Categorización Automática de Documentos

M. Alicia Pérez Abelleira, Alejandra Carolina Cardoso[1]

[1] Facultad de Ingeniería e Informática e IESIING. Universidad Católica de Salta
Campo Castañares s/n, A4400 Salta, Argentina
{aperez,acardoso}@ucasal.net

Abstract. La clasificación de documentos de texto es una aplicación de la minería de textos que pretende extraer información de texto no estructurado. Su interés se justifica porque se estima que entre el 80% y el 90% de los datos de las organizaciones son no estructurados. Por otro lado, la búsqueda semántica permite al usuario especificar en una consulta no solamente términos que deben aparecer en el documento, sino conceptos y relaciones, que pueden detectarse mediante el análisis de texto. El objetivo de este trabajo es implementar un buscador semántico que aproveche el resultado de algoritmos de aprendizaje automático supervisado y semi-supervisado para la categorización o clasificación de documentos. El dominio de aplicación es un corpus de más de 8000 documentos que contienen nueve años de resoluciones rectorales de la Universidad Católica de Salta en distintos formatos.

Keywords: categorización de documentos, buscador semántico, aprendizaje semisupervisado, minería de texto, UIMA.

1 Introducción

El conocimiento es cada vez más un recurso de importancia estratégica para las organizaciones y su generación, codificación, gestión, divulgación aportan al proceso de innovación. Todos estos aspectos se incluyen en la llamada *gestión del conocimiento*. La cantidad de documentos de diversos tipos disponibles en una organización es enorme y continúa creciendo cada día. Estos documentos, más que las bases de datos, son a menudo un repositorio fundamental del conocimiento de la organización, pero a diferencia de éstas la información no está estructurada. La minería de textos tiene como objetivo extraer información de texto no estructurado, tal como entidades (personas, organizaciones, fechas, cantidades) y las relaciones entre ellas. Por otro lado, la búsqueda semántica permite al usuario especificar en una consulta no solamente términos que deben aparecer en el documento, sino esas entidades y relaciones extraídas mediante el análisis de texto.

La categorización de documentos de texto es una aplicación de la minería de texto que asigna a los documentos una o más categorías, etiquetas o clases, basadas en el contenido. Es un componente importante de muchas tareas de organización y gestión de la información. El enfoque tradicional para la categorización de textos en que los expertos en el dominio de los textos definían manualmente las reglas de clasificación ha sido reemplazado por otro basado en técnicas de aprendizaje automático, o en combinaciones de éste con otras técnicas.

Nuestro trabajo se centra en desarrollar técnicas para la categorización automática de documentos según su contenido que avancen el estado del arte en nuestro medio, aplicando el aprendizaje automático a la minería de texto. El objetivo final es implementar un buscador semántico que aproveche el resultado de algoritmos de aprendizaje para la clasificación de documentos. El dominio de aplicación es un corpus de más de 8000 documentos que contienen 9 años de resoluciones rectorales de la Universidad Católica de Salta en distintos formatos (Word, texto plano, PDF).

Este artículo comienza describiendo la información no estructurada, repositorio fundamental del conocimiento de una organización, y las arquitecturas para la gestión de información no estructurada. La Sección 3 muestra nuestra instanciación del modelo general para el problema de la clasificación y búsqueda de resoluciones rectorales. En la Sección 4 se exploran diferentes algoritmos para la categorización de documentos y se describen los experimentos realizados para determinar su adecuación a nuestro dominio. Concluye el trabajo con el funcionamiento del motor de búsqueda semántica (Sección 5) y algunas conclusiones.

2 Información Estructurada y No Estructurada

La información estructurada se caracteriza por tener un significado que pretende no ser ambiguo y que está representado explícitamente en la estructura o formato de los datos. El ejemplo típico es una base de datos relacional. De la información no estructurada podría decirse que su significado no está implicado en su forma y por tanto precisa interpretación para aproximar o extraer el mismo. Los documentos en lenguaje natural o hasta de voz, audio, imágenes, etc. entran en esta categoría. El interés por extraer significado de la información no estructurada se debe a que se estima que entre el 80% y el 90% de los datos de las organizaciones son no estructurados [1]. Aunque muchas organizaciones han invertido en tecnologías para minería de datos estructurados procedentes de sus bases de datos y sistemas transaccionales, en general no han intentado capitalizar sus datos no estructurados o semi-estructurados.

Una aplicación de gestión de la información no estructurada (UIM por sus siglas en inglés) típicamente es un sistema que analiza grandes volúmenes de información no estructurada con el fin de descubrir, organizar y entregar conocimiento relevante al usuario final. La información no estructurada puede ser mensajes de correo electrónico, páginas web o documentos generados con una variedad de procesadores de texto, como en el caso de las resoluciones rectorales de nuestra universidad. Estas aplicaciones utilizan para el análisis una variedad de tecnologías en las áreas del procesamiento del lenguaje natural, recuperación de la información, aprendizaje automático, ontologías y hasta razonamiento automático.

El resultado del análisis generalmente es información estructurada que se hace accesible al usuario mediante aplicaciones adecuadas. Un ejemplo puede ser la generación de un índice de búsqueda y la utilización de un buscador que facilita el acceso a documentos de texto por tema, ordenados según su relevancia a los términos o conceptos de la consulta del usuario.

Existen diversas arquitecturas para el desarrollo de aplicaciones UIM. Para nuestro trabajo hemos utilizado UIMA [2], una arquitectura software basada en componentes que surgió como proyecto de investigación de IBM y fue puesta a disposición de la comunidad como software libre.

