Obtención de descripciones significativas para una memoria corporativa

Cristal Karina Galindo Durán[1], R. Carolina Medina-Ramírez[2], y Mihaela Juganaru Mathieu[3]

[1,2] Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana, Unidad Iztapalapa, San Rafael Atlixco 186, Vicentina, 09340 Ciudad de México, D.F., e-mail[1]-cdgalindod@gmail.com, e-mail[2]- cmed@xanum.uam.mx
[3] Institut H. Fayol, École Nationale Supérieure des Mines de Saint Etienne, 158, cours Fauriel, 42023, Saint Etienne Cedex 2, France, e-mail - mathieu@emse.fr.

Resumen. En el presente artículo mostramos una metodología y una serie de pruebas para obtener descripciones significativas para los recursos de tipo documento de una memoria corporativa. Este enfoque se centra en la extracción de las características relevantes relacionadas al contenido de documentos. De este proceso se pueden obtener elementos que pueden servir como parte de índices de información que contribuyan a una recuperación rápida de información; así como la generación de matrices de frecuencias para la aplicación de diversos algoritmos de agrupamiento o minería de texto.

Palabras clave: Memoria corporativa, descripciones significativas, índices, minería de texto.

1. Introducción

El conocimiento dentro de una organización se denomina memoria corporativa (MC) o memoria organizacional (MO) y se define como: la representación explícita, tácita, consistente y persistente del conocimiento en una organización [1]

La memoria de una organización va creciendo y evolucionando cada día, esta memoria se compone de datos provenientes de las bases de datos, documentos textuales (imprimibles), documentos multimedia, personas, por mencionar algunos.

La gestión de los documentos[1] de una organización en el sentido más amplio impone poder guardar de manera permanente los documentos, poder consultarlos fácilmente a cada momento y también poder buscar información en colecciones particulares de documentos, o buscar documentos o personas sobre un tema, un concepto o buscar documentos/personas relacionados (vinculados) entre ellos [2].

[1] Por documento vamos a comprender una unidad conceptual de tratamiento completo; varios documentos van a formar colecciones y las colecciones conforman la memoria de la organización.

Para facilitar las actividades dentro de una organización se hace necesaria la implementación de aplicaciones inteligentes que con la ayuda de recursos semánticos del tipo ontología, descripciones semánticas de los recursos de una memoria [3], así como axiomas y razonadores mejoren los procesos para la gestión de recursos de información dentro de la organización, sin importar si ésta se encuentra centralizada o distribuida (nube corporativa).

Para ello es necesaria la implementación de índices que describan e identifiquen el documento[4], de tal forma que se pueda registrar ordenadamente los temas de los que trata dicho documento con el fin de permitir una clasificación, consulta y recuperación de los mismos.

Por otro lado la indexación semántica [5] implica que adicionalmente se utilicen herramientas de la gestión del conocimiento (técnicas de procesamiento de lenguaje natural, vocabularios controlados, tesauros y ontologías) y herramientas de la Web semántica [6], para encontrar los conceptos que representen con mayor precisión los documentos, de tal forma que su indexación permita consultas con un alto grado de relevancia (utilidad) para el usuario. En nuestro caso proponemos un índice semántico conformado por tres partes importantes: datos bibliográficos, contenido y vínculos de los términos contenidos en un documento en relación con otros términos contenidos en otros documentos, esto realizado con Datos Abiertos Enlazados (DAE). En la figura 1 se presenta la propuesta de nuestro índice semántico, el cual sirve para caracterizar el documento en tres ejes principales.

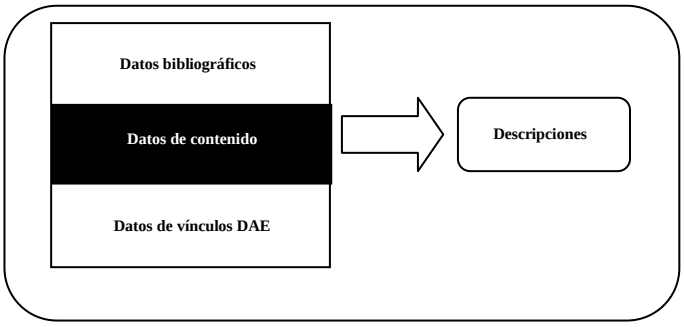

Fig. 1. Elementos que conforman el índice semántico propuesto

El interés de este artículo es centrarse en la obtención de los datos de contenido que describan el recurso de información, la cual se realiza mediante técnicas de procesamiento natural de lenguaje, así como minería de texto, la cual nos permite obtener patrones importantes en el recurso de forma automática y las cuales se pueden utilizar para la caracterización del mismo.

