ISSN 1695-5498

HIPERTEXT.NET
Revista Académica sobre Documentación Digital y Comunicación Interactiva
DEPARTAMENTO DE COMUNICACIÓN · GRUPO DE INVESTIGACIÓN DIGIDOC

upf. Universitat Pompeu Fabra Barcelona

Número 15 · Noviembre de 2017

Tesis doctoral – Síntesis. Exploración de procedimientos semiautomáticos para el proceso de indexación en el entorno web

Mari Vàllez
Universitat Pompeu Fabra; Universitat Oberta de Catalunya. Barcelona
mari.vallez@gmail.com
https://marivallez.com

PhD dissertation – Summary. Exploration of semiautomatic procedures for the indexing process in the web environment

RESUMEN

La ingente cantidad de información que existe actualmente hace necesario el desarrollo de herramientas, métodos y procesos que faciliten el acceso a la información. Las técnicas de indexación cuentan con una larga tradición en este ámbito. Sin embargo, su aplicación a gran escala y en el contexto de la Web no siempre es viable por la magnitud y la heterogeneidad de la información presente en ella. En esta tesis se presentan dos propuestas para facilitar el proceso de indexación de documentos en Internet. La primera se caracteriza por el uso de técnicas de indexación semiautomáticas basadas en aspectos de posicionamiento web, que se aplican a través de una herramienta propia denominada DigiDoc MetaEdit. La segunda propone un modelo para la actualización de vocabularios controlados a partir del procesamiento de los logs de las búsquedas formuladas por los usuarios en los buscadores.

PALABRAS CLAVE

Indexación, Anotación semántica, Vocabulario controlado, Web semántica, Metadatos, Recuperación de información, Logs de consultas, Palabras clave

ABSTRACT

The vast amount of information that currently exists necessitates the development of tools, methods and processes that facilitate access to it. Indexing techniques have a long tradition of promoting the improvement of these systems. However, its application on a large scale and in the context of the Web is not always feasible because of the magnitude and diversity of the information in it. This thesis presents two proposals to facilitate the process of indexing documents on the Internet. The first is characterized by the use of semiautomatic indexing techniques based on aspects of SEO, and applied through a proprietary tool called DigiDoc MetaEdit. The second proposes a model for updating controlled vocabularies from the processing of logs of searches made by users on search engines.

KEYWORDS

Indexing, Semantic annotation, Controlled vocabulary, Semantic Web, Metadata, Information retrieval, Query logs, Keywords

Vàllez, M (2017). Tesis doctoral – Síntesis. Exploración de procedimientos semiautomáticos para el proceso de indexación en el entorno web. *Hipertext.net*, n. 15, p. 91-99. DOI: 10.2436/20.8050.01.50

https://dx.doi.org/10.2436/20.8050.01.50

1. Introducción

Internet es un gran universo de contenidos donde no siempre resulta fácil localizar la información pertinente para cada necesidad. Ello se debe principalmente a dos de sus singularidades: la facilidad para la creación de contenidos y la enorme cantidad de información disponible. Hay que tener en cuenta que actualmente el volumen de información digital se duplica cada dos años (IDC, 2014).

A pesar de todo, aún falta un gran camino por recorrer para poder satisfacer las necesidades de información más complejas, aunque los motores de búsqueda cada día ofrecen una experiencia de búsqueda al usuario más gratificante. La solución actual es ofrecer un listado de resultados para que el usuario los ojee y elija aquel que considere más relevante para resolver su necesidad de información. Probablemente, la situación ideal sería proporcionar directamente una respuesta o dar una lista acotada de resultados totalmente pertinentes para la necesidad de información formulada.

La propuesta de la Web semántica de Berners-Lee (2001) iba por este camino. Proponía un cambio de paradigma: transformar la actual web basada en el lenguaje natural en una Web estructurada, donde los contenidos sean etiquetados semánticamente de forma explícita para conseguir que los programas informáticos puedan interpretarlos. De esta forma, se facilitaría el procesamiento de los contenidos y la recuperación de información sería más pertinente (Ding et al., 2005).

La asignación de metadatos es uno de los elementos básicos del proyecto de la Web semántica. Implica una nueva forma de crear contenidos, donde los metadatos deben incorporarse para facilitar su posterior procesamiento. En este contexto surge la necesidad de herramientas que faciliten la anotación semántica y la asignación de meta-información de calidad.

La representación del contenido de un documento con metadatos es una práctica con una larga trayectoria. Desde sus inicios, los sistemas de recuperación de información han utilizado este método para facilitar el acceso a la información, ya que es una forma compacta y eficiente de representar el contenido de un documento.

Este proceso se conoce con el nombre de indexación o indización, y aunque las propuestas de sistemas automáticos están presentes desde hace décadas (Spärck Jones, 1974; Stevens, 1970), aún existe una arraigada tradición de realizar el proceso manualmente porque con frecuencia se cuestiona la eficiencia de muchas propuestas automáticas.

La investigación presentada en esta tesis intenta fusionar estos dos ámbitos, la indexación y la anotación semántica. Los dos comparten un mismo objetivo aunque enfocado desde distintas perspectivas (Hou, Gu, & Zhou, 2015; Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004). La anotación semántica está vinculada a la lingüística y al procesamiento del lenguaje natural (Vàllez, 2009; Vàllez, & Pedraza-Jiménez, 2007). En cambio, la indexación está asociada a la disciplina de la Biblioteconomía y la Documentación.