3 Arquitectura del Sistema

Conceptualmente suele verse a las aplicaciones de UIM con dos fases: una de análisis y otra de entrega de la información al usuario. En la fase de análisis se recogen y analizan colecciones de documentos. Los resultados del análisis se almacenan en algún lenguaje o depósito intermedio. La fase de entrega hace accesible al usuario el resultado del análisis, y posiblemente el documento original completo mediante una interfaz apropiada. La Fig. 1 muestra la aplicación de este esquema a nuestro dominio, en el que partimos de más de 8000 resoluciones rectorales en archivos de texto de distinto tipo: Word, PDF, texto plano. Previo al análisis, se procede a la extracción del texto de cada archivo utilizando las herramientas de software libre POI (poi.apache.org) y *tm-extractors* (www.textmining.org). También se divide en partes la resolución extrayendo el encabezado (texto que contiene el número y la fecha de la resolución) y el cuerpo con la mayor parte de la información, y descartando en lo posible el texto "de forma".

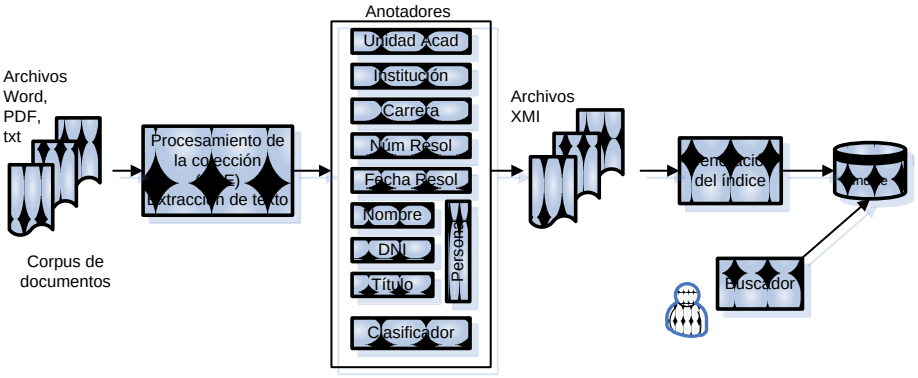

Fig. 1. Arquitectura del sistema.

La fase de análisis incluye tokenización y detección de entidades en documentos individuales tales como personas, fechas, organizaciones, unidades académicas y datos sobre la resolución (fecha y número). Además con la ayuda de un clasificador aprendido automáticamente del corpus de resoluciones, como se explica en la Sección 4, se anota cada documento con una categoría. Existen 21 categorías que fueron obtenidas del personal especializado en la elaboración de resoluciones. Algunos ejemplos son: designación de planta docente, convenio de pasantías, convenio de colaboración, llamado a concurso docente, o designación de tribunal de concurso.

El resultado de la fase de análisis es un conjunto de archivos en formato XMI (Sección 3.3). Estos archivos contienen, además de las partes relevantes del texto original, metadatos en forma de anotaciones correspondientes a las entidades y a la

categoría de documentos. Estos archivos serán procesados para construir el índice de un motor de búsqueda que contiene los tokens (en nuestro caso, las palabras que aparecen en el texto) y las entidades y categorías extraídas automáticamente.

En la fase de entrega existe una interfaz para hacer consultas de búsqueda en el índice de forma que el usuario pueda buscar documentos que contengan combinaciones booleanas de tokens, entidades y categorías mediante un motor de búsqueda semántica.

3.1 Análisis a Nivel de Documento

En UIMA, el componente que contiene la lógica del análisis se llama anotador. Cada anotador realiza una tarea específica de extracción de información de un documento y genera como resultado anotaciones, que son añadidas a una estructura de datos denominada CAS (*common analysis structure*). A su vez, esas anotaciones pueden ser utilizadas por otros anotadores. Los anotadores pueden ser agrupados en anotadores agregados.

La mayoría de los anotadores de nuestro sistema realizan reconocimiento de entidades con nombre (NER), a saber: personas, unidades académicas, carreras, instituciones, fechas, número y año de las resoluciones. Además, para detectar entidades correspondientes a personas se agregan otras (nombres propios, DNIs y títulos) obtenidas por los anotadores correspondientes.

Las técnicas utilizadas para el reconocimiento de entidades son [3,4]:

- Equiparación con expresiones regulares que capturan el patrón que siguen las entidades (ejemplos son la detección de DNIs, fecha y número de las resoluciones).
- Equiparación con diccionarios y *gazetteers* (ejemplos son las carreras, unidades académicas, instituciones, títulos y nombres propios). El diccionario de nombres propios consta de más de 1300 nombres y fue extraído automáticamente del sistema de gestión de alumnos. El enfoque basado en componentes de UIMA nos ha permitido adaptar el *Gazetteer Annotator* de Julie Lab [5] basado en la implementación que hace *Lingpipe* del algoritmo Aho-Corasick [3].
- Equiparación con plantillas: para detectar entidades correspondientes a personas se utiliza una plantilla que describe a la persona mediante los siguientes atributos: nombre1, nombre2, apellido(s), DNI, título. Sólo nombre1 y apellido(s) son obligatorios. Todos los atributos son a su vez entidades detectadas por anotadores.

Además de los anotadores mencionados, se utiliza el anotador de UIMA que detecta tokens usando una sencilla segmentación basada en los espacios en blanco, y crea anotaciones con tokens y sentencias.

Aparte de todos estos anotadores, existe otro que asigna la categoría de documento en base al modelo aprendido automáticamente, como se describe en la Sección 4

3.2 Análisis a Nivel de Colección

Además se realiza análisis al nivel de la colección de documentos. Un caso particular es un bucle de realimentación, que produce recursos estructurados a partir del análisis de una colección de documentos y luego usa esos recursos para permitir el