hasta la determinación del Dewey del libro. Esta presentación sólo intenta catalogar títulos nuevos que posean mayor afinidad sintáctica con los contenidos mínimos asociados a las carreras del departamento informática.

A modo de ejemplo se utilizan 5 libros nuevos con sus respectivos índices temáticos:

- *"Algoritmo + Estructuras de Datos = Programas"*
- *"Fisicoquímica versión si"*
- *"Introducción a la Bioestadística"*
- *"Matemtica Discreta y Lógica"*
- *"Sistemas de Información para la Administración".*

Un primer paso de pre procesamiento se realiza mediante el módulo Document Process de RM. Todo el material de texto a procesar: Contenidos Mínimos de las diferentes áreas de conocimiento de la FCEFN, Títulos ya catalogados con su Dewey y nuevo títulos. El módulo de RM se encarga de: separar en palabras (tokenize), eliminar palabras carentes de significado (Filter Stopwords), filtrar palabras de cierta cantidad de caracteres (Filter Token), reducir los términos a una forma base o raíz (stem), regenerar los documentos con cadenas de hasta una cierta cantidad de palabras (Generate n-Grams) y por último calcular los TF/IDF asociados a cada documento.

El siguiente paso consiste en determinar la pertinencia de material bibliográfico sobre las áreas de la FCEFN (Informática, Geofísica/Astronomía, Geología, Biología). Este esquema modular se observa en la Figura 1. El módulo Cross Distance mide, mediante la métrica de coseno considerada como la que mejor se adapta a la aplicación en **[9]**, la similitud sintáctica entre títulos y áreas de conocimiento y se pueden apreciar en la Figura 2.

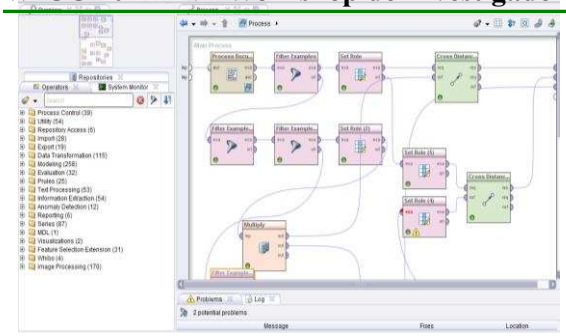


Figura 1. Parte del esquema Modular de la Aplicación en el entorno Rapid Miner 5.3.015

La figura 2 evidencia que, por ejemplo, *Algoritmo + estructuras de datos = programas* (**0,218)** y *Sistemas de información para la administración* (**0,398)** poseen mayor similitud sintáctica o afinidad con los contenidos mínimos asociados a las carreras de departamento informática. Filtrados todo los títulos bibliográficos con mayor similitud sintáctica hacia los contenidos mínimos del departamento Informática y dado que los mismos poseen el CDD, se trata de encontrar, para las titulaciones por catalogar, la mayor afinidad sintáctica hacia este conjunto de datos. Los resultados se muestran en la figura 3.

Result Overview | ExampleSet (Multiply (4)) | ExampleSet (Cross Distances (3)) | ExampleSet (Cr

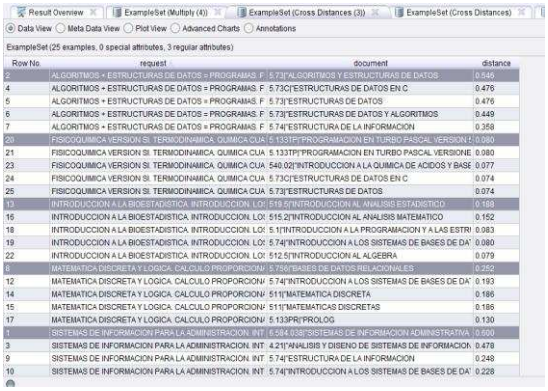
Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (20 examples, 0 special attributes, 3 regular attributes)

