Diseño de un instrumento de investigación basado en el cálculo de pesos de términos a partir del factor TF/IDF: resultados preliminares

Vuotto Andrés[1]
Bogetti Celeste
Universidad Nacional de Mar del Plata. Argentina

Resumen

Inicialmente se desarrolla una breve introducción de la actividad referente al procesamiento de los textos y su manipulación en sistemas automatizados de gestión de información, estableciendo una relación directa con la actividad del profesional documentalista y su aporte en el diseño de instrumentos de investigación vinculados a la evaluación semántica de los documentos. A partir de las técnicas de tratamiento documental señaladas se explica el trabajo implementado en una investigación cuyo principal objetivo fue determinar el nivel de presencia de un tópico específico en una colección documental; trabajando en un desarrollo a medida que aplica el factor TF/IDF como indicador de relevancia terminológica, en conjunto con un tesauro y la configuración de un árbol de pertenencias semánticas. Finalmente se detallan los resultados obtenidos y se analiza el trabajo que desde la bibliotecología es posible aportar a otras disciplinas que centren sus investigaciones en la evaluación terminológica y sus usos a partir de grandes volúmenes de información.

Introducción

La planificación e implementación del procesamiento automatizado de los textos para su indexación y modelización, y posterior incorporación en un sistema que permita la extracción de conocimiento; requiere el diseño de muchos procesos que generalmente no son de total control y/o desarrollo por parte de un experto documentalista.

Estas operaciones suelen quedar en una considerable proporción bajo la gestión de programas informáticos preparados para tal fin, cumpliendo muchas veces con aspectos generales, pero difícilmente configurables para atender particularidades de cada caso; no siendo menor el desconocimiento de las reglas y parámetros que su motor de procesamiento tiene programados para ofrecernos un resultado óptimo en la construcción de índices y cálculos específicos.

En la actualidad el procesamiento automatizado de la información dista bastante de su esencia nativa, como fue y es la de indexar y buscar documentos útiles en una colección (Baeza Yates, 1999). Hoy por hoy; por motivos fundados en el

[1] *Vuotto Andrés; Bogetti Celeste* - E-mail: avuotto@gmail.com
Departamento de Documentación. Universidad Nacional de Mar del Plata.

crecimiento de la información y documentos, como también de las necesidades del investigador y en los cambios de las propiedades documentales; incluye la modelización, clasificación y categorización de documentos, arquitectura de sistemas, interfaces de usuario, visualización y filtrado de datos, lenguajes, etc.

En este contexto, una característica fundamental y que diferencia ampliamente el tratamiento de la información de la correspondiente a los datos; es que la materia prima sobre la que se trabaja es el lenguaje natural incluido en las necesidades de información como también en los contenidos de los documentos; cuya estructura es débil y carece de exactitud semántica. Ambas cualidades obligan al desarrollo de instrumentos terminológicos y herramientas de procesamiento semántico que reduzcan la problemática a un nivel considerable y ofrezcan una respuesta con acierto aceptable.

El trabajo y procesamiento de los textos consisten en una serie de técnicas que persiguen principalmente los siguientes objetivos:

- Desarrollar economía de palabras, desechando aquellas de escaso aporte semántico a los tópicos abordados (bajo nivel de contenido).
- Definir los conceptos principales, como también el peso de cada uno de ellos en la descripción temática del documento (ponderación de palabras).

El tratamiento previo de los textos permite lograr una representación semántica de los contenidos construyendo de esta forma una visión lógica del documento.

Estos procesos durante mucho tiempo han constituido operaciones de tipo manuales o con escasa ayuda de herramientas tecnológicas, operando sobre sectores de contenido puntuales del documento con representación lógica reducida.

Con el advenimiento del texto completo y su consideración como cuerpo de significado esencial para la construcción de índices, la relación costo/beneficio de los métodos tradicionales de trabajo no eran posibles. Hecho potenciado por el crecimiento exponencial de la cantidad de documentos sobre todo en formatos digitales alojados en espacios de almacenamientos externos a los del propio sistema. Esta situación no sólo exigió una adaptación de los motores de almacenamiento, modelización y búsqueda, sino también la incorporación de algoritmos y scripts que permitan automatizar determinados procesos que ya no iban a poder ser tan controlados o dirigidos por el cerebro humano.

