Biblios
ISSN 1562-4730 (online)

Aplicación del factor TF-IDF en el análisis semántico de una colección documental

Andrés Vuotto
Celeste Bogetti
Gladys Fernández

Universidad Nacional de Mar del Plata – MDP, Argentina.

ARTICLES

Resumen

Objetivo. Describe la aplicación de una herramienta para el análisis semántico de una colección documental, basada en el uso de la frecuencia de término – frecuencia inversa de documento (TF-IDF).
Metodología. Se desarrolla un sistema, basado en lenguaje PHP y bases de datos MySQL, para la gestión de un tesauro, del cálculo TF-IDF (como indicador de peso semántico) y para el desarrollo de un árbol de relevancia (conformado por aquellos conceptos más relevantes del tema analizado). Se evaluó la herramienta en el análisis semántico de una colección documental de Psicología.
Resultados. El sistema logró identificar el nivel de presencia del tema: deontología profesional, en una colección los documentos del programa de Psicología.
Conclusiones. La experiencia descrita confirma la viabilidad de la herramienta para el análisis semántico de una colección documental. Destaca la pertinencia y las capacidades de los profesionales de la información para el desarrollo de herramientas para el tratamiento de información. Los autores sugieren un especial abordaje técnico a partir del uso de scripts y de flujos de la información.

Palabras clave

Análisis semántico; TF-IDF; Recuperación de información; Minería de datos; Extracción de información en bases de datos

Application of TF-IDF factor in the semantic analysis of a documentary collection

Abstract

Objective. This paper describes the application of a tool for the semantic analysis of a document collection based on the use of term frequency–inverse document frequency (TF – IDF).
Methodology. A system based on PHP and MySQL database for the management of a thesaurus, the calculation of TF – IDF (as an indicator of semantic weight) and for development a relevance tree (consisting of those concepts is developed most relevant issue analyzed). The tool was tested to the semantic analysis of a documentary collection of Psychology.
Results. The system was able to identify the level of track presence: professional ethics, in a collection of documents Psychology program.
Conclusions. The experience described confirms the viability of the tool for the semantic analysis of a documentary collection. It underlines the relevance and capacities of information professionals to develop this kind of tools for processing information. The authors suggests a special technical approach for use of scripts and information flows.

Keywords

Semantic analysis; TF - IDF; Information retrieval; Data mining; Knowledge discovery

1 Introducción

La recuperación de la información (de aquí en adelante RI) puede definirse de forma resumida como el proceso de indexar y buscar documentos útiles en una colección respondiendo a necesidades del usuario (Baeza Yates, 1999). Por motivos fundados en el crecimiento de la información y documentos, como también en las necesidades del usuario y en los cambios de las propiedades documentales; hoy más que nunca no se debe obviar en su concepción a la modelización, clasificación y categorización de documentos, la construcción de arquitectura de sistemas e interfaces de usuario orientados a la visualización y filtrado de datos, y principalmente la participación de los lenguajes documentales derivados de XML.

Con el advenimiento de la web, sobre el inicio de los años 90, cambiaron radicalmente los actores que intervienen en el proceso de producción e investigación documental, rediseñando rápidamente los conflictos y desafíos concernientes a la actividad de la RI, donde el centro de la escena y el principal acervo documental convergen en un espacio virtual cuyo modelo de datos y estructura de la información son de difícil definición.

Una característica fundamental, y que diferencia ampliamente la RI de la recuperación de datos, es que la materia prima sobre la que se trabaja es el lenguaje incluido en las necesidades de información como también en los contenidos de los documentos. Esta herramienta que utilizan las personas para expresarse, denominada lenguaje natural, posee propiedades que merman la efectividad de los sistemas de recuperación de información textual. Estas propiedades son la variación y la ambigüedad lingüística. Cuando hablamos de la variación lingüística nos referimos a la posibilidad de utilizar diferentes palabras o expresiones para comunicar una misma idea. En cambio, la ambigüedad lingüística se produce cuando una palabra o frase permite más de una interpretación (Vallez & Pedraza-Jimenez, 2007).