El resto del documento se estructura de la siguiente manera: la sección 2 presenta la metodología utilizada para la obtención de descripciones. En la sección 3 se presenta pruebas realizadas. Finalmente, la sección 4 presenta algunas conclusiones y trabajo futuro.

2. Metodología utilizada para la obtención de descripciones

La metodología aplicada corresponde a una parte del proceso global de la minería de texto, la cual consiste en 4 grandes etapas [8]: 1) Preparación de texto 2) Búsqueda de información, 3) Extracción de información y 4) Minería de texto.

Las etapas no son precisamente secuenciales, el resultado de la minería impone, generalmente, cambios de parámetros (como distancias, criterios) o de métodos dictados por la interpretación de los resultados en la fase de post-minería, pero la fase de preparación de texto es muy cara en términos de tiempo de tratamiento[2]. Porque la fase de preparación de texto es esencial, lo mejor es guardar de manera permanente toda la información que puede servir.

La fase del preparación de texto tiene como objetivo convertir la base de recursos documentales de entrada a un conjunto de palabras (*tokens*) significativas reduciendo así, el número de datos a analizar. El contenido de un documento se transforma en valores, para poder aplicar técnicas de minería de texto o en índices para poder extraer o encontrar fácilmente los documentos relacionada con una búsqueda de información.

La secuencia de los pasos que sigue la aplicación desarrollada para la preparación del texto no es fija, los pasos que se presentan se pueden realizar en orden, sin embargo se pueden regresar algunos pasos anteriores, para realizar los ajustes necesarios. Estos pasos son:

1. Conversión a texto plano: Conversión de los recursos documentales de la memoria corporativa a un formato de texto plano (.txt).
2. Análisis léxico: Basado en herramientas de procesamiento de lenguaje natural como tokenizadores, los cuales separan las palabras. En esta fase se puede eliminan los caracteres especiales como: #, $, %, :-, entre otros.
3. Aplicación de la lematización (*stemming*): Permiten la reducción de las palabras a su forma básica o raíz, por ejemplo, eliminando las partes no esenciales de los términos como prefijos y sufijos o derivando las palabras en plural a su raíz en singular.
4. Eliminación de palabras vacías (*stopword removal*): Busca eliminar los términos con poco significado en la recuperación de información como: pronombres, partículas interrogativas y ciertas preposiciones.
5. Generación de matrices de frecuencia: Se genera un archivo con el tipo de matriz que se desee (frecuencias originales, matriz de términos, matriz inversa de términos, etc.), dependiendo de la métrica requerida.

En la figura 2 se pueden observar de forma gráfica los pasos descritos anteriormente para la preparación de texto de los documentos.

Dichos pasos conforman el proceso de preparación de texto que permite generar datos de calidad y conducir a patrones o reglas de calidad [9].

[2] El tiempo para la preparación de texto es largo en razón del tamaño de las colecciones, del uso de herramienta de análisis gramatical y de la construcción de los índices.

Para los pasos 1, 2,4 y 5 se realizan mediante una implementación en Java y para el paso 3 se hace necesaria la utilización de TreeTagger, la cual es una herramienta de análisis léxico que determina la categoría gramatical de las palabras, así como la forma invariante de dichas palabras [10]. Para obtener la matriz de frecuencia se realiza de igual forma con Java, sobre esta matriz se pueden obtener reglas de agrupamiento y clasificación, árboles de decisión, modelos de regresión o tendencias [11].