Las principales actividades que se realizan en la etapa de tratamiento previo de los documentos son:

1) Análisis léxico

Consiste en la transformación de un documento en palabras o autoridades para describir el contenido del texto.

El trabajo léxico plantea variadas disyuntivas sobre cuáles términos pueden ser descriptores o no del documento, para lo cual deben tomarse decisiones prestando principal atención en la frecuencia de los términos, los pesos de estos en el conjunto del texto y de la colección, y finalmente cómo se resuelve el tratamiento de los números, y caracteres especiales.

2) Eliminación de palabras vacías

Las palabras vacías constituyen aquellos términos ausentes de relevancia semántica, generalmente conformados por los artículos, preposiciones, y conjunciones; las cuales por lo general mantienen una presencia de un 80% en el total de palabras de la colección. Eventualmente también pueden incluirse otros términos.

3) Procesos de stemming (reconocimiento de stems)

Consiste en el proceso de eliminación automática de partes no esenciales de los términos (sufijos, prefijos) para reducirlos a su parte esencial (*lema*) y ayudar a la correcta indización.

Existen variadas teorías y algoritmos, las técnicas más identificativas son la *lematización*flexiva (que elimina plurales,género y terminaciones verbales) frente a *lematización*derivativa (que elimina además sufijos derivativos).

4) Detección de grupos nominales

Son conjuntos de términos que se agrupan en función de un sustantivo que cumple el rol de núcleo en el documento.

El armado de grupos nominales plantea cierto conflicto y no siempre se implementa; ya que requiere la determinación de sustantivos, adjetivos, verbos, artículos, conectores, etc. en un texto, y paso siguiente identificar el núcleo en cada caso, y los roles restantes como los *determinantes*, *complementos*, etc.

Aun así, reducir esta tarea sólo a la selección de sustantivos puede ser fundamental para la elección de los términos de indización y simplificar los índices del sistema, principalmente cuando se trabaja a texto completo.

5) Construcción de tesauros

El tesauro, del latín y griego Thesaurus cuyo significado es tesoro de palabras, refiere a un conjunto seleccionado de términos de peso semántico en un área del saber, estableciendo relaciones terminológicas entre ellos. De esta forma al buscar un término en el tesauro se puede conocer de manera inmediata qué relaciones mantiene con otros términos pertenecientes al mismo tópico.

El Tesauro se perfeccionó con el paso del tiempo hasta ser una herramienta de control terminológico muy precisa gracias al desarrollo que cobraron con las últimas décadas del S. XX los estudios de Semántica y Sintaxis.

Esta herramienta no sólo normaliza el vocabulario al tratar su sinonimia, sino que permite identificar niveles jerárquicos dentro de un mismo tema y moverse en esa matriz en el proceso de descripción semántica y de búsqueda de información.

Las relaciones que generalmente trata un tesauro son:

- Descriptor o término.
- USE: Termino autorizado para utilizar
- TR, TE y UP: términos relacionados (sinónimos), términos específicos, y términos que usan el término buscado (usado por).
- TG: términos de mayor nivel genérico.

Existen muchos tesauros automatizados en las principales áreas del conocimiento; a los efectos del pre-procesamiento de los textos el objetivo es poder compatibilizar estas herramientas con los motores de indización creados. La construcción de un tesauro no es tarea específica de la modelización de los documentos, pero sí lo es la instrumentación de estos en dicho proceso.

Las operaciones anteriormente señaladas permiten, en su correcto desarrollo, lograr una visión lógica del documento que describa semánticamente su contenido y pueda intervenir en un proceso de minería de datos e información concreta.