Esta tesis tiene por objetivo explorar procedimientos semiautomáticos adaptados al entorno web para:

- mejorar el proceso de indexación
- actualizar los vocabularios controlados utilizados en el proceso de indexación

Concretamente, se pretende identificar técnicas que ayuden a optimizar el proceso de indexación, así como a mejorar el mantenimiento de las principales herramientas que intervienen en el proceso. En este sentido, se exploran técnicas para la extracción de palabras clave de documentos web que puedan utilizarse como descriptores. Además, también se formula una propuesta para actualizar y mantener los vocabularios controlados procesando un conjunto de datos (query logs) obtenidos a partir de la interacción de los usuarios con los documentos indexados.

Las principales contribuciones que se presentan en esta tesis se pueden resumir en:

1. Técnicas para actualizar y mejorar los procesos de indexación teniendo en cuenta las características del entorno web.
2. Modelo para actualizar y mejorar los vocabularios controlados a partir del análisis de los logs de consultas.
3. Aportaciones conceptuales al ámbito teórico de la indexación en el entorno web.

A continuación se citan los cuatro artículos que configuran la investigación realizada y que forman la tesis:

1. Vàllez, M., Rovira, C., Codina, L., & Pedraza-Jiménez, R. (2010). Procedimientos para la extracción de palabras clave de páginas web basados en criterios de posicionamiento en buscadores. *Hipertext.net*, núm.8.
2. Vàllez, M. (2011). *Keyword Research*: métodos y herramientas para identificar las palabras clave. *BiD: textos universitaris de biblioteconomia i documentació*, núm. 27.
3. Vàllez, M., Pedraza-Jiménez, R., Blanco, S., Codina, L., & Rovira, C. (2015). A semi-automatic indexing system based on embedded information in HTML documents. *Library Hi Tech, 33*(2), 195-210.
4. Vàllez, M., Pedraza-Jiménez, R., Blanco, S., Codina, L., & Rovira, C. (2015). Updating controlled vocabularies by analysing query logs. *Online Information Review, 39*(7).

La Figura 1 muestra de forma gráfica la conexión que existe

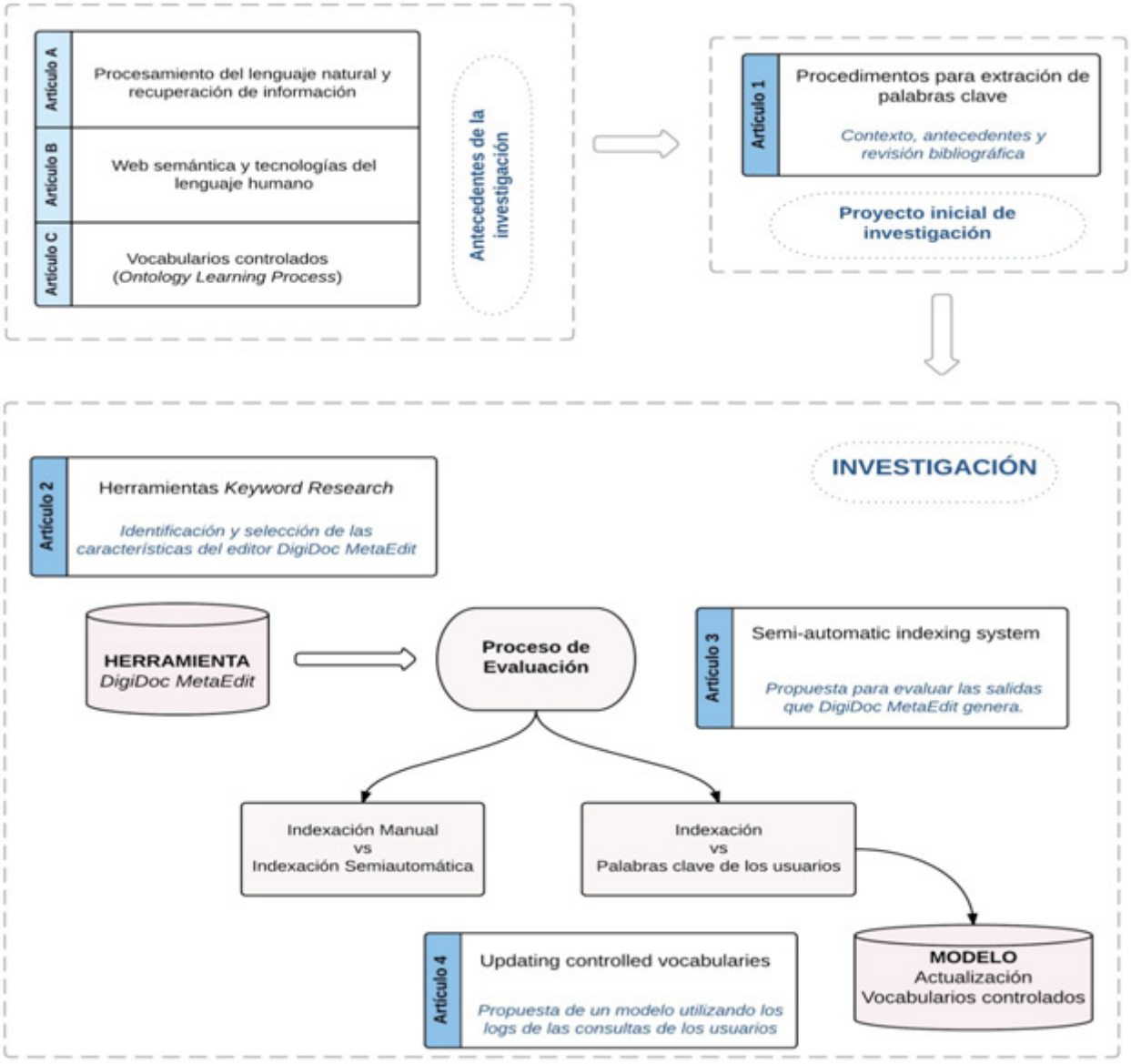

Figura 1: Estructura de la investigación realizada en la tesis.

entre los diferentes artículos que configuran la tesis y las investigaciones anteriores que son los antecedentes.