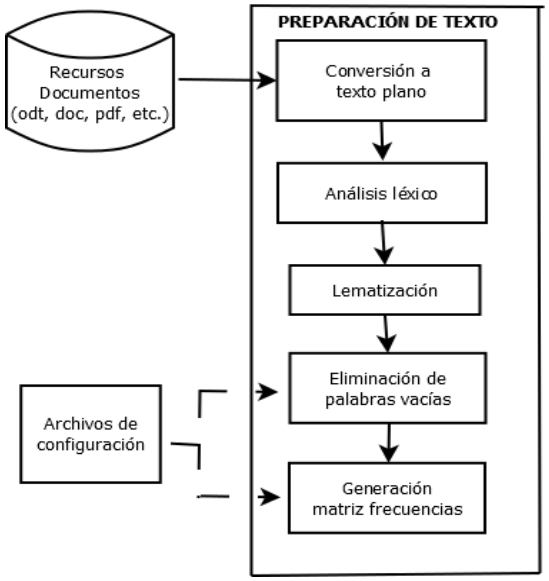

Fig. 2. Etapas de la preparación de texto.

3. Pruebas realizadas

Las pruebas se realizaron sobre dos colecciones de documentos, la primera colección poseía 34 documentos sobre el área de computación específicamente sobre ontologías, la segunda colección posee 150 documento y es más heterogénea abarcado diferentes dominios, ya que posee documentos de minería (texto, datos y XML), Web Semántica y sobre Cursos Masivos en Línea y Abiertos todos los documentos considerados se encuentran en idioma inglés. Se realizaron alrededor de 4 pruebas antes de saber qué *tokens* eran importantes. El tiempo promedio de tratamiento de un documento en formato pdf de 55Kbytes es alrededor de 1.2 minutos para la primera prueba, para la cuarta prueba ese mismo documento es tratado en 50 segundos; aunque la aplicación en la cuarta prueba se realiza una reducción de tiempo esto está relacionado con la reduccion del número de los tokens que se guardan. A continuación se describen dichas pruebas.

3.1.1. Primera prueba

Se consideró utilizar la primera colección, discriminando algunos tipos gramaticales como: preposiciones o conjunción subordinante (*until, before, after, etc.*), conjunción de coordinación (*and, but, nor, or, yet, etc.*), determinador (*a, an, every, not, the,* etc.), determinador wh (*which),* Pronombre wh (*who, what and whom*), adverbío wh (*how, where, why, etc*.) pronombres personales, pronombres posesivos, símbolos (expresiones que no pertenecen al idioma inglés), así como las palabras desconocidas (<*Unknown*>) y la preposición *to.*
En la tabla 1 se presentan las frecuencias más altas obtenidas en la primera prueba, en la primera columna se encuentran los tokens y en la segunda columna las frecuencias.

Tabla 1. Frecuencias más altas de la primera prueba

Tokens	**Frecuencia**
Be	19106
Ontology	4281
Use	3896
Annotation	3046
Have	2888

3.1.2. Segunda prueba

Para esta prueba se considero la segunda colección de 150 documentos, discriminando los mismos tipos gramaticales que en la prueba anterior; sin embargo nos pudimos dar cuenta que aún con la discriminación de esos tipos, obteníamos tipos gramaticales no deseados como: existencial *there*, modales (*can, might, may, should, must,* ought, shall, etc.), terminaciones posesivas, superlativos y comparativos. Por lo que se procedió únicamente a considerar 4 tipos gramaticales: adjetivos, pronombres, adverbios y verbos, ya que de acuerdo al diccionario Británico [12] y la Universidad de Oregon [13] de los ocho tipos gramaticales del idioma inglés, los básicos e importantes son los antes mencionados.

3.1.3. Tercera prueba

Durante las pruebas anteriores nos dimos cuenta que cuando TreeTagger procesaba los sustantivos, en muchos de los casos la forma base era desconocida, es decir, del 100% de los sustantivos el 40% eran desconocidos, por lo que resultaban frecuencias muy altas con el lema *Unknown*, ya que no eran reconocidos. Para solucionar este problema procedimos a indicarle en la aplicación en Java, que cuando fuese un sustantivo tomara la palabra, tal cual.

