RECUPERACIÓN DE OBJETOS DE APRENDIZAJE EN REPOSITORIOS: UNA APLICACIÓN CON BÚSQUEDA SEMÁNTICA

RETRIEVAL OF LEARNING OBJECTS IN REPOSITORIES: AN APPLICATION WITH SEMANTIC SEARCH

AUTOR
GERMÁN A. OSORIO ZULUAGA
Ph.D (c) en Ingeniería
*Universidad Nacional del Colombia – Sede Manizales
Profesor Asociado
Departamento de Informática y Computación
gaosorioz@unal.edu.co
COLOMBIA

AUTOR
NÉSTOR DARÍO DUQUE MÉNDEZ
Ph.D en Ingeniería
*Universidad Nacional del Colombia – Sede Manizales
Profesor Asociado
Departamento de Informática y Computación
ndduqueme@unal.edu.co
COLOMBIA

***INSTITUCIÓN**
Universidad Nacional de Colombia
– Sede Manizales
UNAL
Campus La Nubia - Manizales
COLOMBIA

INFORMACIÓN DE LA INVESTIGACIÓN O DEL PROYECTO: El trabajo presentado en este artículo fue costeado parcialmente por el proyecto de investigación financiado por COLCIENCIAS titulado "RAIM: Implementación de un framework apoyado en tecnologías móviles y de realidad aumentada para entornos educativos ubicuos, adaptativos, accesibles e interactivos para todos" con contrato 0205-2013.

RECEPCIÓN: Mayo 04 de 2015 **ACEPTACIÓN:** Septiembre 12 de 2015

TEMÁTICA: Ingeniería del software

TIPO DE ARTÍCULO: Artículo de Investigación Científica e Innovación

Forma de citar: Osoario Zuluaga, G. A. (2015).Recuperación de objetos de aprendizaje en repositorios: Una aplicación con búsqueda semántica. En R, Llamosa Villalba (Ed.). Revista Gerencia Tecnológica Informática, 14(40), 43-54. ISSN 1657-8236.

RESUMEN ANALÍTICO

En los últimos años ha crecido el número de recursos educativos almacenados en repositorios de objetos de aprendizaje. Para su recuperación se usan generalmente métodos de búsqueda tradicionales de coincidencia de términos de consulta con los metadatos de los objetos de aprendizaje. La precisión en los resultados de búsqueda con estos métodos sigue siendo baja. En este sentido, este trabajo se enfocó en mejorar el indicador de precisión en las búsquedas, usando la técnica de Análisis Semántico Latente sobre los metadatos que describen el objeto de aprendizaje. Esta técnica permite aproximaciones por su significado. En el experimento realizado, se evidenció una mejora en la precisión en la búsqueda, a medida que se ingresaron más términos en la consulta. La implementación se puede extender a búsquedas de texto completo de objetos de aprendizaje textuales, si se tiene acceso al contenido textual completo del objeto de aprendizaje.

PALABRAS CLAVES: Análisis Semántico Latente, LSA, objetos de aprendizaje, repositorios de objetos de aprendizaje, recuperación de información.

ANALYTICAL SUMMARY

In recent years, the number of educational resources stored in repositories of learning objects has increased. For recover them, generally traditional methods searching query terms that match the metadata of the learning objects are used. Precision in search results with these methods remains low. In this sense, this work focused on improving the precision indicator by applying Latent Semantic Analysis technique (LSA) over metadata that describe to learning object. This technique allows approximations for its meaning. In the experiment, it is shown an improvement in precision in the search, as more terms are entered in the query. The implementation could be extended to full-text searches of textual learning objects, if we have access to the full text of the learning object content.

KEYWORDS: Latent Semantic Analysis, LSA, learning objects, learning object repositories, information retrieval

INTRODUCCIÓN

En la era actual de las computadoras, hay varios recursos que recopilan información relacionada con un tema determinado, como las estaciones de radio, televisión, Internet, entre otros. Internet es considerado como factor importante para la obtención de cualquier información sobre un dominio dado (Jain & Singh, 2013). La recuperación de información de la Web, es obtenida a través de motores de búsqueda de propósito general.