2. Marco teórico

Partiendo del contexto anterior, el marco teórico en que se sustenta la investigación desarrollada en esta tesis se encuadra en tres ámbitos diferentes pero estrechamente vinculados entre sí: i) el proceso de indexación, ii) la anotación semántica, y iii) los vocabularios controlados.

2.1. El proceso de indexación

La teoría de la indexación intenta establecer cuál es el proceso de indexación más eficaz, para que éste sea ejecutado como una ciencia más que como un arte (Borko, 1977; Hjørland, 2011). La literatura divide el proceso de indexación en dos etapas principales: primera, la identificación de los temas del documento; y segunda, la representación de ellos en un vocabulario controlado (Mai, 2001).

La indexación manual implica un proceso intelectual para usar un vocabulario controlado, y por tanto este sistema resulta complejo, lento y costoso. Además conlleva un elevado número de incongruencias tanto externas, cuando la tarea se lleva a cabo por múltiples indexadores, como internas, cuando un solo indexador realiza el trabajo en diferentes momentos (Olson & Wolfram, 2008; White, Willis, & Greenberg, 2014; Zunde & Dexter, 1969).

En cuanto a la indexación automática, ésta se puede abordar principalmente desde dos perspectivas. La primera es la extracción de palabras clave, basada en la presencia de la palabra clave en el texto y en el conjunto de la colección (Frank, Paynter, Witten, Gutwin, & Nevill-Manning, 1999; C.

Zhang, 2008). La segunda técnica es la asignación de palabras clave, basada en la concurrencia entre los términos del texto y del tesauro o algún otro vocabulario controlado (Moens, 2002; Yang, Zhang, Li, Yu, & Hao, 2014). Ambas perspectivas presentan inconvenientes. La primera técnica, la extracción de palabras clave, puede presentar resultados erróneos, en especial cuando se trata de palabras formadas por varios términos (es decir, cuando los sistemas utilizados deben identificar N-gramas). En el caso de la asignación, los principales problemas radican en la dificultad de disponer de lenguajes controlados que cubran la diversidad temática de los documentos, y en la necesidad constante de actualización del vocabulario.

2.2. La anotación semántica

La anotación semántica realizada con metadatos es la forma de dotar de contenido semántico a los documentos y de conseguir que los ordenadores puedan procesar e interpretar la información (Aguado de Cea, Álvarez de Mon Rego, & Pareja-Lora, 2002). El primer paso para llevar a cabo la anotación semántica es la 'extracción de información'. Éste es el término utilizado en el ámbito del procesamiento del lenguaje natural para referirse a la actividad de extraer automáticamente información específica de textos en lenguaje natural. Existen diferentes aproximaciones para realizar este proceso, las dos principales son: los sistemas de aprendizaje automático (*machine learning*) y los sistemas basados en reglas y patrones (Flynn, Zhou, Maly, Zeil, & Zubair, 2007).

Una vez extraída la información, el siguiente paso es la propia anotación. Las herramientas de anotación semántica permiten convertir en metadatos el contenido semántico extraído (Liao, Lezoche, Panetto, Boudjlida, & Loures, 2015; Uren et al., 2006). Existen diferentes aproximaciones para realizar la anotación semántica que pueden agruparse en tres categorías: la lingüística, la basada en ontologías, y la basada en vocabularios controlados.

2.3. Los vocabularios controlados

La ANSI/NISO Z39.19-2005, revisada en 2010, establece las directrices y convenciones para los vocabularios controlados. Los agrupa en cuatro niveles según su grado de complejidad: lista de descriptores, anillos de sinónimos, taxonomías y tesauros (NISO, 2010, p. 16). Los primeros son simplemente un conjunto limitado de descriptores que forman una lista alfabética, también denominada "lista de selección". Los anillos de sinónimos incluyen además descriptores equivalentes para un concepto. Las taxonomías van un paso más allá, contemplan una organización jerárquica de los conceptos (Milne, 2007). Y por último, los tesauros que incorporan también las relaciones semánticas entre los conceptos (López-Huertas, 1999).

El uso de los vocabularios controlados evita los problemas que conlleva el lenguaje natural en la recuperación de información. En concreto, soluciona el problema de la polisemia y la homonimia; es decir, cuando un término puede tener más de un significado. Y también el de la sinonimia, cuando un concepto puede designarse con diferentes palabras. En el primer caso se incrementa la precisión, pues se eluden los problemas de ambigüedad y se recuperan sólo los documentos pertinentes. En el segundo, caso se incrementa la cobertura, pues se incluyen los contenidos que se designan con términos alternativos.

Además, los vocabularios controlados pueden ser utilizados en diferentes fases del proceso de búsqueda de información y por ello pueden ser compatibles con la búsqueda por palabras clave. Por ejemplo, permiten acotar los resultados por temas o expandir la consulta, e incluso facilitar sistemas de recomendación (Murphy et al., 2003). Actualmente, estas funcionalidades están presentes en entornos cerrados de búsqueda, donde se requiere un elevado grado de efectividad, como pueden ser las bases de datos especializadas (Kharazmi, Karimi, Scholer, & Clark, 2014; McKenzie, 2001) o los repositorios (Haniewicz, 2012; White, 2013).

