# ¿Cómo hacer perfiles de documentos?

## *How to profile documents?*

**Antonio Guillén Espejo**
Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
Carretera San Vicente del Raspeig S/N, 03690 San Vicente del Raspeig
aguillen@dlsi.ua.es

**Resumen:** En la actualidad muchos usuarios navegan por Internet a través de gran cantidad de sitios webs. En ocasiones, se realiza con el objetivo de encontrar documentos específicos (por ejemplo, noticias, blogs, etc.) o documentos de una categoría concreta. ¿Cómo hacer que los motores de búsqueda puedan identificar apropiadamente aquellos documentos de acuerdo con las necesidades de los usuarios? Se hace necesario encontrar una forma correcta de extraer la información que pueda caracterizar a los documentos. Esta información puede ser: tópicos, polaridad, áreas de interés, etc. Las Tecnologías del Lenguaje Humano (TLH) son capaces de obtener automáticamente esta información y representarla como meta-datos. En esta investigación, proponemos definir un modelo de perfiles de documentos capaz de apoyar principalmente la búsqueda de documentos (entre otros objetivos) mediante el estudio de las TLH y los documentos más comunes en Internet.
**Palabras clave:** Perfil de Documento, Perfil de Usuario, Clasificación de Documentos, Extracción de Información

**Abstract:** Nowadays, users browse on the Internet through a huge amount of websites. In many cases, it is performed with the aim of finding specific documents (e.g. news, blogs, etc.) or depending on a concrete category. How to improve search engines to be able properly find out documents taking into account the users' needs? To this aim, it is necessary to find out a proper way of extracting information from documents and offering useful meta-data for their profiling. For example, these meta-data can be topics, sentiment polarity, subject areas and many other features that can be extracted thanks to the support of Human Language Technologies (HLT). In this research, we propose the definition of a document profiling model through the study of HLT in conjunction with different document types commonly found on the Internet.
**Keywords:** Document Profile, User Profile, Document Classification, Information Extraction

### *1 Justificación de la investigación propuesta*

Internet está creciendo considerablemente desde la Web 2.0, uno de los motivos es la posibilidad que tiene cualquier usuario para crear y publicar sus propios contenidos. Estos contenidos pueden ser opiniones, ya sea en blogs, redes sociales, pero también opiniones acerca productos comprados por Internet, vídeos, noticias, etc. Este tipo de contenidos se pueden considerar como documentos de los cuales podemos obtener información útil usando las Tecnologías del Lenguaje Humano (TLH). Esta información asociada al documento en forma de meta-datos puede tener diversos propósitos: el apoyo a la clasificación de documentos, el apoyo a los motores de búsqueda, mejorar los sistemas de recomendación, pero también el apoyo a algunas áreas del Procesamiento del Lenguaje Natural (PLN) como el análisis de sentimientos.

El objetivo de la tesis es el estudio de las TLH apropiadas para creación de perfiles de documentos (análisis semántico, polaridad, complejidad de lectura, etc.). En nuestro trabajo, definimos el término documento como una unidad de información y contenido proveniente de Internet, ya sea de mayor o menor tamaño, y de diversos dominios. Por ejemplo, algunos de los documentos a tratar pueden ser noticias online, posts en redes sociales o foros, blogs, comentarios sobre productos, etc.

Pretendemos que la creación de este perfil sea capaz de representar un documento a fin de dar utilidad y apoyo a ciertas áreas o sistemas concretos. Para ello, se llevará a cabo la creación de un modelo donde se establece qué documentos concretos vamos a tratar, qué información queremos representar, qué TLH se van a usar para extraer esta información, así como los detalles que se han de tener en cuenta en la aplicación de las TLH sobre ciertas clases de documentos (tipo de contenido, tamaño del documento, etc.). Asimismo, se pretende crear una ontología para el adecuado almacenamiento de los meta-datos del perfil, y la implementación de un prototipo capaz hacer un uso automático de esta ontología. Esto ayudará en la tarea de evaluación del modelo y la verificación de utilidad real de los perfiles generados.

## 2 Origen y trabajo relacionado

Podríamos considerar antecedentes de nuestro trabajo los sistemas de integración de TLH. Estos tratan de facilitar el acceso y uso de estas tecnologías por parte de investigadores. En este sentido, existen herramientas como InTime (Gómez Soriano, 2008) que trata de integrar gran cantidad de herramientas TLH para su uso de forma remota e independientemente del sistema operativo. Otra herramienta de integración es TLH Suite (Guillén, Lloret, y Gutiérrez, 2016) que además trata de vincular la información anotada usando diferentes herramientas con el fin de obtener un paquete semántico. Otra aproximación vinculada con los paquetes semánticos (Lloret, Gutiérrez, y Gómez, 2015) trata de la representación del conocimiento diseñando una ontología. Nuestro trabajo de investigación pretende dar un paso más allá de la integración y anotación semántica, aunque estas aproximaciones nos pueden servir como base al modelo de perfiles de documentos que se está diseñando.