Objetivos de la investigación

El presente trabajo desarrolla el diseño e implementación de un instrumento de investigación de aplicación automatizada cuyo principal objetivo es determinar el nivel de presencia que un tópico mantiene, en una colección documental determinada, en relación a un campo del saber o un área temática superior. El instrumento explicado en los siguientes apartados se utilizó, en conjunto con otras técnicas principalmente de orden bibliométrico, en un proyecto de investigación dirigido por los investigadores Ana Hermosilla y Gustavo Liberatore destinado a indagar y evaluar la importancia que las asignaturas de la carrera de Licenciatura en Psicología de la Universidad Nacional de Mar del Plata le asignan a la temática Deontología y Ética Profesional en los desarrollos de sus Planes de Trabajos de los equipos Docentes (PTD).

El tratamiento que reciben los textos es de orden semántico y terminológico, haciendo uso de las técnicas anteriormente señaladas. El cual se completa con la obtención de cálculos de factores, obtenidos de forma automática por medio de

algoritmos desarrollados específicamente para este caso, que revelan con un alto nivel de acierto el peso que un concepto mantiene en una colección; como son el caso del factor TF e IDF; ambos ampliados en la explicación metodológica.

Metodología

La investigación tiene un diseño no experimental, exploratorio-descriptivo, motivo por el cual no se formularon hipótesis.

La misma se encuentra dividida en dos fases, las cuales implicaron diferentes muestras e instrumentos de recolección de datos. La que corresponde a esta presentación es la primer parte, en la cual se procedió a realizar una evaluación curricular tomando como unidad de análisis los seis programas de las asignaturas del área de investigación de la carrera de psicología. En cuanto a la segunda se está trabajando con una muestra no probabilística de docentes pertenecientes a cada una de las asignaturas.

Con respecto a la primera fase del plan de trabajo, la fuente de datos se compone de dos grupos principales de contenidos.

En primera instancia se constituyó un listado de términos o autoridades con elevada presencia semántica en el cuerpo teórico "deontología profesional". Para su conformación se trabajó con una herramienta terminológica que permitiera identificar un árbol jerárquico de términos y sus relaciones. En este caso se utilizó el "Tesauro ISOC de Psicología" en su versión en línea desarrollado por el Instituto de Estudios Documentales Sobre Ciencia y Tecnología[2]. De esta forma se construyó un árbol de pertenencias[3] cuya representación conforma una jerarquía de temas y subtemas pertenecientes al tópico mayor (deontología profesional), diferenciando en casa caso las distintas relaciones existentes entre los niveles semánticos, tomando para este caso las determinadas por el tesauro, como son TR (término relacionado), TG (término genérico), UP (usado por), TE (término específico), Familia (define la familia temática en la que se incluye cada concepto) y USE (indica el término que debe usarse como descriptor en una indexación). El objetivo del árbol de pertenencias es configurar un mapa del tópico estudiado donde quede plasmado en detalle todos los términos intervinientes con el fin de identificar los distintos niveles de representación dentro de la temática como también claros en términos cuantificables para el cálculo de frecuencias. Luego cada término fue sometido a un proceso de lematización, utilizando sólo su raíz para el cálculo de su peso dentro de los textos.

[2]Tesauro ISOC de Psicología. http://thes.cindoc.csic.es/intro_PSICO_esp.php
[3]Se puede consultar el árbol de pertenecías desarrollado en el siguiente link: http://www.psicologiayetica.com.ar/congresos/bogetti-vuotto/bogetti-vuotto-arbol-pertenecias-2013.pdf

El otro objeto de contenido que también conforma la fuente de datos para el estudio se constituye por los apartados correspondientes a los “Contenidos” y la “Bibliografía” de los planes de estudios intervinientes, cuyos textos han sido sometidos a un pre-procesado anterior a su almacenamiento en la base de datos de tipo MySQL desarrollada para tal fin. El trabajo con los textos constituyó las siguientes actividades:

- Eliminación de aquellos elementos que no representan un nivel semántico para la indexación o stripping (encabezados, notas, etc.).
- Normalización, en este caso sólo se aplicó la detección de palabras vacías (stopwordslists) por medio de la eliminación de aquellas “palabras función”, como artículos, pronombres, preposiciones, etc. Para este caso se trabajó con un algoritmo de búsqueda y eliminación de palabras desarrollado particularmente para esta investigación con el lenguaje interpretado PHP.