Asimismo, hay que dedicar un apartado a la revisión de las diferentes aproximaciones que se utilizan para obtener los descriptores de un vocabulario controlado; es decir, las unidades léxicas que se utilizan para designar conceptos en un dominio restringido temáticamente. Los métodos de extracción de terminología se clasifican en tres grupos: lingüístico, estadístico e híbrido (Estopà, 1999; Pazienza, Pennacchiotti, & Zanzotto, 2005; Chunxia Zhang, Niu, Jiang, & Fu, 2012).

3. Metodología

La investigación desarrollada en esta tesis es principalmente exploratoria, donde el tema central es estudiado desde un enfoque multidisciplinar. Además, se ha utilizado el método empírico para abordar la investigación. En primer lugar, se ha definido el objeto de estudio y el objetivo a alcanzar; después, se han establecido las hipótesis y las preguntas de investigación; y por último, se ha comprobado la validez de la propuesta con los resultados obtenidos de los experimentos.

El enfoque metodológico principal es el cuantitativo, ya que es el sistema utilizado para evaluar la viabilidad de los sistemas propuestos. En algún caso concreto se recurre al método cualitativo para valorar la calidad de los resultados obtenidos.

El corpus utilizado está formado por una selección aleatoria 100 artículos científicos en formato HTML de la revista *BiD: textos universitaris de biblioteconomia i documentació*, indexada en el portal de revistas académicas *Temaria*.

Las herramientas utilizadas durante la tesis se clasifican en dos grupos: i) programas creados ad hoc para realizar las

comparaciones, y ii) software libre que ha permitido procesar los corpus. Se han programado una serie de rutinas en Python para procesar los documentos y ejecutar las comparaciones. Estas rutinas procesan los archivos, identifican las palabras coincidentes y presentan los resultados. Para procesar los logs de consultas se utilizan dos herramientas de software libre: Natural Language Toolkit (NLTK) y Ngram Statistics Package (NSP).

A efectos de la evaluación de los resultados de la indexación, se considera que la indexación realizada por los expertos humanos es el modelo de referencia ya que identifica las palabras clave que mejor describen los documentos. Asimismo, para realizar la evaluación de los procesos de indexación automática se utilizan las medidas de exhaustividad, precisión y valor-F, que son las empleadas habitualmente en el ámbito (Medelyan & Witten, 2005; Verberne, D'hondt, van den Bosch, & Marx, 2014):

$$Precisión = \frac{\#\ PC\ asignadas\ correctamente}{\#\ PC\ asignadas}$$

$$Exhaustividad = \frac{\#\ PC\ asignadas\ correctamente}{\#\ PC\ asignadas\ manualmente}$$

$$Valor - F = 2\ \frac{Precisión * Exhaustividad}{Precisión + Exhaustividad}$$

Para la propuesta del modelo de actualización de los vocabularios controlados (explicado en 4.4) se procesan los logs de las consultas recogidos con *Google Analytics*. De esta forma se identifican qué consultas (palabras clave) se han utilizado en los buscadores para acceder al corpus de documentos detallado anteriormente. Es decir, se procesan todas las consultas formuladas por los usuarios desde los buscadores para acceder a los 100 artículos de la revista *BiD: textos universitaris de biblioteconomia i documentació* que forman el corpus estudiado.

4. Resultados

Los principales resultados de las investigaciones realizadas en esta tesis se agrupan en: i) descripción de las herramientas *Keyword Research*; ii) presentación de la herramienta DigiDoc MetaEdit desarrollada para experimentar con el proceso de indexación; iii) proceso de evaluación al que se somete la herramienta para valorar el proceso de indexación semiautomático; y ultimo, iv) modelo para actualizar los vocabularios controlados.

4.1. Características de las herramientas Keyword Research

En el artículo "Keyword Research: métodos y herramientas para identificar palabras clave" (Artículo 2) se ha estudiado cómo funcionan este tipo de herramientas y también se han analizado las más destacadas. Se trata de herramientas muy

