TÉCNICAS DE RECUPERACIÓN DE INFORMACIÓN EN LA DETERMINACIÓN DE PERTINENCIAS BIBLIOGRÁFICAS

Mag. Raúl Klenzi, Lic. Laura Gutierrez Lic. Viviana Villafañe,
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, gutierrez.laura, villafane.viviana}@gmail.com

Resumen

En este trabajo se analizan comparativamente métricas de similitud entre documentos de texto con el objetivo de evaluar la pertinencia de títulos bibliográficos, pertenecientes a una biblioteca universitaria, respecto a las áreas de conocimiento de la unidad académica en que aquella está contenida. La instancia de comparación se realiza entre los títulos bibliográficos (consulta o request) en circulación durante un ciclo lectivo en el ámbito de la biblioteca de la FCEFN-UNSJ, versus los contenidos mínimos, perfil profesional, e incumbencias de las diferentes titulaciones que brindan los Departamentos o áreas de conocimiento de la FCEFN (referencia).
Se trata de determinar la afinidad o pertinencia de aproximadamente 700 títulos bibliográficos mediante la aplicación de la herramienta de software libre RapidMiner (RM) **[6]** utilizando sus módulos de modelado y minería de texto (TextMining –TM-). Esta herramienta, como medida de similitud sintáctica entre documentos, permite la utilización de diferentes métricas y tareas de segmentación que serán comparadas desde el punto de vista de la calidad del resultado, al contrastarlas a su vez, con pertinencias brindadas por docentes de las diferentes carreras a modo de expertos.

Palabras clave: TextMining, Recuperación de Información, Métricas de Similitud, Pertinencias Bibliográficas.

Contexto

La línea de investigación se enmarca en el proyecto bianual 2011-2012 "**MINERÍA DE DATOS EN LA DETERMINACIÓN DE PATRONES DE USO Y PERFILES DE USUARIO**" código 21/E889 que se desarrolla en el ámbito de la FCEFN-UNSJ, aprobado por el Consejo de Investigaciones Científicas Técnicas y de Creación Artística (CICITCA), financiado por la propia Universidad y ajustado a evaluación externa.
Los datos sobre los que se trabaja en el proyecto, son relativos a las áreas de salud y farmacia como así también al área educación. En esta última área se tratará con datos generados en el marco de la acreditación de las carreras del DI, del análisis de rendimiento académico de alumnos, y datos inherentes a la Biblioteca de la FCEFN. "*Las Bibliotecas Universitarias se enfrentan a un entorno básicamente digital, global y cada vez más competitivo y deberán mejorar sus servicios con el afán de sobrevivir*". **(2001, Rowena Cullen, *Library Trends, 49:4*)**
En este contexto, toda posibilidad de mejora en los servicios ofrecidos por la biblioteca es bien valorada. Desde allí, surgen los datos de circulación bibliográfica en diferentes ciclos lectivos, de encuesta a usuarios de la misma y el análisis de logs de la página web de la mencionada biblioteca. Particularmente la presente propuesta de trabajo y línea de investigación se centra en la determinación de pertinencias de títulos bibliográficos en

circulación durante un ciclo lectivo respecto de los planes de estudio de las carreras que se dictan en la FCEFN-UNSJ.

Introducción

El éxito de la revolución digital y el crecimiento de Internet aseguran que grandes volúmenes de datos multimedia de alta dimensión, están disponibles en todo lo que nos rodea. Esta información se mezcla con la participación de diferentes tipos de datos tales como texto, imagen, audio, voz, hipertexto, gráficos y componentes de vídeo entremezcladas unas con otras. Sin embargo, la mayor parte de estos datos no son de mucho interés para la mayoría de los usuarios.

La Minería de Datos (Data Mining –DM-) se refiere al proceso de extracción de conocimiento, mediante modelos, resúmenes y valores derivados de una gran colección de datos, que es de interés para el usuario. **[1]**

Cuando los datos para las tareas de minería, provienen de documentos de texto, forma en que se encuentra el 80% de la información, surge la (TM).

TM consiste en la búsqueda de regularidades o patrones que se encuentran en un texto, a partir de técnicas de aprendizaje automático; por lo tanto, se considera como una de las muchas ramas de la lingüística computacional. **[4]**

Actualmente se vive el fenómeno de "sobrecarga de información", la misma se encuentra en diferentes formatos, a veces difícil de procesar y resulta importante tratar esa información, disponible electrónicamente, para que pueda servir a diferentes personas en distintos contextos. Por ello surge la Recuperación de Información (Information Retrieval –IR-) **[7]**

En el presente trabajo se investiga y hace uso de técnicas de DM, TM, IR y uso de medidas de similitud, con el fin de encontrar pertinencias de Títulos bibliográficos, aplicando para ello el módulo Text Procesing de la herramienta de software RM el cual servirá para llevar a cabo las tareas de preprocesamiento y determinación de pertinencias.