Con el objetivo de determinar el nivel de presencia de la temática analizada dentro de los planes de estudios se trabajó con la técnica utilizada por los sistemas de recuperación de información (de aquí en adelante SRI) para determinar qué documento o texto responde mejor a una necesidad de información planteada o consulta de usuario. En este caso el cálculo del factor TF y el factor IDF podrá determinar el peso de cada término del árbol de pertenencias en cada plan de estudios y por consiguiente el tratamiento, o nivel de ausencia, que la “deontología profesional” representa en los programas de las cátedras en cuestión.

En este caso se decidió resolver el cálculo diseñando un pequeño sistema que emule los procesos llevados a cabo por los SRI de gran escala, donde la colección de documentos está conformada por los planes de estudios procesados y almacenados en la base de datos, la temática “deontología profesional” constituye la necesidad del usuario del SRI, el árbol de pertenencias representa la jerarquía de términos que describen la necesidad del usuario y para la ejecución de la búsqueda se desarrolló un algoritmo que cumpla con los siguientes cálculos:

- Factor TF de cada elemento del árbol en cada documento: Corresponde a la capacidad de representación del término en un documento a través de la obtención de su frecuencia de aparición. Su fórmula es: $Tf(n)=\sum^{D1}_{(n)}$ Frecuencia de aparición de un término (n) en un documento (D1), es la suma de sus ocurrencias.
- Factor IDF de cada elemento del árbol en cada documento: Es el coeficiente que determina la capacidad discriminatoria del término de un documento con respecto a la colección.
- $IDF_{(n)}= log_{10} N/DF_{(n)}+1$: Donde N es el número total de documentos, DF es el número de documentos donde aparece el término n. El logaritmo se utiliza para obtener un coeficiente bajo de fácil manejo, y el +1 funciona como factor correctivo del resultado.

- Cálculo de la Ponderación TF-IDF de cada término: Corresponde al producto de ambos factores. Los resultados son una representación de la importancia del término en cada documento y por consiguiente (en función de la jerarquía armada) de la presencia del tópico en cada plan de estudio.

Considerando que se trabajó a partir de una estructura jerárquica de términos en función del peso de su significado dentro del tópico abordado, no se puede evaluar la ponderación igual para todos. En este caso un peso TF-IDF de valor 2 no representa lo mismo para un término identificado como descriptor (término autorizado para utilizar como descriptor en la indexación) que para uno que ocupa el lugar de TR (término relacionado al descriptor) en la estructura. Es por ello que se ha establecido la siguiente calificación en función del dato "nivel del término" a partir del lugar que este ocupa en el árbol de pertenencias (Algunos términos se encuentran en más de una ubicación en función de los diferentes descriptores, en esos casos se decidió el nivel del mismo a partir de su relación con la temática principal, como es *deontología profesional*):

- Descriptor o término autorizado: nivel 3
- USE: nivel 3
- TR, TE y UP: nivel 2
- TG: nivel 1

Para completar la ponderación e incluir los niveles detallados en los resultados se completará la fórmula de la siguiente manera:

Ponderación del término= TF-IDF*nivel del término

El siguiente aspecto refiere a la salida estructurada de la información acopiada en un arquitectura compatible con los cálculos a realizar y con las posibilidades de lectura de los software's intervinientes, para lo cual se han trabajado algoritmos de stemming o lematización de los textos, como también de eliminación de palabras vacías y en un menor grado técnicas para la detección de grupos nominales.

Resultados

El árbol de pertenencias se encuentra constituido por 103 conceptos, de los cuales 23 sólo mostraron peso o ponderación superior a 0 en los textos analizados. Del subgrupo de términos señalado encontramos representación de diferentes niveles en función de las diferencias jerárquicas. A continuación se incluye una versión resumida de la tabla obtenida, en la cual se puede observar por cada plan de estudios

la representación calculada de los términos del árbol de pertenencias por sus niveles[4]. Es importante señalar que la columna "ponderación por niveles" muestra una sumatoria de todas las ponderaciones obtenidas en ese nivel por cada cátedra.