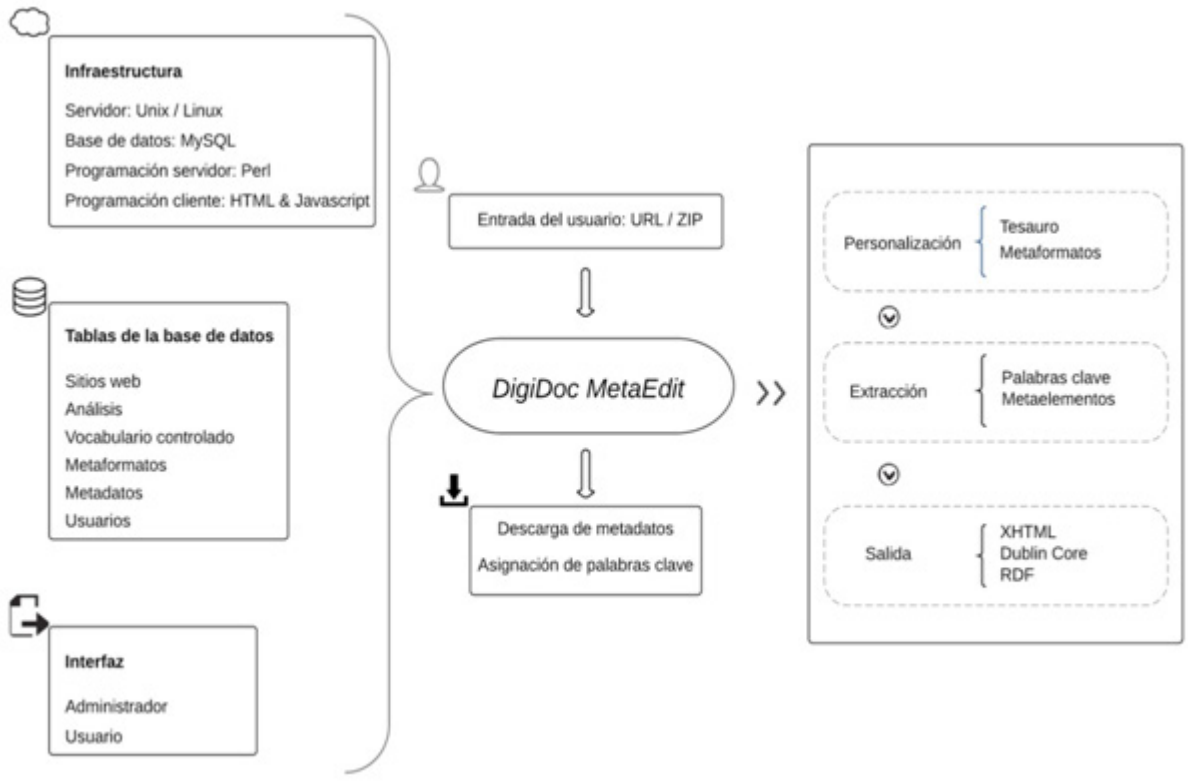

Figura 2: Estructura de la herramienta DigiDoc MetaEdit.

extendidas en el ámbito del SEO y del SEM que ayudan a identificar palabras clave con las que posicionar un sitio web. Así, estas herramientas muestran que los usuarios acostumbran a utilizar combinaciones de tres o más palabras cuando realizan una búsqueda. Esto implica que se emplean formas diferentes y precisas para acceder a los contenidos web. Por lo tanto, la identificación de palabras clave alternativas resulta muy útil, y de aquí proviene el éxito de estas herramientas. El estudio y análisis de estas herramientas ha permitido obtener información muy valiosa para definir las prestaciones a incorporar en la aplicación DigiDoc MetaEdit.

4.2. Herramienta DigiDoc MetaEdit

En los artículos "Procedimientos para la extracción de palabras clave de páginas web basados en criterios de posicionamiento en buscadores" (Artículo 1) y "A semiautomatic indexing system based on embedded information in HTML documents" (Artículo 3) se presentan las principales características de la herramienta DigiDoc MetaEdit. Esta herramienta desarrollada por el grupo de investigación DigiDoc, es el sistema utilizado para experimentar con los procesos de indexación semiautomáticos.

DigiDoc MetaEdit es un editor de metadatos (Pedraza-Jiménez, Codina, & Rovira, 2008; Vàllez, Rovira, Codina, & Pedraza-Jimenez, 2010) que permite la descripción de documentos HTML con alto contenido informacional. La herramienta fue creada con la misión de ayudar a la asignación de metadatos (anotación semántica), aunque la funcionalidad que se ha experimentado en esta tesis es la de identificar las palabras clave más significativas de un documento web para favorecer su indexación. El metaeditor permite definir qué criterios aplicar para asignar a un documento palabras clave de un vocabulario controlado. Se caracteriza por su gran versatilidad para adaptarse a cada contexto. Una vez las palabras clave de un documento web han sido identificadas, la herramienta genera un informe con todas las palabras clave halladas y también un archivo RDF (u otros formatos) con los metadatos.

En cuanto a las características técnicas, DigiDoc MetaEdit se ha desarrollado como una aplicación de software libre con licencia GLP. Se ha diseñado en Perl usando MySQL para el almacenamiento de datos. Su estructura es modular, lo que hace que sea más fácil agregar nuevas funcionalidades. Tiene tres módulos principales:

1. Módulo de personalización: su objetivo es permitir la personalización de la herramienta en base al vocabulario controlado, al formato de los metadatos, a los valores de las variables, etc.