Ambas cualidades obligan al desarrollo de instrumentos terminológicos y de procesamiento semántico que reduzcan la problemática a un nivel considerable y ofrezca una respuesta en la recuperación con acierto aceptable.

Las actividades de indexación se realizan principalmente sobre el texto completo de los documentos, pero también se aplican técnicas similares para procesar la consulta en el sistema de recuperación de información (SRI), ingresada por el usuario. Como se observa en la Figura 1, el proceso desde que el usuario plantea en lenguaje natural su consulta al sistema hasta que este último ofrece una respuesta representada en documentos incluye varios pasos intermedios:

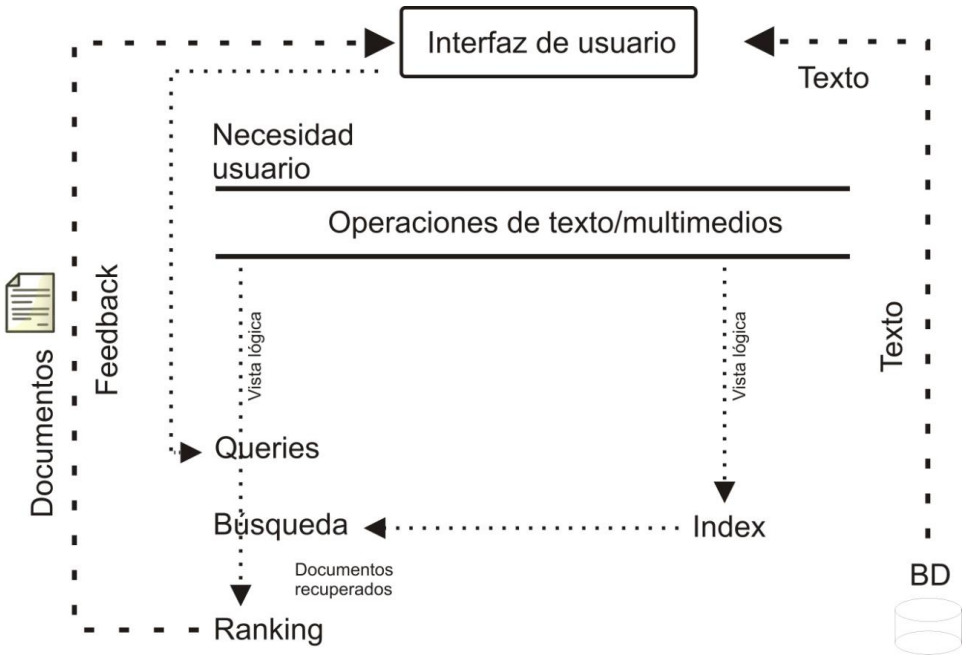

Figura 1 - Proceso de indexación
Fuente: Los autores (2015).

No 60 (2015) • http://biblios.pitt.edu/ • DOI 10.5195/biblios.2015.227

El trabajo y procesamiento de los textos, previamente a ser ingresados a las bases de datos de un SRI, 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.

Actualmente el texto completo es considerado un cuerpo semántico esencial para la construcción de índices, modificando la relación costo/beneficio de los métodos tradicionales de trabajo. 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 SRI. Esta situación no sólo exigió una adaptación de los motores de almacenamiento 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. Ya que en muchos casos pueden ser de gran utilidad para describir un contenido y en otros pueden presentar mucha vaguedad al respecto.

2. **Eliminación de palabras vacías**

   Las palabras vacías son inútiles en el momento de la recuperación y no describen semánticamente un texto; usualmente conformadas por el grupo de 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. Su eliminación reduce considerablemente el peso de los textos a almacenar y agiliza el trabajo de los motores de almacenamiento y de búsqueda.

   La eliminación de palabras vacías, al igual que el trabajo léxico, debe realizarse con cuidado para evitar minimizar la posibilidad de acierto discriminando términos significativos.

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 y recuperación.

   La reducción de una palabra a su lema se realiza no sólo con los términos de indización resultante del trabajo léxico, sino también se realiza sobre la consulta del usuario en el momento de la búsqueda.

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).