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**, $\mathbf{IDF=\log (N/m_t)}$ y requiere menor esfuerzo de cómputo.

Así, el valor **(n(d, t)/ || d_1 ||)×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]**

Diversas formas de medida se proponen para contrastar documentos. Una de las más conocidas es similaridad 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). Otra forma es el producto punto entre los vectores **q** y **dj**. **[3]**

Luego, a partir de una consulta dada es posible devolver una lista de documentos ordenados por distancia (los más relevantes primeros). Seguidamente, se procede a realizar los cálculos algebraicos para determinar la semejanza (Por ej: mediante el producto escalar, coseno, etc.) entre el vector consulta y cada uno de los vectores que representan documentos de la colección. **[5]**

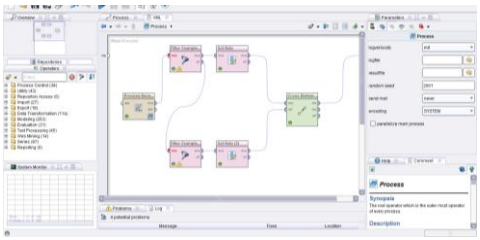

Figura1: Esquema modular de la aplicación en el entorno de software RapidMiner.

En la Figura1 se ve parte de la implementación contenida en el entorno de la herramienta de software utilizada. En ella se observa las diferentes áreas en que se divide el entorno visual, y desde donde se pueden elegir los operadores a aplicar y definir los parámetros de los distintos algoritmos implementados.

Los documentos, titulos 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 preprocesados 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.

Las métricas de similitud consideradas, poseen un rango de valores posibles que oscila en forma continua entre 0 o -1, cuando los documentos comparados son sintácticamente diferentes, y 1 cuando reflejan una similitud total.

Las métricas de similitud contrastadas se observan en la Tabla 1. En la misma, a modo de ejemplo, figuran los valores

alcanzados por cada una de ellas para un título bibliográfico específico "**ANALISIS Y DISENO DE SISTEMAS DE INFORMACION**" de los casi 700 que se compararon.
La Tabla 1 permite apreciar, con valores subrayados, una mayor afinidad "**pertinencia**" entre el título bibliográfico considerado y los planes de estudios correspondientes al DI.

Medidas De Similitud / Plan De Estudio	Coseno	Prod. Interno	Prod. Máximo	Correla-ción	Dice	Jaccard	Overlap
GEOF-ASTRON	0.0586	0.0586	0.0324	0.0511	0.0029	0.0015	0.0841
INFORM.	0.2696	0.2696	0.0820	0.2738	0.0183	0.0092	0.3168
BIOLOG.	0.0574	0.0574	0.0316	0.0500	0.0030	0.0015	0.0838
GEOLOG.	0.0330	0.0330	0.0191	0.0223	0.0015	0.0008	0.0480

Tabla1: Valores de similitud logrados por diferentes métricas.

Profundizando el análisis y tras la determinación de la pertinencia de cada título bibliográfico para con cada Departamento, mediante técnicas de segmentación se encuentra el grado de pertinencia (Total, Alta, Media, Baja y Nula) **[2]** que cada título posee con un Departamento específico desde la división del material asignado a ese departamento en cinco segmentos.
La Figura 2 muestra el esquema modular que, mediante la utilización del algoritmo de segmentación k-Means, realiza la segmentación de los títulos bibliográficos ya asignados a un área de conocimiento.

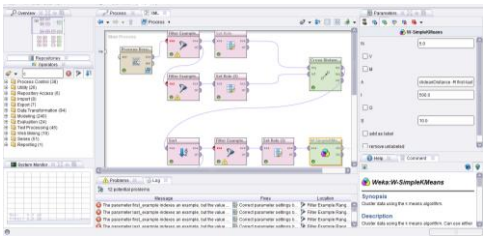

Figura 2: Esquema de módulos que permite, mediante Segmentación, determinar el grado de pertenencia.