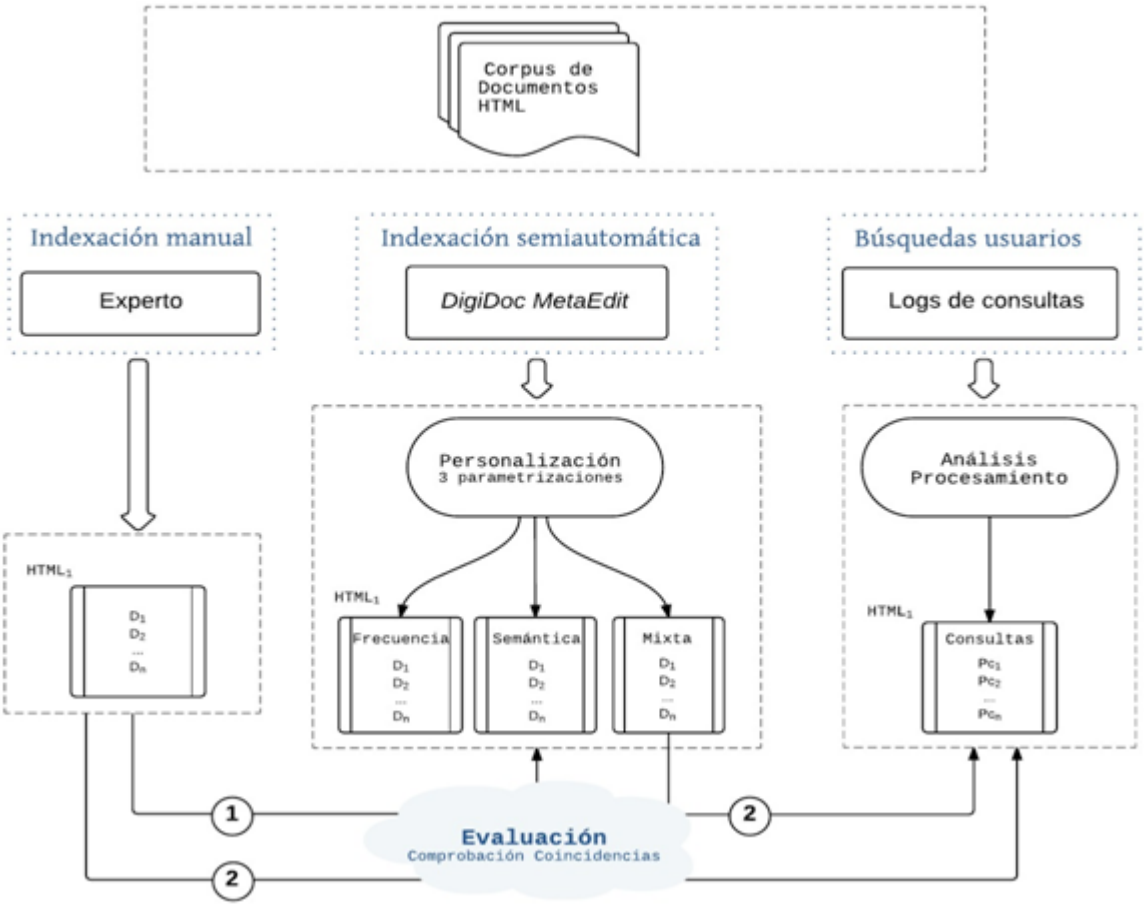

Figura 3: Proceso de evaluación de la herramienta DigiDoc MetaEdit.

2. Módulo de extracción: su objetivo es extraer las palabras clave y metaelementos de los documentos HTML.

3. Módulo de salida: su objetivo es presentar los metadatos extraídos y generar fragmentos de código con los metadatos según diferentes estándares, RDF, Dublin Core o XHTML.

La Figura 2 muestra un resumen de la estructura del DigiDoc MetaEdit.

4.3. Sistema de evaluación de la indexación semiautomática

La evaluación de los resultados de la indexación semiautomática se presentan en ls artículos "A semi-automatic indexing system based on embedded information in HTML documents" (Artículo 3) y "Updating controlled vocabularies by analysing query logs" (Artículo 4). El proceso de evaluación se divide en dos fases: primero, se compara la indexación manual frente a la indexación semiautomática obtenida con la herramienta; y segundo, se comparan los términos utilizados en el proceso de indexación manual y semiautomático frente a las palabras clave utilizadas por los usuarios en los buscadores. La Figura 3 muestra de forma gráfica los elementos que forman parte del proceso de evaluación al que se somete la herramienta, donde se han identificado las dos fases.

El corpus de documentos está indexado tanto por expertos humanos como por la herramienta DigiDoc MetaEdit que utiliza tres parametrizaciones diferentes. Además, se cuenta con las palabras clave utilizadas por los usuarios para acceder a los documentos. De este modo, el proceso de evaluación se ha realizado en dos etapas. Primero se ha comprobado el grado de solapamiento entre las indexaciones manual y semiautomática (fase 1). Después se han examinado las coincidencias entre los descriptores utilizados en el proceso de indexación, tanto manual como semiautomático, y las palabras clave utilizadas por los usuarios en los buscadores (fase 2).

4.4. Modelo para actualizar los vocabularios controlados

El artículo "Updating controlled vocabularies by analysing query logs" (Artículo 4) se presenta el modelo para obtener términos candidatos a incorporarse al vocabulario controlado. Los N-gramas obtenidos después del procesamiento tienen asociada diferente información: frecuencia del término en el corpus de consultas, número de palabras que lo forman, número de documentos a los que se accede con él, visitas generadas, y duración media de las visitas en segundos.

Esta lista de palabras clave obtenida no pasa diferentes de filtrado. Primero se excluyen los unigramas ya que la mayoría son demasiado genéricos. Después de descartan los términos que se utilizan para acceder a un único documento ya que se consideran poco representativos. Por último, se aplica la siguiente fórmula para asignar un grado de notoriedad de los términos candidatos a partir de la información asociada a cada término; es decir, identificar los términos son más relevantes y por tanto mejores candidatos.

$$\frac{Frec.\ Término}{Documentos}\ X\ Visitas\ X\ Duración\ visitas\ X\ Long.\ n\_grama$$

Los elementos de la fórmula se describen a continuación. La primera parte de la fórmula:

$$\frac{Frec.\ Término}{Documentos}$$

está formada por el número de veces que la palabra clave aparece en el corpus de consultas (frecuencia del término) y por el número de documentos en los que aparece. Este cálculo permite identificar las palabras clave más relevantes para el corpus de consultas de los usuarios. De este modo, una alta frecuencia de la palabra clave en las consultas denota su importancia, aunque ésta es compensada al dividirse por el número de documentos a los que permite acceder la pala-

N-grama	Frec.	Long.	Doc. accedidos	Visitas	Duración media visitas
evaluación por competencias	133	3	2	477	633
redes sociales	86	2	5	972	586
identidad digital	30	2	2	159	642
web 2.0	50	2	7	231	590
sistemas de evaluación	19	3	2	42	852
evaluación de competencias	21	3	2	64	488
sistemas de gestión	33	3	6	78	752
instrumentos de evaluación	24	3	5	67	673
repositorios institucionales	32	2	9	109	635
educación superior	24	2	4	68	530

Tabla 1. Lista final de candidatas a incorporarse al vocabulario controlado.

bra clave. Por tanto, se resta importancia a aquellas palabras clave más comunes y que por ello tienen menor poder de discriminación. Se trata de una adaptación de la medida del Tf-idf al corpus de documentos utilizado, las consultas de los usuarios.