La aplicación de técnicas de stemming puede resultar contradictoria a los beneficios de la recuperación cuando no se realiza correctamente, ya que debe ir acompañada con un trabajo terminológico que ofrezca al sistema el listado de lemas y de implementación de algoritmos efectivos, como es el caso del aportado por Porter para eliminación de sufijos que con simplicidad arroja resultados comparables a aquellos algoritmos de mayor complejidad.

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 en un SRI, 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.

Los SRI más complejos incluyen la automatización de esta operación, constituyendo un resultado altamente favorable para la RI, pero es necesario evaluar el costo en el procesamiento de los textos y en los recursos afectados.

Aún 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 refiere a un conjunto seleccionado de términos de peso semántico en un área específica del saber, estableciendo relaciones terminológicas entre ellos. De esta forma al buscar un término en el tesauro se puede conocer de forma inmediata qué relaciones mantiene este con otros términos pertenecientes al mismo tópico.

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; a los efectos del pre-procesamiento de los textos el objetivo es poder compatibilizar estas herramientas con los motores de indización creados. La creación 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 búsqueda por significado y/o necesidades temáticas.

2 Objetivos del estudio realizado

El presente estudio forma parte de dos proyectos de beca de investigación del Consejo Interuniversitario Nacional (Beca CIN 2012-2013 y 2013-2014) en los que se indagaban distintos aspectos correspondientes al área de Investigación en psicología de la carrera Licenciatura en Psicología de la Universidad Nacional de Mar del Plata. Esta investigación complementaria a las becas señaladas, persigue el objetivo de poder obtener de forma automatizada una representación estadística basada en técnicas de procesamiento del lenguaje natural que permitan identificar el nivel de presencia de un tópico determinado en los programas de estudios de cada una de las asignaturas analizadas. De esta forma sería posible evaluar este aspecto cada cátedra, tomando como fuente los textos de cada programa, en función de aquellos contenidos que se consideran centrales para la formación del profesional y deben ser incluidos en niveles aceptables en la formación del licenciado en psicología. Como objetivo secundario se persigue la portabilidad del instrumento de investigación diseñado, permitiendo replicar de forma simple su aplicación en distintas cátedras y pudiendo variar con pocos pasos intermedios el tópico a analizar.

En este trabajo se ejemplifica y describe cómo se logra el objetivo indicado tomando como tema a indagar en cada cátedra la “formación ético-deontológica”.

3 Tecnologías empleadas para el modelado de textos y esquema de trabajo con cada una de ellas.

El tratamiento de los textos consiste en un proceso que requiere de la instrumentación de recursos intelectuales y tecnológicos. Las tareas a realizar deben ser “comprendidas” y ejecutadas por motores virtuales que permitan operar con grandes colecciones y en tiempos reducidos a milésimas de segundos en cada etapa, asegurando resultados altamente ponderables ya sea para el almacenamiento inicial como para cada búsqueda y recuperación.

La construcción de sistemas automatizados incluye de forma excluyente la comprensión completa por parte de todo el equipo de cada operación a resolver con los textos, y como tareas no menores la selección de las tecnologías adecuadas en función de estrictos criterios de evaluación de variables de rendimiento y el desarrollo de algoritmos que resuelvan cada proceso de forma integrada.

En lo que refiere a las tecnologías, las opciones son variadas. Es posible encontrar sistemas ya desarrollados para estos casos; los cuales pueden completar todo el proceso o partes de este; obligando a utilizar, sincronizar y compatibilizar programas informáticos para distintas funciones.

La propuesta que se intenta explicar en este trabajo no adopta la modalidad de operar con sistemas ya desarrollados, principalmente si nos referimos a casos no libres y programados con lenguajes poco conocidos. Si bien es cierto que esa opción puede significar el camino más rápido hacia los objetivos, la idea de abordar el desarrollo propio se funda en un pensamiento evolutivo y de control total de los textos, materia prima que requiere un riguroso cuidado ya que cuando las dimensiones son intangibles hechos como pérdida de información, transmisión de ruido y/o silencio en el proceso de búsqueda, y demás amenazas pueden suceder invisibles a los gestores del sistema, y aún peor, a los usuarios.