En esta nueva instancia, y a modo de ejemplo, se observa en la Figura 3 que en el cluster (cluster_0) de mayor afinidad "pertinencia Total" para con las carreras del DI aparece el título "**ANALISIS Y DISENO DE SISTEMAS DE INFORMACION** " que coincide a su vez, y sirve a modo de validación con lo expresado por los docentes del Departamento, quienes desde su experiencia rotularon al libro de la misma manera.

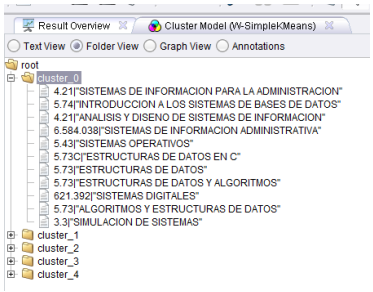

Figura 3: Asignación de títulos bibliográficos a segmentos.

Líneas de investigación y desarrollo

En el marco del proyecto que contiene la presente línea de investigación se pretende, en lo atinente a los datos de la biblioteca y entre otros, seguir los siguientes pasos:

- Dado que el material bibliográfico circulante tiene asociado su correspondiente número Dewey, el cual hace referencia a la ubicación en biblioteca de ese material según el área de conocimiento al que este asociado, y cuya obtención es manual, proponer un proceso automático de obtención para los títulos bibliográficos que carezcan de él, o para el material que genere la UNSJ.
- Correlacionar datos numéricos y comentarios textuales desde encuesta a usuarios de la biblioteca mediante TM.
- Determinar perfiles de usuarios desde el análisis de logs generados en la página web de la biblioteca.

Resultados y Objetivos

Con los objetivos iniciales de comparar métricas de similitud utilizando el módulo de Text Processing de RM, y encontrar pertinencias bibliográficas se han logrado los siguientes resultados preliminares:

- De la comparación entre métricas surge, desde esta primer aproximación, que la medida de similitud del Coseno es la mejor medida, ya que ante documentos similares su valor se aproxima a 1, sus valores fluctúan entre 0 (sin afinidad) y 1 (afinidad total), y logra una mejor separación entre los valores de pertinencia que un título bibliográfico posee respecto de todos los departamentos o áreas de conocimiento.
- Se logró una buena aproximación entre la tarea de segmentación que encuentra los grados de pertinencia y las consideraciones vertidas por docentes.

La continuidad y profundidad de esta línea de investigación nos lleva a:

- Constatar complejidades algorítmicas teóricas, respecto del consumo de procesador/es, memoria y tiempo y optar, según estas nuevas medidas, por aquella métrica que más se aproxime a lo considerado por los expertos.
- Con el objeto de mejorar las medidas de similitud entre documentos, plantear el uso de sinonimias.
- Comparar los resultados alcanzados mediante RM, con otras que también poseen, módulos de TM y WebMining (WM), por ej: KNIME (Konstanz Information Miner).

Formación de Recursos Humanos

Si bien en el marco del proyecto 21/E889 hay otros recursos humanos en formación, en particular, desde los datos inherentes a biblioteca y aplicaciones con TM se ha concluido un trabajo final de grado en Licenciatura en Sistemas de Información, se está dirigiendo otro trabajo final de grado en la carrera Licenciatura en Ciencias de la Información y una tesis de maestría en Informática de la Universidad de la Matanza. Es de estacar que el conocimiento adquirido en cada línea de investigación, es posteriormente volcado en las carreras del DI

Referencias

[1] Kantardzic, M (2003) “Data Mining: Concepts, Models, Methods, and Algorithms” ISBN:0471228524 John Wiley & Sons © (343 pages)

[2] Klenzi, R. Tesis de posgrado de maestría “Aplicación de minería de datos a la gestión bibliotecaria”. Biblioteca FCEFN-UNSJ. 2008.

[3] Liu B., “Web DataMining. Exploring Hyperlinks, Contents, and Usage Data” Springer-Verlag Berlin Heidelberg 2007

[4] Manning C, Prabhakar R. Hinrich & Hinrich Schütze, “An Introduction to Information Retrieval”, Cambridge University Press. 2009.

[5] Min, S; Yi-Fang B. Handbook of Research on Text and Web Mining Technologies -Information science reference- Editorial Advisory Board 2009.

[6] Rapid-I. http://rapid-i.com/api/rapidminer-5.1/com/rapidminer/tools. 2011.

[7] Tolosa G y Bordignon F., “Introducción a la Recuperación de Información. Conceptos, modelos y algoritmos básicos” UNLu, Arg. 2007.