La segunda parte de la fórmula:

Visitas X Duración visitas

hace referencia al número de visitas a un documento generadas por una palabra clave, y la duración media de las visitas. Su objetivo es determinar la relevancia de las palabras clave introducidas por el usuario de acuerdo a las 'necesidades de información'. Un alto número de visitas indica que una palabra clave es muy relevante. Del mismo modo, se asume que las visitas de más duración reflejan un mayor interés de los usuarios en el documento visitado. Por tanto, la fórmula multiplica el número de visitas generadas por una palabra clave por la duración media de cada visita.

La última parte de la fórmula:

Long. n_grama

toma en consideración la relevancia semántica de las palabras clave utilizadas por los usuarios. A tal fin se contempla la longitud de las palabras clave como una variable importante. En el ámbito de los vocabularios controlados es requisito básico la desambiguación de los descriptores, y esto sólo es posible con palabras clave precisas que acostumbran a ser compuestas. Por ello en la fórmula se considera un factor de relevancia positivo que las palabras clave identificadas estén formadas por varios elementos.

Tras aplicar la fórmula de notoriedad, se obtiene una lista final de palabras clave candidatas a incorporarse al vocabulario controlado. La Tabla 1 muestra los diez primeros términos candidatos obtenidos con su aplicación. Al revisar esta lista de palabras se observan términos candidatos a descriptores que hacen referencia a conceptos relativamente nuevos: 'redes sociales', 'identidad digital' o 'web 2.0'. Por tanto, el modelo permite identificar la nueva terminología del ámbito temático. Por otro lado, también se localizan diferentes variantes de un mismo concepto ('evaluación por competencias' o 'evaluación de competencias') y se ofrece al responsable del vocabulario controlado información complementaria para decidir qué descriptor incorporar.

5. Conclusiones

La necesidad de acceder a información relevante y pertinente es una de las situaciones más habituales tanto desde un punto de vista profesional como de ocio. Por tanto, cada vez es más apremiante encontrar mecanismos para facilitar la recuperación de información y así cubrir mejor las necesidades de información de los usuarios. La investigación presentada en la tesis intenta contribuir en la mejora de este panorama.

A modo de resumen se pueden citar tres aportaciones principales de la tesis:

1. El uso del marcado semántico, a través de las etiquetas HTML, y de los vocabularios controlados resulta una combinación ventajosa para la indexación de contenidos web.

2. La indexación semiautomática puede utilizarse como una herramienta complementaria para el indexador humano.

3. El uso de la información obtenida de la analítica web debe tenerse en cuenta para mejorar los procesos de recuperación de información.

Estas propuestas se caracterizan por su sencilla implementación y fácil adaptación al entorno de aplicación, además de su eficiencia y eficacia.

Bibliografía y referencias

Aguado de Cea, G., Álvarez de Mon Rego, I., & Pareja-Lora, A. (2002). Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de Web Semántica: OntoTag. *Revista Iberoamericana de Inteligencia Artificial*, 17, 37–49.

Berners-Lee, T., Hendler, J., Lassila, O., & others. (2001). The semantic web. *Scientific American*, 284(5), 34–43.

Borko, H. (1977). Toward a theory of indexing. I*nformation Processing & Management*, 13(6), 355-365.

Ding, L., Finin, T., Joshi, A., Peng, Y., Pan, R., & Reddivari, P. (2005). Search on the semantic web. *Computer*, 38(10), 62-69.

Estopà, R. (1999, julio 26). *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, Barcelona, Spain. http://www.tdx.cat/bitstream/handle/10803/7489/treb1de2.pdf?sequence=1

Flynn, P., Zhou, L., Maly, K., Zeil, S., & Zubair, M. (2007). Automated template-based metadata extraction architecture. En D. Goh; T. Cao; I. Sølvberg; & E. Rasmussen (Eds.), *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*. Berlin, Germany: Springer, 327–336.

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-specific keyphrase extraction. *Proceedings of the 16th International Joint Conference on Artificial Intelligence.* San Francisco, CA, USA: Morgan Kaufmann Publishers, 668-673.

Haniewicz, K. (2012). Local controlled vocabulary for modern web service description. En L. Rutkowski; M. Korytkowski; R. Scherer; R. Tadeusiewicz; L. A. Zadeh; & J. M. Zurada (Eds.), *Artificial Intelligence and Soft Computing*. Berlin: Springer, 639-646.

Hjørland, B. (2011). The importance of theories of knowledge: Indexing and information retrieval as an example. *Journal of the American Society for Information Science and Technology*, 62(1), 72–77.

Hou, S., Gu, J., & Zhou, Z. (2015). A novel concept index in semantic Web search. *Journal of Computational Information Systems*, 11(9),

3347-3356.