A continuación se detallan, en términos generales, las características que se deben considerar a la hora de elegir las tecnologías intervinientes:

- Trabajar con tecnologías de tipo software libre, primordial para acceder a un mundo de desarrollo con mayores permisos y a una serie innumerable de recursos ya probados y estudiados por una comunidad internacional particularmente activa.
- El aprendizaje y la escritura de los lenguajes intervinientes debe ser simple, la complejidad en este aspecto no asegura un mayor rendimiento y entorpece considerablemente el trabajo cotidiano.

- La velocidad de procesamiento debe ser veloz con un mínimo de recursos informáticos; es necesario considerar la cantidad de información en simultáneo con la que se estará trabajando, traducida en bits ocupan espacios de memoria en el computador y en el microprocesador que mal administrados pueden interrumpir el proceso.

- Las posibilidades de almacenamiento de información debe ser considerablemente amplia, ya que será necesario almacenar textos completos de los documentos y desarrollar sobre ellos tareas de altas, bajas y modificaciones constantemente.

- Todas deben ser herramientas multiplataforma y de fácil migración, ya que incurrir en dependencias de estructuras tecnológicas puede ocasionar en el futuro obstáculos importantes hasta la posibilidad de tener que desechar todo el desarrollo recorrido.

- Los lenguajes de programación deben trabajar, de forma “embebida”, con lenguajes de marcado; agilizando y ampliando las posibilidades de visualización en pantalla y en el uso de los browsers como ventanas a los contenidos y motores de búsqueda.

Este trabajo propone, como conjunto de tecnologías que cumplen de forma comprobada cada lineamiento anteriormente indicado. El uso de Apache como servidor web robusto, PHP (Hypertext Pre-processor) en la función de lenguaje de programación con amplias propiedades para el procesamiento de texto y grandes cantidades de información en tiempos reducidos, MySQL para la gestión de bases de datos y XML (lenguaje de marcado que define el HTML) y ofrece la posibilidad de generar nuevos lenguajes en función de necesidades específicas. Para la escritura de código se utilizó el software NotePad++, los lenguajes informáticos intervinientes se escriben en editores de textos simples o planos, en este caso el NotePad++ es un potente editor de texto con funciones para el trabajo con los lenguajes anteriormente señalados.

La metodología de trabajo no requiere de una estructura tecnológica onerosa, por el contrario, se puede resolver en una única computadora bajo el esquema denominado “todo en uno”, donde el servidor y el cliente (en este caso el cliente sería el navegador web o browser que permite visualizar la ejecución de los procesos) se encuentran en la misma computadora. Para las etapas de prototipo y desarrollo, este esquema suele resultar el más cómodo, una vez finalizado el software ya es posible incluirlo en un servidor de acceso público en el caso de que esa sea la intención. En la Figura 2 pueden observarse las distintas etapas de este proceso.

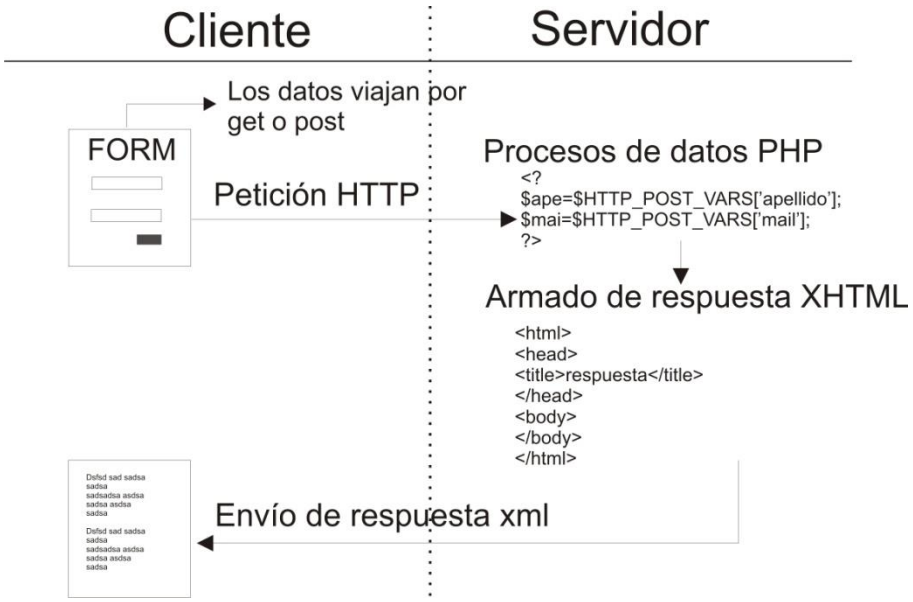

Figura 2 - Prototipo y desarrollo del sistema

Fuente: Los autores (2015).