Row No.	request	document	distance
2	ALGORITMOS + ESTRUCTURAS DE DATOS = PROGRAMAS. FUNDAI	CONT-INFO	0.216
9	ALGORITMOS + ESTRUCTURAS DE DATOS = PROGRAMAS. FUNDAI	CONT-BIO	0.053
13	ALGORITMOS + ESTRUCTURAS DE DATOS = PROGRAMAS. FUNDAI	CONT-GEO-ASTR	0.051
17	ALGORITMOS + ESTRUCTURAS DE DATOS = PROGRAMAS. FUNDAI	CONT-GEOL	0.040
12	FISICOQUIMICA VERSION SI. TERMODINAMICA. QUIMICA CUANTICA.	CONT-GEOL	0.051
14	FISICOQUIMICA VERSION SI. TERMODINAMICA. QUIMICA CUANTICA.	CONT-BIO	0.045
16	FISICOQUIMICA VERSION SI. TERMODINAMICA. QUIMICA CUANTICA.	CONT-GEO-ASTR	0.041
20	FISICOQUIMICA VERSION SI. TERMODINAMICA. QUIMICA CUANTICA.	CONT-INFO	0.021
1	INTRODUCCION A LA BIOESTADISTICA. INTRODUCCION. LOS DATO	CONT-BIO	0.080
7	INTRODUCCION A LA BIOESTADISTICA. INTRODUCCION. LOS DATO	CONT-INFO	0.058
15	INTRODUCCION A LA BIOESTADISTICA. INTRODUCCION. LOS DATO	CONT-GEO-ASTR	0.042
18	INTRODUCCION A LA BIOESTADISTICA. INTRODUCCION. LOS DATO	CONT-GEOL	0.040
3	MATEMATICA DISCRETA Y LOGICA. CALCULO PROPOCIONAL. INDI	CONT-INFO	0.147
8	MATEMATICA DISCRETA Y LOGICA. CALCULO PROPOCIONAL. INDI	CONT-BIO	0.053
10	MATEMATICA DISCRETA Y LOGICA. CALCULO PROPOCIONAL. INDI	CONT-GEO-ASTR	0.052
19	MATEMATICA DISCRETA Y LOGICA. CALCULO PROPOCIONAL. INDI	CONT-GEOL	0.039
1	SISTEMAS DE INFORMACION PARA LA ADMINISTRACION. INTRODUC	CONT-INFO	0.398
4	SISTEMAS DE INFORMACION PARA LA ADMINISTRACION. INTRODUC	CONT-GEO-ASTR	0.089
5	SISTEMAS DE INFORMACION PARA LA ADMINISTRACION. INTRODUC	CONT-BIO	0.082
11	SISTEMAS DE INFORMACION PARA LA ADMINISTRACION. INTRODUC	CONT-GEOL	0.052

Figura 2. Resultados con la pertinencia de los libros nuevos hacia un área de conocimiento.

En la Figura 3 se aprecia que: *Algoritmo + Estructuras De Datos = Programas* es más afín a ALGORITMO Y ESTRUCTURA DE DATOS (similitud: **0.546**), **Dewey de 5.73**, por ello el sistema propondría esta asignación CDD. A su vez, *Sistemas De Información Para La Administración* tiene mayor similitud con el título: SISTEMAS DE INFORMACION ADMINISTRATIVA (similitud: **0,600**) **Dewey 6.584.038** el cual sería propuesto como CDD de la publicación a incorporar en la biblioteca de la FCEFN.



Result Overview | ExampleSet (Multiply (4)) | ExampleSet (Cross Distances (3)) | ExampleSet (Cross Distances)

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (25 examples, 0 special attributes, 3 regular attributes)

Row No.	request	document	distance
2	ALGORITMOS + ESTRUCTURAS DE DATOS = PROGRAMAS. F	5.73C|"ALGORITMOS Y ESTRUCTURAS DE DATOS	0.546
4	ALGORITMOS + ESTRUCTURAS DE DATOS = PROGRAMAS. F	5.73C|"ESTRUCTURAS DE DATOS EN C	0.476
5	ALGORITMOS + ESTRUCTURAS DE DATOS = PROGRAMAS. F	5.73J|"ESTRUCTURAS DE DATOS	0.476
6	ALGORITMOS + ESTRUCTURAS DE DATOS = PROGRAMAS. F	5.73J|"ESTRUCTURAS DE DATOS Y ALGORITMOS	0.449
7	ALGORITMOS + ESTRUCTURAS DE DATOS = PROGRAMAS. F	5.74J|"ESTRUCTURA DE LA INFORMACION	0.358
20	FISICOQUIMICA VERSION SI. TERMODINAMICA. QUIMICA CUA	5.133TP|"PROGRAMACION EN TURBO PASCAL VERSION F	0.080
21	FISICOQUIMICA VERSION SI. TERMODINAMICA. QUIMICA CUA	5.133TP|"PROGRAMACION EN TURBO PASCAL VERSIONE	0.080
23	FISICOQUIMICA VERSION SI. TERMODINAMICA. QUIMICA CUA	540.02|"INTRODUCCION A LA QUIMICA DE ACIDOS Y BASE	0.077
24	FISICOQUIMICA VERSION SI. TERMODINAMICA. QUIMICA CUA	5.73C|"ESTRUCTURAS DE DATOS EN C	0.074
25	FISICOQUIMICA VERSION SI. TERMODINAMICA. QUIMICA CUA	5.73J|"ESTRUCTURAS DE DATOS	0.074
13	INTRODUCCION A LA BIOESTADISTICA. INTRODUCCION. LOS	519.5|"INTRODUCCION AL ANALISIS ESTADISTICO	0.169
16	INTRODUCCION A LA BIOESTADISTICA. INTRODUCCION. LOS	515.2|"INTRODUCCION AL ANALISIS MATEMATICO	0.152
18	INTRODUCCION A LA BIOESTADISTICA. INTRODUCCION. LOS	5.1|"INTRODUCCION A LA PROGRAMACION Y A LAS ESTRU	0.083
19	INTRODUCCION A LA BIOESTADISTICA. INTRODUCCION. LOS	5.74J|"INTRODUCCION A LOS SISTEMAS DE BASES DE DA	0.080
22	INTRODUCCION A LA BIOESTADISTICA. INTRODUCCION. LOS	512.5|"INTRODUCCION AL ALGEBRA	0.079
8	MATEMATICA DISCRETA Y LOGICA. CALCULO PROPOCIONAL.	5.756|"BASES DE DATOS RELACIONALES	0.252
12	MATEMATICA DISCRETA Y LOGICA. CALCULO PROPOCIONAL.	5.74J|"INTRODUCCION A LOS SISTEMAS DE BASES DE DA	0.193
14	MATEMATICA DISCRETA Y LOGICA. CALCULO PROPOCIONAL.	511|"MATEMATICA DISCRETA	0.186
15	MATEMATICA DISCRETA Y LOGICA. CALCULO PROPOCIONAL.	511|"MATEMATICAS DISCRETAS	0.186
17	MATEMATICA DISCRETA Y LOGICA. CALCULO PROPOCIONAL.	5.133PR|"PROLOG	0.130
1	SISTEMAS DE INFORMACION PARA LA ADMINISTRACION. INT	6.584.038|"SISTEMAS DE INFORMACION ADMINISTRATIVA	0.600
3	SISTEMAS DE INFORMACION PARA LA ADMINISTRACION. INT	4.21|"ANALISIS Y DISEÑO DE SISTEMAS DE INFORMACION	0.478
9	SISTEMAS DE INFORMACION PARA LA ADMINISTRACION. INT	5.74J|"ESTRUCTURA DE LA INFORMACION	0.249
10	SISTEMAS DE INFORMACION PARA LA ADMINISTRACION. INT	5.74J|"INTRODUCCION A LOS SISTEMAS DE BASES DE DA	0.228