En un ámbito más restringido de la Web, como lo es el de su aplicación en educación, se implementan soluciones para manejar sus recursos, comúnmente denominados objetos de aprendizaje (OA). Los repositorios de objetos de aprendizaje son una de esas soluciones que permiten la organización y gestión de estos recursos digitales, facilitando su búsqueda y recuperación.

Con el fin de recuperar información de los repositorios, cada objeto debe ser etiquetado con metadatos, los cuales contienen datos que describen el OA (Barak & Ziv, 2013). Con base en ellos, principalmente, se buscan y recuperan los OAs.

En general, las plataformas de e-learning utilizan tecnologías de recuperación de información tradicionales. Los usuarios suelen ingresar palabras clave en los motores de búsqueda, y como resultado reciben la lista de todas las páginas web que contienen la misma cadena de caracteres de las palabras clave elegidas. Estos motores de búsqueda se basan en términos de indexación sin tener en cuenta la semántica de los contenidos pedagógicos y el contexto (Smine et al., 2012). En ese mismo sentido, Ismail y Joy (2011) plantean que los sistemas de gestión de contenidos y repositorios, han restringido los medios para organizar y comprender las relaciones semánticas captadas entre los objetos de aprendizaje y otros documentos almacenados. En otras palabras, señala que los metadatos no contienen información que permita reconocer los patrones pedagógicos estructurales que tiene el repositorio en su conjunto.

Además y con respecto a la búsqueda y recuperación de OAs, se da la situación de que la relevancia de los objetos de aprendizaje recuperados con base a sus metadatos sigue siendo baja (Tabares Morales et al., 2011). De igual manera, se presenta un elevado porcentaje de errores en el acceso a los repositorios, primordialmente debido a las numerosas ocasiones en que los repositorios, o bien no responden a las consultas, o simplemente no funcionan (Gil et al., 2010).

Por otro lado, el número de recursos de aprendizaje disponibles en la Web se ha incrementado dramáticamente (Alharbi et al., 2012). Hay una disponibilidad crecientemente exponencial de OAs, que genera un aumento en la dificultad de encontrar el recurso adecuado para el usuario, basado en el contexto del aprendizaje o de sus preferencias (Plaza Morales, 2011; Sabitha & Mehrotra, 2013). Esta situación enfrenta a los docentes e investigadores en un espacio educativo, y en general a los usuarios, a realizar un gran esfuerzo para componer material educativo basado en OAs. En este sentido, no es difícil encontrar recursos digitales en diferentes áreas. De hecho, una tarea más difícil es encontrar los recursos adecuados asociados a dominios específicos, entre la enorme cantidad de recursos disponibles (Becerra et al., 2012). Sumado a lo anterior, las herramientas utilizadas para buscar y encontrar objetos de aprendizaje en diferentes sistemas no proporcionan una manera significativa y escalable para comparar, clasificar y recomendar el material de aprendizaje (Gil et al., 2010).

En este trabajo se propone un sistema de recuperación de información en repositorios, basado en Análisis Semántico Latente (LSA), que pretende mejorar la precisión de los resultados de búsqueda y recuperación de información y que tiene en cuenta las relaciones semánticas entre todos los OAs del repositorio, es decir, las similitudes en significado entre los distintos objetos. El algoritmo propuesto para generar la matriz índice de las consultas actúa sobre los metadatos de título, descripción y palabras clave.

1. MARCO CONCEPTUAL

Esta sección presenta los conceptos básicos de objetos de aprendizaje y su soporte. Además de explicar conceptualmente LSA.

1.1 OBJETOS DE APRENDIZAJE Y REPOSITORIOS

1.1.1 Objetos de Aprendizaje

El término fue acuñado en 1992 por Wayne Hodgins, experto futurista y estratega en e-learning de Autodesk Inc (Saum, 2007). Un objeto de aprendizaje se define como cualquier entidad, digital o no digital, que puede ser utilizada para el aprendizaje, la educación o la formación (IEEE, 2002), soportada en la tecnología (Moore, 2013). Se orientan a la instrucción, aprendizaje o enseñanza basada en computador. No son realmente una tecnología; más propiamente dicho es una filosofía, que se fundamenta en la corriente de las ciencias de la computación conocida como orientación a objetos (López Guzmán, 2005).