Según la literatura, la generación de perfiles suele centrarse en los usuarios, tratando aspectos como la identificación de autoría (Sapkota et al., 2015) y el apoyo a los sistemas recomendación (Bobadilla et al., 2013). A pesar de que nuestro trabajo se orienta a la generación de perfiles de documentos, nos pueden ser útiles estos y otros trabajos debido a que tendremos en cuenta aspectos relacionados con la identificación de autoría, por ejemplo, detectar la edad y genero del autor del documento (Rangel y Rosso, 2016). En referencia a los sistemas de recomendación, un enfoque que se acerca a nuestro trabajo es (Li et al., 2010) que trata de mejorar sistemas de recomendación de noticias teniendo en cuenta la información asociada a la propia noticia, justificando el valor añadido que aporta este tipo de información vinculada.

Existen aproximaciones con características muy similares a la nuestra. (Gulla et al., 2014) propone una aproximación para generar perfiles en base a la interacción de un usuario en Internet para la recomendación de noticias. Uno de nuestros dominios de aplicación también es el de noticias, por lo tanto, se tendrá en cuenta la metodología seguida. (Kshirsagar y Deshkar, 2015) propone un sistema de análisis de *reviews* de productos, con el fin de extraer características tales como la polaridad sentimental. En nuestro trabajo también pretendemos tratar con *reviews*, noticias, etc. pero obteniendo más características de estos documentos.

## 3 Descripción de la investigación propuesta e hipótesis

Se propone un trabajo de investigación orientado a la generación de perfiles de documentos. Inicialmente trabajaremos con documentos en inglés y español, ya que muchas de las TLH a estudiar soportan estos dos idiomas. Consideramos documento a una unidad textual de información proveniente principalmente de Internet. Nuestra hipótesis es que la generación de perfiles de documentos usando las TLH, puede ayudar a aspectos como la búsqueda de documentos o recomendaciones de estos a usuarios, seguimiento en tiempo real de la información tratada en redes sociales, o incluso mejorar resultados en tareas del PLN como el análisis de sentimientos o la clasificación de documentos.

Inicialmente la investigación se enmarca en la definición de un modelo de perfiles de documentos. La definición de este modelo consiste en estudiar y especificar los documentos concretos que vamos a tratar, un estudio de la información que queremos obtener de estos (meta-datos) y el estudio de las TLH más apropiadas para el cálculo de estos meta-datos. En la Figura 1 se puede observar la idea del perfil de documentos: dado un documento, se le aplica las TLH para obtener una serie de meta-datos vinculados al mismo.

Para la definición del modelo, la prime-

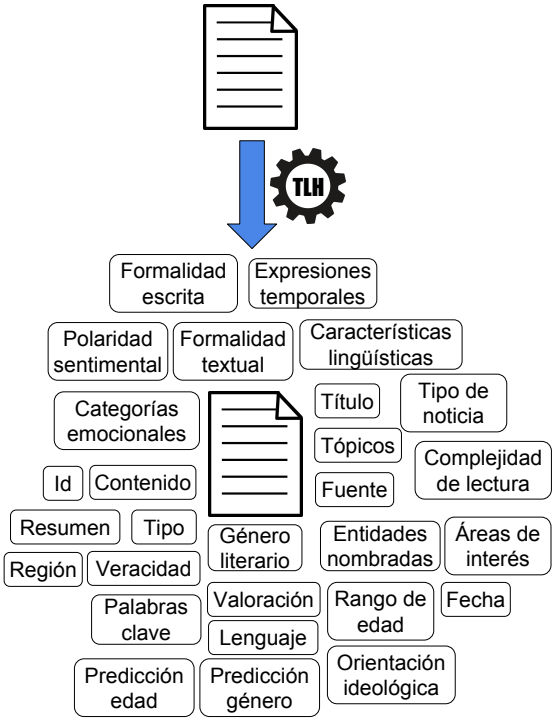


Figura 1: Idea del perfil de documentos.

ra tarea ha sido determinar qué documentos concretos vamos a estudiar. En Internet existen muchos tipos de documentos, en nuestro trabajo hemos querido limitar esta selección a los documentos que habitualmente se consultan en Internet: *noticias online*, *posts en redes sociales*, *reviews de productos o servicios*, *blogs*, *Webs personales*, *documentos académicos*, *documentos científicos*, *manuales de instrucciones*, *tutoriales*, *extractos literarios*, *documentos administrativos/técnicos*.