La mayor representación del tópico estudiado se encuentra en el nivel 3 ya que ese grupo lo conforman los términos de mayor peso semántico, donde observamos una amplia diferencia del programa de la cátedra "Epistemología de la Psicología", como también ocurre en los diferentes niveles. Si bien es útil ver la tabla de cálculos completa, que no pudo ser incluida en este texto por cuestiones de espacio, se puede considerar que un peso o ponderación superior a 50 en el nivel 3 es una presencia aceptable de la temática en un plan de estudios donde el objeto de estudio principal no es justamente la deontología.

Tabla de resumen		
Plan de estudios[5]	**Nivel de términos**	**ponderación por niveles**
Asignatura 1	**3**	**37,49**
Asignatura 2	**3**	**15,05**
Asignatura 3	**3**	**13,24**
Asignatura 4	**3**	**11,20**
Asignatura 5	**3**	**8,86**
Asignatura 6	**3**	**6,77**
Asignatura 1	2	51,96
Asignatura 4	2	32,83
Asignatura 6	2	31,51
Asignatura 5	2	25,56
Asignatura 2	2	23,40
Asignatura 3	2	12,97

[4]La versión completa se puede ver en la siguiente dirección: http://www.psicologiayetica.com.ar/congresos/bogetti-vuotto/bogetti-vuotto-calculo-tf-idf-2013.pdf

[5]Para guardar la confidencialidad de los datos obtenidos de cada asignatura se denominó a las mismas "asignatura N". Los números indicados no refieren al orden en que se encuentran dispuestas las materias en la currícula, fueron asignados al azar.

Asignatura 1	1	11,82
Asignatura 5	1	8,86

Conclusiones

La aplicación de la herramienta diseñada en conjunto con el uso del factor TF/IDF como indicador del peso de los términos autorizados en el total de la colección, concretamente en la temática sobre la cual fue aplicada y la información expresada en la Tabla de Resumen, nos permitió concluir que en la primera fase de la investigación la transmisión de contenidos ético-deontológicos es escasa y prácticamente nula en la mayoría de las asignaturas, tomando como parámetro para esto la información textual incluida en los programas de cada cátedra. Sólo en la asignatura 1 puede plantearse la existencia de un nivel de contenido medio o aceptable, pero en las asignaturas 2 a 6 el grado de transmisión es bajo. Realizando una evaluación global, y no por asignatura, el nivel de transmisión es muy reducido.

Pero el objetivo de este escrito no es profundizar en los resultados obtenidos dentro del contexto de la formación en Psicología. Principalmente es compartir la experiencia, aún en etapa experimental, específica en el terreno de la bibliotecología y la asistencia en materia de gestión y minería de información orientada a la investigación interdisciplinaria.

La calidad en el proceso de extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior depende en gran medida de los textos almacenados en las bases de datos y de las posibilidades de desarrollar algoritmos lógicos que generen respuestas claras en conjunto con los objetivos y metas de investigación. El cotejamiento o comparación de palabras claves o descriptores, tanto de la consulta del usuario como del documento, es un sistema que no siempre arroja resultados favorables. Para evitar altas proporciones de ruido y silencio en la manipulación es necesario desarrollar un pre-procesamiento de los textos almacenados y de la consulta al sistema, trabajando con recursos terminológicos, tecnológicos e intelectuales.

Como ya se señaló, en lo que refiere a la tecnología, las opciones son varias. Programas informáticos que ayudan en esta tarea se consiguen con facilidad y en muchos casos destacados por su calidad de resultados.

Otra alternativa, que desde el trabajo bibliotecológico se puede evaluar, es la de dominar desde el inicio hasta el final cada etapa del proceso desarrollado con los

contenidos, atendiendo a particularidades de la colección y de los objetivos de la investigación.