La Figura 3 muestra la evolución de la modelización de los textos que se lleva a cabo con cada documento:

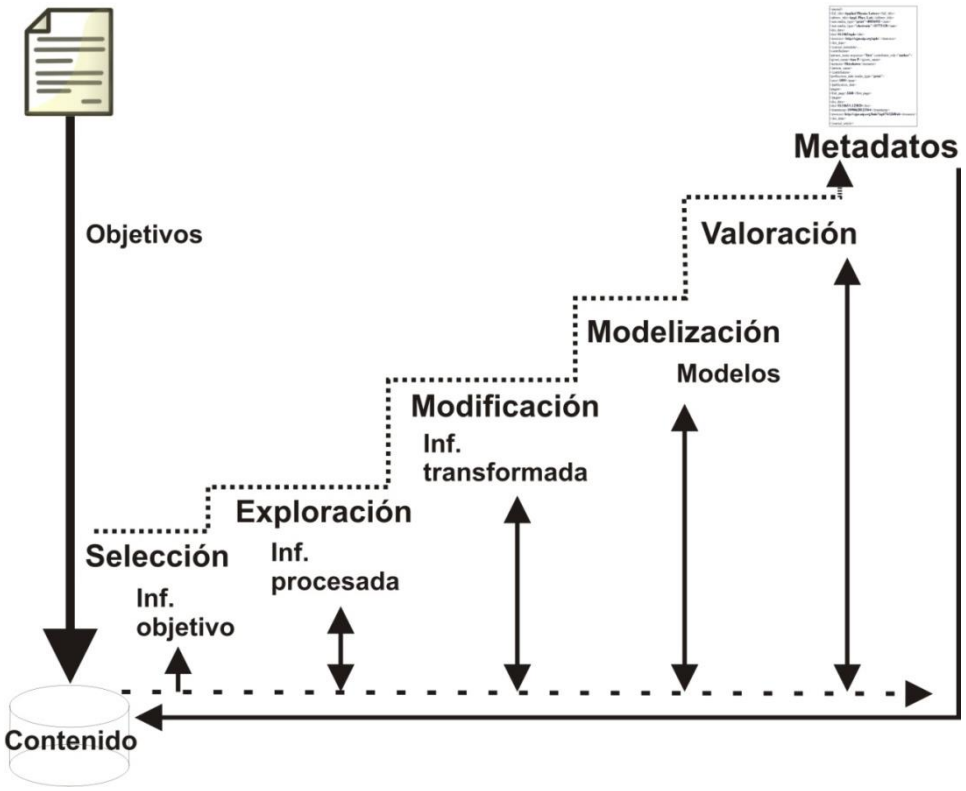

Figura 3 - Proceso de modelado de la información

Fuente: Los autores (2015).

4 Metodología desarrollada

La investigación tiene un diseño no experimental, exploratorio-descriptivo, motivo por el cual no se formularon hipótesis. Constituye en la evaluación de la presencia que un tema específico mantiene en una colección de documentos determinada, para lo cual se trabajó aplicando técnicas de indizado, búsqueda y recuperación; desarrollando un sistema de recuperación de información que cumpla con los lineamientos anteriormente planteados en este escrito y permita transportar la experiencia a investigaciones similares.

La misma se encuentra dividida en dos fases, las cuales implicaron diferentes muestras e instrumentos de recolección de datos. La que corresponde a este trabajo es la primer parte, en la cual se procedió a realizar una evaluación curricular tomando como unidad de análisis los programas de las asignaturas del área de investigación de la carrera de psicología. Conformando la colección de nuestro SRI.