1.1.2 Metadatos

García Aretio, citado por Astudillo (2011), los define como "una estructura detallada del texto, que describe atributos, propiedades y características distribuidos en diferentes campos que identifican claramente al objeto, con el fin de que pueda encontrarse, ensamblarse y utilizarse". Son especialmente útiles en los recursos que no son textuales y en los que su contenido no puede ser indizado por sistemas automáticos, por ejemplo, los multimedios (López Guzmán, 2005).

Los metadatos tienen las siguientes características (Sánchez-Alonso et al., 2007):

- "Dicen algo" sobre el objeto, en sentido general.
- Físicamente son externos al propio recurso: están contenidos en otro archivo o se obtienen de algún servicio diferente.
- Utilizan un formato técnico para su expresión y para su intercambio, generalmente lenguajes definidos sobre XML.
- Utilizan una serie de descriptores, campos o elementos normalizados para conseguir un cierto grado de interoperabilidad entre diferentes sistemas.

Entre los estándares de metadatos más comunmente usados tenemos Dublin Core (1995), LOM (2002), ADL SCORM (2004). También existen adaptaciones de estos estándares denominados *perfiles de aplicación*, tales como CanCore (Canadá), UK LOM Core (Reino Unido), Vetadata (Australia) (Astudillo, 2011) y OBAA (Brasil) (Rodríguez M et al., 2012).

El estándar LOM está compuesto de 45 elementos agrupados en 9 categorías (IEEE, 2002), el cual fue adaptado para la versión colombiana MEN-LOM (Ministerio de Educación Nacional, 2012).

1.1.3 Bibliotecas digitales y Repositorios de Objetos de Aprendizaje ROA

Los repositorios de objetos de aprendizaje son sistemas de información complejos que involucran a conjuntos muy amplios de objetos digitales y sus respectivos

metadatos, además de múltiples estructuras y servicios (por ejemplo, búsqueda, navegación y personalización), y se construyen normalmente por una comunidad de usuarios con intereses específicos (Laender et al., 2008). Conforman la infraestructura clave para el desarrollo, almacenamiento, administración, localización y recuperación de todo tipo de contenido digital (López Guzmán, 2005).

Existen repositorios locales que contienen los OAs propios y los repositorios remotos que son aquellos a los que se accede a través de una red (Rodríguez M et al., 2012). Los repositorios pueden almacenar solo los metadatos o de manera conjunta, también los recursos educativos asociados (Gil et al., 2010).

1.1.4 Federación de Repositorios de OA

Con el fin de centralizar las búsquedas sobre sitios distribuidos, los repositorios de objetos de aprendizaje digitales, se unen en federaciones de repositorios (Rodríguez M et al., 2012).

Es un entorno que consta de instancias paralelas de repositorios existentes unidos en una federación que se comportan como si se tratara de un único repositorio. El deseo de federar repositorios surge en realidad también como resultado de la comprensión de que ninguna biblioteca digital única alberga todos los artefactos que son relevantes para un dominio específico, una comunidad o una aplicación (Van de Sompel et al., 2008).

1.1.5 Federación de Repositorios de Objetos de Aprendizaje Colombia - FROAC

FROAC fue construida en el marco del proyecto "ROAC Creación de un modelo para la Federación de OA en Colombia que permita su integración a confederaciones internacionales" con el fin de ofrecer un punto único de acceso a recursos educativos que se encuentran almacenados en repositorios de las diferentes sedes de la Universidad Nacional de Colombia y otras instituciones nacionales. Centraliza los metadatos de los recursos que se encuentran distribuidos en los ROAs afiliados de forma transparente para el usuario. Pretende convertirse en una plataforma experimental para la comunidad involucrada en su desarrollo y de los grupos que deseen vincularse (Tabares et al., 2014).