Figura 3. Resultados de la pertinencia de los nuevos libros con los demás libros ya catalogados de Área Informática.

Resultados y Objetivos

Esta primera aproximación parece ir en el camino correcto:

- Se logran proponer valores de CDD de bibliografía recientemente incorporada a la biblioteca mediante la aplicación de MT utilizando medidas de similitud sintáctica y tareas de segmentación.
- Si bien por la extensión de la publicación no se evidencia en la misma, el agregado de los índices temáticos a la sola cadena conformada por los títulos bibliográficos, y aún a costa de la ampliación del espacio de búsqueda, permitió mejorar las medidas de similitud realizadas en **[9]**.

Los pasos siguientes consisten en:

- Extender, a la totalidad de los títulos trabajados, sus correspondientes índices temáticos.
- Cotejar con las respuestas que el experto de biblioteca asigna como CDD.
- Obtener no sólo similitudes sintácticas con bibliografía ya catalogada, sino también con lo especificado en el Dewey Decimal Classification, 22st Ed.

Formación de RRHH

Esta presentación es parte de un camino iniciado por un grupo de investigadores y alumnos de la FCEFN, que permitió y permite la realización de trabajos finales de grado y posgrado. Particularmente este escrito se corresponde a un fragmento del trabajo final correspondiente al alumno de la carrera Licenciatura en Ciencias de la Computación Jorge Matías Araya.

Referencias

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

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

[3] Klenzi R. *"Aplicación de minería de datos a la gestión bibliotecaria"* Un caso de estudio Biblioteca Emiliano Pedro Aparicio de la FCEFN-UNSJ. Tesis de Maestría. Universidad Nacional de la Matanza (UNLaM). 2008

[4] Malberti, María Alejandra "*Aplicación de minería de reglas de asociación en una biblioteca universitaria"* Caso de estudio: Biblioteca Universitaria de la FCEFN-UNSJ. Tesis de Maestría. UNLaM 2008.

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

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

[7] 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 & Sons Inc Pub.

[8] Feldman, R; Sanger, J. THE TEXT MINING HANDBOOK. Advanced Approaches in Analyzing Unstructured Data. Cambridge. University Press 2007.

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

[10] Benito, Miguel. *"El sistema de clasificación decimal Dewey"*. [en líne] [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.

[11] Gandhi, Smiti. "*Knowledge management and reference service".* Senior Research Librarian**.** Harcourt Education in Orlando, Florida, United States. The Journal of Academic Librarianship. Vol. 30. Issue 5. September 2004, Pages 368-381.

[12] Lösch M.; Waltinger, U;Wolfram Horstmann, W; Mehler A. *Building a DDC-annotated Corpus from OAI Metadata.* Bielefeld University Library Faculty of Technology. Bielefeld University, Department for Computer Science and Mathematics. 1765-9329-1-PB. 2011.