IDC. (2014). *The digital universe of opportunities: rich data and the increasing value of the internet of things*. Massachusetts, USA: IDC Analyze the Future. http://www.emc.com/leadership/digital-universe/2014iview/index.htm

Kharazmi, S., Karimi, S., Scholer, F., & Clark, A. (2014). A study of querying behaviour of expert and non-expert users of biomedical search systems. *Proceedings of the 19th Australasian Document Computing Symposium*. New York, USA: ACM.

Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. J*ournal of Web Semantics*, 2(1), 49-79.

Liao, Y., Lezoche, M., Panetto, H., Boudjlida, N., & Loures, E. R. (2015). Semantic annotation for knowledge explicitation in a product lifecycle management context: A survey. *Computers in Industry*, 71, 24–34.

López-Huertas, M. (1999). Potencialidad evolutiva del tesauro: hacia una base de conocimineto experto. *La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de la información*. Granada, Spain: Universidad de Granada, 133–140.

Mai, J. E. (2001). Semiotics and indexing: An analysis of the subject indexing process. *Journal of Documentation*, 57(5), 591.

McKenzie, E. M. (2001). Natural language searching: How win works in Westlaw. *Legal Reference Services Quarterly*, 18(4), 39-47.

Medelyan, O., & Witten, I. H. (2005). Thesaurus-based index term extraction for agricultural documents. P*roceedings of 6th Agricultural Ontology Service* (AOS). Vila Real, Portugal: Food and Agriculture Organization of the United Nations, 1122-1129.

Milne, C. (2007). Taxonomy development: assessing the merits of contextual classification. *Records management journal*, 17(1), 7–16.

Moens, M.-F. (2002). Automatic indexing: The assignment of controlled language index terms. En C. Zhai; & M. de Rijke (Eds.), *Automatic indexing and abstracting of document texts*. New York, USA: Springer, 103-132.

Murphy, L. S., Reinsch, S., Najm, W. I., Dickerson, V. M., Seffinger, M. A., Adams, A., & Mishra, S. I. (2003). Searching biomedical databases on complementary medicine: the use of controlled vocabulary among authors, indexers and investigators. *BMC Complementary and Alternative Medicine*, 3(1), 3.

NISO. (2010). *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. ANSI/NISO Z39.19-2005 (R2010). Baltimore, Maryland, USA: National Information Standards Organization. http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf

Olson, H. A., & Wolfram, D. (2008). Syntagmatic relationships and indexing consistency on a larger scale. *Journal of Documentation*, 64(4), 602-615.

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. En S. Sirmakessis (Ed.), *Knowledge Mining.* Berlin: Springer, 255–279.

Pedraza-Jiménez, R., Codina, L., & Rovira, C. (2008). Semantic web adoption: online tools for web evaluation and metadata extraction. En D. Ruan; & J. Montero (Eds.), *Computational Intelligence in Decision and Control* Madrid, Spain: World Scientific Publishing Company, 121-126.

Spärck Jones, K. (1974). Automatic indexing. J*ournal of Documentation,* 30(4), 393-432.

Stevens, M. E. (1970). *Automatic indexing: a state-of-the-art report*. Washington, DC, USA: National Bureau of Standards. http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED041610

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., & Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. J*ournal Web Semantics*, 4(1), 14-28.

Vàllez, M. (2009). La Web semántica y las tecnologías del lenguaje humano. En L. Codina; M.-C. Marcos; & Rafael Pedraza-Jiménez (Eds.), *Web semántica y sistemas de información documental* Gijón, España: Trea, 155-180.

Vàllez, M., & Pedraza-Jiménez, R. (2007). El procesamiento del lenguaje natural en la recuperación de información textual y áreas afines. *Hipertext.net, 5*. http://www.upf.edu/hipertextnet/numero-5/pln.html

Vàllez, M., Rovira, C., Codina, L., & Pedraza-Jimenez, R. (2010). Procedimientos para la extracción de palabras clave de páginas web basados en criterios de posicionamiento en buscadores. *Hipertext.net*, (8). http://www.upf.edu/hipertextnet/numero-8/extraccion_keywords.html

Verberne, S., D'hondt, E., van den Bosch, A., & Marx, M. (2014). Automatic thematic classification of election manifestos. *Information Processing & Management*, 50(4), 554-567.

White, H. (2013). Examining scientific vocabulary: mapping controlled vocabularies with free text keywords. *Cataloging & Classification Quarterly*, 51(6), 655-674.

White, H., Willis, C., & Greenberg, J. (2014). HIVEing: The effect of a semantic Web technology on inter-indexer consistency. *Journal of Documentation*, 70(3), 307-329.

Yang, S., Zhang, B., Li, S., Yu, C., & Hao, Q. (2014). Keyword extraction using multiple novel features. *Journal of Computational Information Systems*, 10(7), 2795-2802.

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 4(3), 1169-1180.

Zhang, C., Niu, Z., Jiang, P., & Fu, H. (2012). Domain-specific term extraction from free texts. 2*012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 1290-1293.

Zunde, P., & Dexter, M. E. (1969). Indexing consistency and quality. *American Documentation*, 20(3), 259–267.

CV

Mari Vállez es Doctora en Comunicación Social, Licenciada en Filología Hispánica y en Documentación, y además tiene un Máster en Lingüística Computacional y el DEA en Informática Aplicada. Trabaja en la Universitat Oberta de Catalunya como bibliotecaria y formadora, y desde el 2006 es profesora de la Universitat Pompeu Fabra en el Departamento de Comunicación. Su investigación se centra en la identificación de sistemas que faciliten el acceso o a la información desde diferentes perspectivas: sistemas de indexación, extracción de información, SEO académico, etc.