Posteriormente, se especifica la información concreta (meta-datos) que queremos obtener de los documentos usando las TLH. Esta selección contempla un amplio conjunto de información con el fin de representar adecuadamente al documento, y además, sea capaz de aportar un valor añadido a los propósitos de este trabajo. Por ejemplo, conocer la *complejidad de lectura* o la *formalidad escrita* de un documento ayuda a la identificación de las personas más apropiadas para leer dicho documento.

Se tiene en cuenta que no todos los meta-datos son adecuados para todos los documentos propuestos. Para solventar esto se ha creado el esquema de documentos mostrado en la Figura2. En este esquema en forma de árbol aparecen los documentos propuestos como hojas, y se establecen relaciones entre ellos a través de documentos conceptuales. Los meta-datos son distribuidos en todo el árbol de manera que solo se asocien a la rama adecuada de documentos. Por ejemplo, para los documentos subjetivos se asocia la información referente el análisis de sentimientos e ideología, ya que esta información solo tiene sentido obtenerla con estos tipos de documentos de carácter subjetivo. Asimismo, el esquema de documentos se ha creado con la intención de diseñar una ontología con la que facilitar el almacenamiento y consulta de los perfiles.

Para calcular el valor de los meta-datos se usan algunas de las TLH más relevantes: *Extracción de información*, *Detección de expresiones temporales*, *Detección de entidades nombradas*, *Detección de dominios*, *Clasificación de polaridad*, *Análisis de legibilidad*, etc. Se debe realizar un estudio de las herramientas que hay actualmente disponibles para el cálculo de los meta-datos, teniendo en cuenta criterios como la fiabilidad de la herramienta y el grado de automatización. La fiabilidad se puede medir comprobando los trabajos relacionados de estas herramientas y los resultados obtenidos de su evaluación. Sin embargo, no todas las herramientas disponen de evaluaciones publicadas. El grado de automatización se definirá dado el tipo de herramienta que se trate, por ejemplo, una herramienta en formato *Servicio Web* tendrá un alto grado de automatización ya que es fácilmente incorporable en un prototipo o aplicación.

## 4 Metodología y experimentos propuestos

El estado actual del trabajo de investigación comprende la definición del modelo descrito en la sección anterior y el estudio de herramientas TLH que se podrían incluir en nuestro trabajo. También se está preparando un conjunto de datos (dataset) usando documentos reales de cada tipo del esquema, generando el perfil de cada uno siguiendo el Algoritmo 1.

En este algoritmo, primero se obtiene el contenido textual a partir de la fuente del documento. De este contenido se genera un resumen en el caso de que sea un documento extenso. Este resumen servirá para aquellas herramientas TLH que requieren texto corto para un mejor funcionamiento. Al identificar el tipo de documento de los definidos en el modelo, se obtendrán los meta-datos corres-

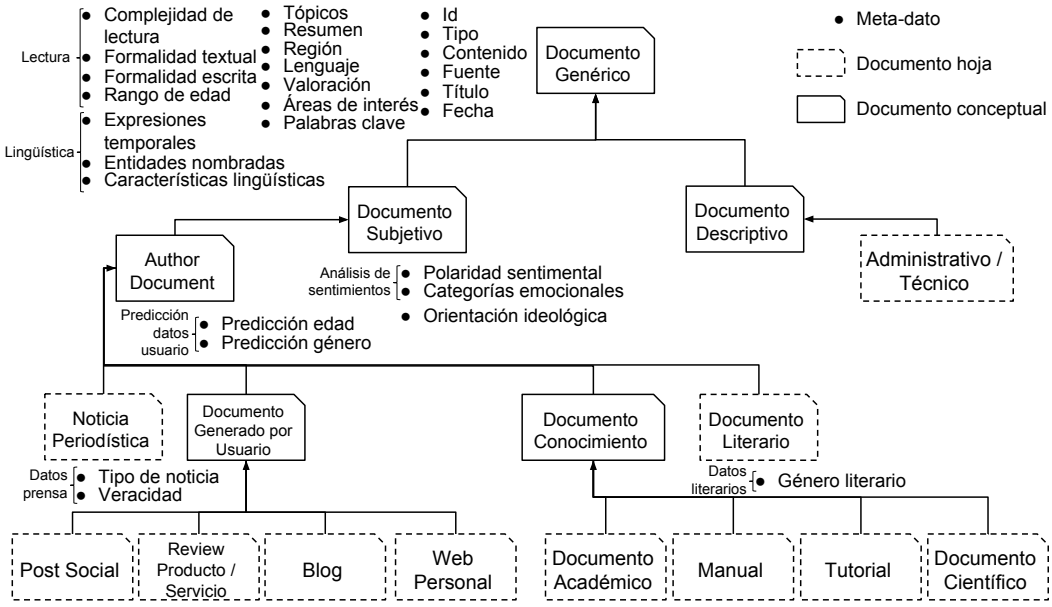


Figura 2: Esquema de documentos y distribución de los meta-datos.