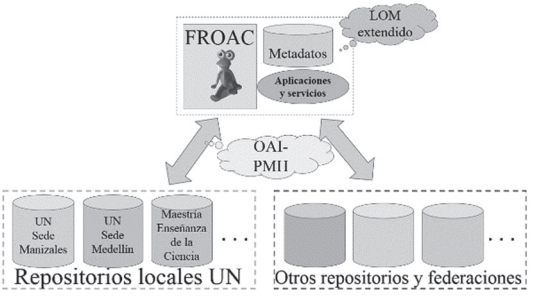
La Figura 1 muestra el esquema adoptado para FROAC. El modelo está compuesto por un grupo de repositorios locales de la Universidad Nacional de Colombia y construidos con la herramienta ROAp, además se prevé la integración con repositorios y federaciones externos. Los metadatos se centralizan en la base de datos de la federación. El estándar básico de metadatos utilizado es IEEE LOM y también tiene la opción de agregar nuevos metadatos para cubrir necesidades específicas relacionadas con el desarrollo de aplicaciones y servicios que podrán ser ofrecidos desde la federación. Los repositorios afiliados no necesariamente deben utilizar este mismo estándar, ya que se realiza un proceso de mapeo de equivalencias para hacer la conversión de otros estándares (Tabares et al., 2014).

1.2 ANÁLISIS SEMÁNTICO LATENTE

Análisis Semántico Latente es una teoría y un método para extraer y representar el significado contextual de las palabras por cálculos estadísticos aplicados a un amplio corpus de textos (Landauer et al., 1998).

Está basado en el concepto de Modelos de Espacio Vectorial (VSM), un enfoque que usa el álgebra lineal para recuperación de información efectiva de manera automática. Fue desarrollado para manejar la recuperación de texto desde grandes bases de datos en donde el texto es heterogéneo y varía el vocabulario. El modelo matemático formal subyacente del modelo de espacio vectorial define vectores únicos para cada término (palabra o concepto) y documento, y las consultas son llevadas a cabo comparando la representación de la consulta a la representación de cada documento en el espacio vectorial. Las similitudes consulta-documento están basadas en conceptos o contenidos semánticos similares (Martin & Berry, 2007).

Figura 1. Esquema de la Federación de Objetos de Aprendizaje



Fuente: Elaboración propia adaptada de Tabares et al. (2014)

La premisa del método es que la información acerca de contextos en los que una palabra en particular aparece o no aparece, proporciona un conjunto de restricciones que determina la similitud de los significados de las palabras entre sí. LSA ha demostrado que aborda los problemas de polisemia y sinonimia bastante bien, lo cual es importante en relación con el problema de localización de características, ya que los usuarios (en esta caso los desarrolladores de sistemas de software)

suelen construir consultas sin saber con precisión el vocabulario del sistema de destino. Es muy adecuado para hacer frente a este tipo de situaciones, ya que no utiliza una gramática o vocabulario predefinido. Los significados de las palabras se derivan de su uso, en lugar de partir de un diccionario o tesauro, que es una ventaja sobre un enfoque tradicional de lenguaje natural, donde deben desarrollarse un subconjunto de la gramática y un diccionario Inglés (Poshyvanyk et al., 2007). En este contexto, cuando se habla aquí de localización de características o conceptos, se hace referencia a la actividad de identificar una ubicación inicial en el código fuente que implementa funcionalidad dada en un sistema de software (Dit et al., 2013).

Basado en evidencias empíricas, se sabe que LSA produce medidas de relaciones palabra-palabra, palabra-pasajes de texto y pasajes de texto-pasajes de texto, que se correlacionan bien con varios fenómenos cognitivos humanos que involucran asociación o similitud semántica. Estas correlaciones demuestran gran parecido entre resultados de LSA y la forma de representación de significado que las personas reflejan de lo que han leído y oído. Como consecuencia práctica de esta correspondencia, LSA permite llevar a cabo juicios de similitud de significados entre palabras, muy aproximados a los elaborados por humanos y predecir objetivamente la similitud global basada en palabras, entre pasajes de texto (Landauer et al., 1998).

Su objetivo es identificar las dimensiones semánticas ocultas en datos de texto y luego usar el mapeo de las palabras originales de estas dimensiones semánticas para obtener una mejor medida de similitud entre documentos y / o consultas (Mihalcea & Radev, 2011). LSA es una evolución de una técnica de recuperación de información bibliográfica denominada Indexación Semántica Latente - LSI (Deerwester et al., 1990).

A continuación se describe matemáticamente.