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 utilizó el “Tesauro ISOC de Psicología” en su versión en línea desarrollado por el Instituto de Estudios Documentales Sobre Ciencia y Tecnología(1). De esta forma se construyó un árbol de pertenencias(2), el cual constituye una técnica analítica que subdivide un amplio tema en subtemas cada vez menores. El resultado es una representación pictórica con una estructura jerárquica que indica cómo un tema determinado puede subdividirse en niveles de detalle cada vez mayores (The Futures Group, 1999). En consecuencia se obtuvo una jerarquía de temas y subtemas pertenecientes al tópico mayor (deontología profesional), diferenciando en cada 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 conceptos 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.

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 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 documentos se trabajó en el cálculo del factor TF y el factor IDF. El método TF-IDF general listas de palabras claves con una calificación o peso que indica qué tan relevante es la palabra con respecto al documento seleccionado y al corpus en general (Cabrera, 2011). De esta forma fue posible establecer una ponderación de cada término del árbol de pertenencias en cada documento de la colección; y por consiguiente el tratamiento, o nivel de ausencia, que la "deontología profesional" representa en los programas de las cátedras en cuestión.

La determinación de una evaluación de términos a partir de cálculos automatizados se presenta en este trabajo como un aspecto central para el logro de los objetivos de la investigación, la adopción del factor TF-IDF se sustenta principalmente en los óptimos resultados observados en investigaciones anteriores en las que se trabajó con este indicador bajo esquemas de procesamiento de los textos similares. Cobo Ortega, Rocha Blanco y Alonso Martínez (2009), en su trabajo "Descubrimiento de conocimiento en repositorios documentales mediante técnicas de Minería de Texto y Swarm Intelligence" utilizan el cálculo TF-IDF para establecer un peso de los términos de la colección y a partir de los resultados logran establecer una representación vectorial de las palabras incluidas en los textos. Trabajando con un listado de términos previamente establecidos de cada área disciplinar, a modo de herramienta lingüística, se obtuvo un peso respecto de la presencia semántica de cada concepto y se redujo significativamente el número de representaciones en función de rasgos insignificantes por estar presentes en un excesivo número de documentos de la colección. Otra investigación que presentó una aplicación de la metodología empleada, pero en este caso para el establecimiento de un ranking de resultados en procesos de recuperación de documentos, es la investigación de Pérez-Iglesias, Fresno y Pérez-Agüera (2008) "Funciones de Ranking basadas en Lógica Borrosa para IR estructurada" donde se trabaja en ponderaciones de términos tanto en la consulta realizada al sistema como también en los contenidos de los documentos que conforman la colección. Trabajos más recientes que abordan el cálculo automático del factor TF-IDF para identificar relevancia semántica es el presentado por Vargas Rosales (2015), donde a partir de un proceso inverso busca generar una taxonomía a partir de un dominio determinado y una comunidad de documentos específicos; también Ropero Montejo (2014) considera pertinente el cálculo de la frecuencia de un documento al desarrollar un método de evaluación automática de organización de textos, haciendo uso del factor inverso para aplicar técnicas de aprendizaje informáticas al procesamiento de lenguaje natural (PNL).

Para la implementación práctica de este caso, en función de las herramientas con las que se contaba para el diseño de un instrumento adecuado, 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)=∑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)= log10 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.

5 Resultados obtenidos

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 (3). Es importante señalar que la columna "ponderación por niveles" muestra una sumatoria de todas las ponderaciones obtenidas en ese nivel por cada documento.

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 documento "Asignatura 1", 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 30 en el nivel 3 es una presencia aceptable de la temática en una colección de documentos donde el objeto de estudio principal no es justamente la deontología (tópico seleccionado como necesidad de información en esta investigación).

Tabla 1 - Cálculo de ponderación

Plan de estudios(4)	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
Asignatura 1	1	11,82
Asignatura 5	1	8,86

Fuente: Los autores (2015).

6 Análisis y discusión de la metodología empleada y los resultados logrados

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 identificar el nivel de presencia en la formación en

psicología que mantiene el tópico analizado; tomando como colección los documentos de los programas de cada asignatura, y como necesidad de información la deontología profesional.

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 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 Knowledge Discovery 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 y sus usos 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 y cálculos de pesos semánticos.