3.1.4. Cuarta prueba

En esta prueba procedimos a discriminar los correos electrónicos, las referencias y las citas en estilo Chicago y APA, ya que muchos de los documentos tenían esta variedad de citas. También se procedió a discriminar el tipo gramatical Adverbio, por considerar que no aporta ninguna información para obtener las palabras importantes, por lo que los tipos gramaticales que se están considerando son 3: sustantivos, verbos y adjetivos. Por otra parte se validó el caso particular cuando en las líneas se presenta una palabra segmentada por un guión, por lo que se procedió a eliminar el guión y concatenar las siguientes sílabas para formar así una sola palabra.

En la tabla 2 se presentan los resultados obtenidos de las frecuencias más altas, en la primera columna se encuntran los tokens y en la segunda la frecuencia

Tabla 2. Frecuencias más altas de la cuarta prueba

Tokens	**Frecuencia**
Data	2765
Ontology	1775
Web	1614
Information	1533
MOOC's	1344

4. Conclusiones y trabajos futuros

En este documento se describe una propuesta para la obtención de descripciones significativas de los recursos de información (documentos y personas) de una memoria corporativa. Se enfatiza en el pre-tratamiento de las colecciones de una memoria corporativa la cual contiene documentos e informaciones en varios dominios. Este particular, pone de manifiesto la importancia de extraer las características más relevantes del contenido de un documento, ya que sin la discriminación adecuada se corre el riesgo de obtener resultados erróneos para las etapas futuras (conformación de descripciones, aplicación de algoritmos de agrupamiento, obtención de conceptos, etc.).

Como trabajo futuro se considera realizar pruebas sobre una memoria educativa cuyos recursos documentales estan en diferentes idiomas: español, inglés y francés. Por otro lado, aplicar la metodología descrita en este documento para la obtención automática de conceptos representativos de un dominio y compararlos con vocabularios conceptuales (ontologías) existentes. Lo anterior con el fin de poner a punto la generación de índices semánticos en donde las descripciones significativas son una parte medular de los mismos.

Referencias

1. Gandon L., Fabien. "Ontology Engineering: a Survey and a Return on Experience". Technical Report RR-4396, INRIA, Marzo 2002.
2. Gilli, Juan J. "Diseño organizativo: estructura y procesos". Ed. Granica, 2007.
3. Erik Alarcón Zamora. "Integración Semántica de los Recursos de una Memoria Corporativa". Tesis de maestría en Tecnologías de la Información, UAM-Iztapalapa.
4. Métodos Para el Análisis de Documentos, determinación de su Contenido y Selección de los Términos de Indización NC- ISO 5963: 2000. 2014. URL: http://www.sld.cu/galerias/pdf/sitios/centromed/nc_iso_5963_metodos_para_el_analisis_de_documentos,_determinacion_de_sucontenido_y_seleccion_de_terminos_de_indizacion.pdf.
5. Niño Zambrano, M.A., Jimena, Pérez, D., Pezo, D.M. "Procesamiento para la Construcción de Índices Semánticos Basados en Ontologías de Dominio Específico". Entramado. Vol. 9, No.1, pp. 262-287. Enero 2013.
6. Galindo Durán, C. K., Medina-Ramírez, R. C., Juganaru-Mathieu, M. "Using Linked Open Data to Enrich a Corporate Memory of Universities". Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government EEE'14.Hamid R. Arabnia Azita Bahrami Fernando G. Tinetti. ISBN 1-60132-268-2 2014.
7. Metadata Standards. 2014. URL: http://www.metadataetc.org/bookwebsite/readings/appendixaschemas.htm
8. Montes y Gómez, Manuel. "Minería de texto: un nuevo reto computacional", 2014. URL: http://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf
9. Liu, H., Motoda, H., "Feature Extraction, Construction and Selection: A Data Mining Perspective". Kluwer Academic, 1998.
10. Sitio oficial TreeTagger, 2014. URL: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
11. Srivastava, A. N. y Mehran, S. "Text Mining Classification, Clustering" and Applications. CRC Press Taylor & Francis Group, 2009.
12. Sitio oficial Cambridge Dictionaries, 2014. URL: http://dictionary.cambridge.org/dictionary/british/part-of-speech
13. A brief grammatical sketch of English, 2014. URL: http://pages.uoregon.edu/tpayne/engram/Engramsection03.pdf