La tarea de programación no es nativa de la actividad del documentalista, y tampoco se expresa que así deba serlo. Muy lejos de ello lo que se plantea es la posibilidad de construir pequeños programas de bajo nivel (script's) y no sistemas integrales.

La tarea del documentalista no deja de ser técnica con respecto a esta operación, y considerando la cantidad de información y la necesidad de trabajar sobre el texto completo, se vuelve difícil de sostener en el tiempo. Es por ello que consideramos que el abordaje técnico del bibliotecario debe ubicarse principalmente en un lugar de dominio de los scripts y de los flujos de la información, y no exclusivamente de usuario avanzado de herramientas existentes que en muchos casos pueden ser de gran utilidad pero frente a eventualidades nos obliga a depender de la solución aportada por un tercero o de abandonar el esquema de trabajo para obtener uno nuevo con otras herramientas.

El trabajo orientado al descubrimiento de conocimiento a partir de grandes volúmenes de información, denominado KDD o KnowledgeDiscovery in Databases, exige abordajes de índole puramente técnicos como también niveles de creatividad en el diseño de instrumentos; tareas que en la investigación científica se vuelven muchas veces centrales para dar paso a la elaboración de cuerpos teóricos, conclusiones y comprobación de hipótesis. En este escenario el profesional de la información puede intervenir participando activamente no sólo en las tareas frecuentes vinculadas a la recolección de datos e información y/o preparación de los textos. Principalmente por su conocimiento en materia de herramientas terminológicas, automatización de grandes volúmenes de textos y estadística aplicada al impacto de la información es que su contribución sin lugar a dudas puede centrarse en el desarrollo de herramientas que permitan modelar la evolución de variables con fines descriptivos y predictivos, con gran valor para toda investigación que requiera una implementación de minería de textos para el logro de sus objetivos.

Bibliografía

Andres T. Hohendahl; José F. Zelasco (2006). Algoritmos eficientes para detección temprana de errores y clasificación idiomática para uso en procesamiento de lenguaje natural y texto. Buenos Aires: RedUNCI. ISBN: 950-9474-35-5

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.

Cueva Martín, Alejandro de la (2000). Acceso y utilización de tesauros en Internet. Documentación Digital. Curso de posgrado. UniversitatPompeuFabra. http://161.116.140.71/pub/fburg/docs/cueva.pdf

De Andrea, N. (2011, diciembre). *Percepción de alumnos/as de psicología de universidades públicas con respecto a situaciones ético-dilemática en sus prácticas*. Ponencia presentada en V Congreso Marplatense de Psicología. Mar del Plata, Argentina.

De Vicente Rodríguez, P. (2006). Formación práctica del estudiante universitario y deontología profesional. *Revista de Educación*, 339, 711-744.

García Figuerola, L. C. (2000). La investigación sobre recuperación de la información en español. En C. Gonzalo García y V. García Yedra (Eds.) "Documentación, Terminología y Traducción" (pp. 73-82). Madrid, Síntesis.

Garrido Medina, J. (1991). Indexación automatizada de publicaciones lingüísticas: el proyecto CAPLE. Procesamiento del lenguaje natural. N. 9 (enero 1991); pp. 107-121.

Liberatore, G., &Lizondo, L. (2009). Representación semántica de un catálogo de tesis por medio de una interfaz de visualización gráfica basada en la metodología TopicMaps. Biblios, (34), 1-13.

Mehdi, A. ;Friedhelm, B.; Antony, D.; y otros. (2012). *Manual de PHP.* http://www.php.net/manual/es/

Moreiro González, J. A (2004). El contenido de los documentos textuales. Su análisis y representación mediante el lenguaje natural. España: Trea.

TheFuturesGroup (1999). Arbol de pertinencias y análisis morfológico.

ValcarcelAsensios, V. (2004). Data mining y el descubrimiento del conocimiento. *Ind. Data* . Jul/Dic. vol.7, no.2, pp-83-86 ISSN 1810-9993.

Vallez, M y Pedraza-Jimenez, R (2007). El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. Hipertext.net, 5, 2007.Disponible en:http://www.upf.edu/hipertextnet/numero-5/