Referencias

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

Cabrera Diego, L. A. (2011). *TF-IDF para la obtención automática de términos y su validación mediante Wikipedia.* Tesis para la obtención del título de Ingeniero en Computación. Facultad de Ingeniería, Universidad Nacional Autónoma de México, México D. F., México.

Cobo, A.; Rocha, R. & Alonso, M. (2009). Descubrimiento de conocimiento en repositorios documentales mediante técnicas de minería de texto y swarm intelligence. *Revista Electrónica de Comunicaciones y Trabajo de Asepuma, 10, 105-124*.

Liberatore, G; Vuotto, A; Bogetti, C. & Hermosilla, A. (2011). *Análisis de las relaciones existentes entre las asignaturas de grado de ética y deontología de las carreras de psicología de Argentina mediante la técnica del apareo bibliográfico (bibliographic coupling)*. Ponencia presentada en el V Congreso Marplatense de Psicología. La Psicología en el porvenir de la cultura. El semejante: entre el enemigo y el desamparado. Mar del Plata, Argentina.

Pérez-Iglesias, J., Fresno, V., & Pérez-Agüera, J. R. (2008). Funciones de Ranking basadas en Lógica Borrosa para IR estructurada. *Procesamiento del lenguaje Natural*, 41, 173-180.

Ropero Montejo, F. T. (2014) *Método para la evaluación automática de la organización de textos argumentativos*. Tesis para la obtención del título Magister en Ingeniería de Sistemas y Computación. Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Bogotá, Colombia.

The Futures Group (1999). Arbol de pertinencias y análisis morfológico.

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

Vargas Rosales, A. A. (2015). *Desarrollo de una herramienta que permita la extracción de una taxonomía de un conjunto de documentos de un dominio específico usando CFinder para la extracción de conceptos clave*. Tesis para la obtención del título de Ingeniero Informático. Facultad de Ciencias e Ingeniería, Pontificia Universidad Católica del Perú, Lima, Perú.

Vuotto, A & Bogetti, C (2014, octubre). *Diseño de un instrumento de investigación basado en el cálculo de pesos de términos a partir del factor TF-IDF: resultados preliminares*. Ponencia presentada en el X Encuentro de Directores y IX de Docentes de Bibliotecología y Ciencia de la Información del MERCOSUR, Buenos Aires, Argentina.

Notas

1- Tesauro ISOC de Psicología. http://thes.cindoc.csic.es/intro_PSICO_esp.php

2- Se puede consultar el árbol de pertenecías desarrollado en el siguiente link: http://www.infeidon.com.ar/investigacion/art-biblios/arbol-pertenecias.pdf

3- La versión completa se puede ver en la siguiente dirección: www.infeidon.com.ar/investigacion/art-biblios/calculo-tf-idf.pdf

4- 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.

Datos de los autores

Andrés Vuotto
Jefe de Trabajos Prácticos con dedicación exclusiva a la docencia e investigación, perteneciente al área de Procesamiento de la Información, Departamento de Documentación, Universidad Nacional de Mar del Plata.
avuotto@gmail.com

Celeste Bogetti
Becaria en investigación por la Facultad de Psicología de la Universidad Nacional de Mar del Plata. Desde sus inicios como miembro del grupo de investigación PSICOLOGIA Y ETICA. CIENCIA Y PROFESION - OCA 1077/07 ha trabajado en las temáticas Formación y Ética profesional, como también Deontología profesional; con destacadas participaciones en congresos nacionales e internacionales y colaboraciones en publicaciones científicas del área.
celes.bogetti@gmail.com

Gladys Fernández
Docente e investigadora del Departamento de Documentación de la Universidad Nacional de Mar del Plata, miembro del equipo de desarrollo de área de educación a distancia. Graduado y docente/investigador del Departamento de Documentación de la Universidad Nacional de Mar del Plata, trabajando en el área Procesamiento de la Información; siempre se dedicó al estudio de la aplicación de las nuevas tecnologías de la información para la representación, análisis y gestión documental; como también para el desarrollo y mejora de sistemas de educación a distancia bajo modalidad virtual.
gvfernan07@gmail.com

Recibido - Received: 2015-07-21

Aceptado - Accepted: 2015-10-19