1.2.1. Matriz de entrada

Para crear un modelo de espacio vectorial para LSA, se debe construir primero la matriz término-documento. Las filas están compuestas por los términos, los cuales son componentes individuales que conforman un documento. Por lo general, son términos pero también pueden ser frases o conceptos dependiendo de la aplicación. Las columnas de la matriz de entrada está compuesto de documentos, los cuales son de un tamaño predeterminado de texto, tales como parágrafos, colección de parágrafos, sentencias, capítulos de libro, libros y así sucesivamente, de nuevo dependiendo de la aplicación. Una colección de documentos compuesta de m términos y n documentos, puede ser representada como una matriz **A** $m \times n$ términos-documentos. Muy frecuentemente $m >> n$, el número de términos es mayor que el número de documentos, sin embargo, hay casos en donde es a la inversa y $n >> m$, por ejemplo, cuando la colección de documentos es de Internet (Martin & Berry, 2007).

Cada elemento A_{ij} no cero de **A**, corresponde a la frecuencia del *i-ésimo* término en el *j-ésimo* documento. Típicamente, la matriz **A** se considera dispersa porque contiene muchas más entradas cero que no cero. Cada documento en una colección tiende solo al uso de un subconjunto de términos del conjunto (Martin & Berry, 2007).

1.2.2 Descomposición de la matriz de entrada en componentes ortogonales

Una vez se crea la matriz de entrada **A**, se transforma en un espacio vectorial de término y documento por descomposiciones ortogonales con el fin de aprovechar el truncado de vectores. Transformar una matriz por el uso de descomposición ortogonal, o matrices ortogonales, preserva ciertas propiedades de la matriz, incluyendo las normas, o longitudes de vector o distancias, de los vectores de fila y columna que forman la matriz **A** $m \times n$ término-documento (Martin & Berry, 2007).

Hay varios métodos para descomponer la matriz **A** en componentes ortogonales. El método más usado para LSA es la descomposición en valores singulares (SVD), por varias razones. La primera, la SVD descompone **A** en componentes ortogonales que representan tanto a los términos como a los documentos. Segundo, la SVD captura suficientemente la estructura semántica subyacente de una colección y permite ajustar la representación de los términos y documentos en el espacio vectorial escogiendo el número de dimensiones. Finalmente, es manejable de la SVD es manejable para grandes conjuntos de datos (Martin & Berry, 2007).

La SVD de una matriz **A** $m \times n$ se define como sigue:

$$\mathbf{A} = \boldsymbol{U}\,\boldsymbol{\Sigma}\boldsymbol{V}^{T} \qquad (1)$$

en donde U es una matriz ortogonal, V es otra matriz ortogonal y Σ es una matriz diagonal con el resto de posiciones en ceros. Las primeras r columnas de la matriz ortogonal U contiene r vectores propios ortonormales asociados con los r valores propios no cero de $\mathbf{AA^T}$. Las primeras r columnas de la matriz ortogonal V contienen r vectores propios ortonormales asociados con los r valores propios no cero de $\mathbf{A^TA}$. Las primeras r entradas de la diagonal de Σ son las raíces cuadradas no negativas de los valores propios no cero de $\mathbf{AA^T}$ y $\mathbf{A^TA}$.

Las filas de la matriz *U* son los vectores de los términos y son llamados vectores singulares izquierdos. Las filas de *V* son los vectores de los documentos y llamados vectores singulares derechos (Martin & Berry, 2007).

1.2.3 Truncado de valores ortogonales

Dado el hecho que **A** puede ser escrita como la suma de matrices de rango 1: $\sum u_i \sigma_i v^T_i$, *r* puede ser reducido a *k* para crear $\mathbf{A_k} = \sum u_i \sigma_i v^T_i$. La matriz $\mathbf{A_k}$ es la aproximación al rango *k* mejor o más cercano (la distancia es minimizada) de la matriz original *A*. La matriz $\mathbf{A_k}$ ($A_k = U_k \Sigma V^T_k$) se crea ignorando o poniendo a cero todos los elementos, excepto los primeros *k* elementos o columnas del vector de términos en *U*, los primeros *k* valores propios en Σ, y los primeros *k* elementos o columnas del vector de documentos en *V*. Para reducir la dimensión de *r* a *k*, se remueve la información extraña y la variabilidad en el uso de términos, referidos como "ruido", el cual se asocia con la base de datos o colección de documentos. Truncando la SVD y creando $\mathbf{A_k}$ es que se captura la estructura semántica subyacente importante de los términos y documentos. Los términos similares en significado, están "cerca" unos de los otros en el espacio vectorial *k*-dimensional, aun si ellos nunca co-ocurren en un documento, y los documentos similares en significado conceptual están cerca los unos de los otros, aun si ellos no comparten términos en común. Este espacio vectorial *k*-dimensional es el fundamento que explota el LSA (Martin & Berry, 2007). En la Figura 2 se puede observar el diagrama de esta transformación.