**Algorithm 1** Generar Perfil Documento

**Require:** *url*, url documento Web

```
c ← obtenerContenido(url)
r ← generarResumen(c)
t ← obtenerTipo(c)
perfil ← newProfile(c, r, t)
listaMt ← obtenerMetadatos(type)
for each mt ∈ listaMt do
    tlh ← mt.obtenerTLH()
    valor ← tlh(c, r)
    perfil.añadir(mt, valor)
```

pondientes al tipo. Cada meta-dato se calcula usando su herramienta TLH asociada, sobre el contenido original del documento o sobre el resumen, según requiera la herramienta. Una vez calculado el valor del meta-dato se añade al perfil. Este algoritmo se implementará en un prototipo para generar automáticamente los perfiles y ser expuestos a evaluación o experimentaciones.

Una posible evaluación del modelo consiste en la realización de una encuesta usando para ello algunos de los perfiles generados en el conjunto de datos y una serie de preguntas para valorar los meta-datos obtenidos (por ejemplo, si son correctos estos valores, que información obtenida es más útil, etc.).

La experimentación del trabajo se puede contemplar desde diversos escenarios. Un escenario podría ser comparar la búsqueda de documentos con el apoyo de los perfiles de nuestra aproximación, con respecto a la indexación de documentos habitual. Otro escenario sería la posible mejora de tareas del PLN como la clasificación de documentos, usando los meta-datos del perfil como características de entrada al sistema de clasificación.

## *5 Elementos de investigación específicos propuestos para discusión*

Nuestra propuesta de investigación tiene cierto carácter novedoso, por lo tanto, surgen algunas cuestiones sobre la definición del modelo que se está haciendo, y las posibles evaluaciones y experimentos que se pretenden realizar. Algunas de estas cuestiones podrían tratar los siguientes aspectos:

- Sobre la selección de documentos ¿Es adecuado el criterio seguido? ¿Faltarían o tendrían que descartarse ciertos documentos?
- Sobre la selección de meta-datos ¿Representa adecuadamente a los documentos? ¿Qué meta-datos son más útiles? ¿Se puede mejorar la distribución de los meta-datos en el esquema de documentos presentado?
- Sobre la evaluación del modelo y los experimentos planteados ¿Son correctas las evaluaciones y experimentos que se proponen? ¿Existen otros escenarios donde sería más interesante experimentar?

### *Agradecimientos*

Esta investigación está parcialmente financiada por la Universidad de Alicante a través de una beca del programa de Formación de Profesorado Universitario (UAFPU2015-5999), así como la Generalitat Valenciana, el Ministerio de Educación, Cultura y Deporte, y las Ayudas Fundación BBVA a equipos de investigación científica 2016, a través de los proyectos: TIN2015-65100-R, TIN2015-65136-C2-2-R, PROMETEOII/2014/001, GRE16-01: "Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet" y Análisis de Sentimientos Aplicado a la Prevención del Suicidio en las Redes Sociales (ASAP).

### *Bibliografía*

Bobadilla, J., F. Ortega, A. Hernando, y A. Gutiérrez. 2013. Recommender Systems Survey. *Knowledge-Based Systems*, 46:109–132.

Gómez Soriano, J. M. 2008. InTiMe: Plataforma de Integración de Recursos de PLN. *Procesamiento del Lenguaje Natural*, 40:83–90.

Guillén, A., E. Lloret, y Y. Gutiérrez. 2016. TLH Suite: herramienta para la anotación semántica de información. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informação*, 2016(18):99–113.

Gulla, J. A., A. D. Fidjestøl, X. Su, y H. Castejon. 2014. Implicit User Profiling in News Recommender Systems. *International Conference on Web Information Systems and Technologies*, páginas 185–192.

Kshirsagar, A. A. y P. A. Deshkar. 2015. Review analyzer analysis of product reviews on weka classifiers. En *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS 2015)*, páginas 1–5.

Li, Q., J. Wang, Y. P. Chen, y Z. Lin. 2010. User comments for news recommendation in forum-based social media. *Information Sciences*, 180(24):4929–4939.

Lloret, E., Y. Gutiérrez, y J. Gómez. 2015. Developing an ontology to capture documents' semantics. En *IC3K 2015 - Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, páginas 155–162.

Rangel, F. y P. Rosso. 2016. On the impact of emotions on author profiling. *Information Processing and Management*, 52(1):73–92.

Sapkota, U., S. Bethard, M. Montes-y Gómez, y T. Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. En *Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*, páginas 93–102.