La mejor selección de *k* permanece como una pregunta abierta. En la práctica, la escogencia de *k* depende de ensayos empíricos, los cuales han mostrado que, en grandes bases de datos, la escogencia óptima para el número de dimensiones está en el rango de 100 a 300 (Martin & Berry, 2007).

1.2.4 Búsqueda semántica

Cuando el usuario ingresa los términos de búsqueda en el buscador, podemos considerar que ha dado un vector que contiene ceros y frecuencias de término correspondientes a los términos especificados en la consulta definido por:

$$\boldsymbol{q}^T = (q_1, q_2, \ldots, q_m) \in \mathbb{R}^{1 \times m},$$

en donde

$$\boldsymbol{q}_I = \{1 \text{ si el termino } T_i \text{ aparece } 0 \text{ en otro caso} \quad (2)$$

Una vez se forma el seudodocumento y se proyecta en el espacio término-documento, se usa una medida de similitud para determinar cuáles términos y documentos están más cerca de la consulta. Se usa comúnmente la medida de similitud del coseno y el coseno del ángulo entre la consulta o seudodocumento y se computan para cada uno de los documentos o términos, como sigue (Gracia, 2002):

$$\cos \varphi_j = \frac{\mathbf{q}^T \mathbf{d}_j}{\|\mathbf{q}\| \|\mathbf{d}_j\|} \quad (3)$$

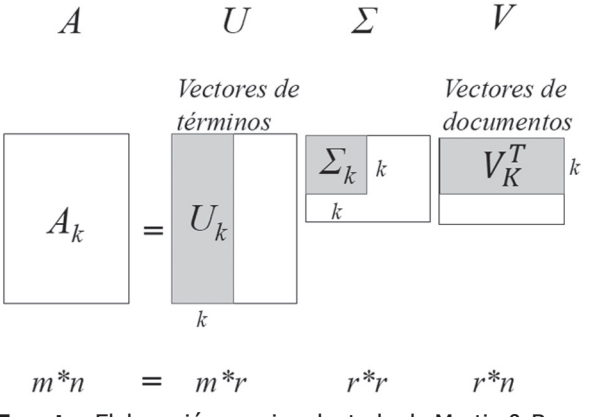
en donde $\mathbf{d}_1$, $\mathbf{d}_2$, ..., $\mathbf{d}_j$, ..., $\mathbf{d}_n$ corresponden a las columnas de la matriz $\mathbf{A_k}$.

Los cosenos asociados a cada documento, se ordenan en orden descendente; así, el documento con el coseno más alto en la consulta será el primero. Solo se listan los documentos que se encuentren por encima de cierto umbral (Martin & Berry, 2007).

1.3 PRECISIÓN

La precisión es uno de los indicadores más típicos en la comunidad académica de Recuperación de Información (Nanba & Okumura, 2005). La precisión hace referencia a la fracción de los documentos recuperados (conjunto *A*) que son relevantes en una consulta realizada por el usuario (Baeza-Yates & Ribeiro-Neto, 2011). En la Figura 3 se puede apreciar gráficamente este indicador. Matemáticamente, la precisión se expresa como (Baeza-Yates & Ribeiro-Neto, 2011):

Figura 2. Diagrama de la descomposición en valores singulares truncado



Fuente: Elaboración propia adaptada de Martin & Berry (2007)

$$\textit{Precisión} = \frac{|\boldsymbol{R} \cap \boldsymbol{A}|}{|\boldsymbol{A}|} \quad (4)$$

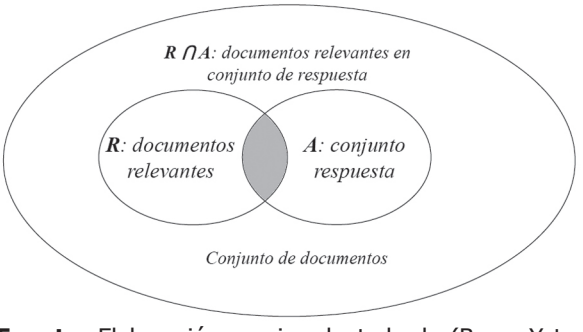
en donde *A* corresponde al conjunto de documentos recuperados y *R* es el conjunto de documentos relevantes.

2. ANTECEDENTES

El método de Análisis Semántico Latente es usado en una amplia gama de aplicaciones.

Una de ellas es la recuperación de imágenes basada en contenido (CBIR). En este sentido, el trabajo de Stathopoulos & Kalamboukis (2015) se muestra que LSA combina eficazmente información visual y textual (anotaciones a imágenes médicas). Comparado con otras técnicas de punta de CBIR, en un mismo conjunto de datos, su desempeño está a la misma altura de los de rendimiento superior.

Figura 3. Representación gráfica de la precisión



Fuente: Elaboración propia adaptada de (Baeza-Yates & Ribeiro-Neto, 2011)

Otra de las aplicaciones es en ingeniería de software, donde existe una comunidad muy activa usando esta técnica. Para reconstruir los enlaces de trazabilidad entre los artefactos de software, en el trabajo de Capobianco et al. (2013) se proponen mejoras en la indexación dejando solo sustantivos en el texto de dichos artefactos. En un amplio estudio empírico, De Lucia et al. (2013) analizan técnicas de recuperación de información para etiquetar código fuente, entre la que se encuentra LSA. Otros trabajos en esta línea son los de Ali et al. (2013), Borg et al. (2014), Wang et al. (2013) y Zamani et al. (2014).

Esta técnica es usada por Novelli & Parente De Oliveira (2012) para la extracción automática de ontologías en documentos y colecciones de documentos. Un trabajo similar es el de Atkinson et al. (2013), en donde LSA es usado para la extracción automática de metadatos de objetos de aprendizaje desde la Web con el fin de identificarlos para usos educativos. De igual manera, Brut et al. (2011) utilizan LSA conjugada con otras técnicas de Procesamiento de Lenguaje Natural para anotación semántica, de manera automática.

En el trabajo de Aswani Kumar et al. (2012), se analizan tres técnicas: Modelo de Espacio Vectorial (VSM), su variante LSA y el Análisis de Concepto Formal (FCA) en las tareas de recuperación de información sobre la base de datos de bibliografía médica y otras bases de datos de salud.

3. PROPUESTA, ALGORITMOS Y EXPERIMENTO

3.1 PROPUESTA

Se propuso realizar la búsqueda de objetos de aprendizaje de un repositorio, usando la técnica del Análisis Semántico Latente, mediante la implementación de los algoritmos correspondientes.

Para ello, se toman tres elementos de la categoría General del estándar LOM: título, descripción y palabras clave, ya que son los metadatos que mejor describen el OA para efectos de recuperación.

3.2 ALGORITMOS

Se implementaron dos algoritmos en Phyton 2.7: uno, que genera la matriz transformada término-documento $\mathbf{A}_{k'}$ mediante la cual se puede cruzar con el vector generado por una búsqueda del usuario (Algoritmo 1); el otro, es el algoritmo que lee el texto de la consulta del usuario y lo cruza con la matriz generada con el primer algoritmo, generando los resultados de mayor similitud semántica (Algoritmo 2). En la Tabla 1 se listan las notaciones usadas en ellos.

En el Algoritmo 1 se parte de los metadatos asociados al título, descripción y palabras clave de cada objeto de aprendizaje del repositorio. En el paso 8 se usó la biblioteca *nltk* de Python (Bird et al., 2009) y para el paso 14, para la implementación del método LSA se usó la biblioteca *numpy* (Oliphant, 2006).

En la determinación del valor de *k* (dentro del paso 14) y por analogía de LSA con el método de Análisis de Componentes Principales (para reducción de dimensión de datos), en el que se calculan vectores propios y autovalores para trabajar en un subespacio de la matriz original (Díaz M, 2007), se usa uno de los criterios para determinar el número de componentes, que consiste en seleccionar los primeros autovalores ordenados de mayor a menor hasta que cubran un valor de varianza acumulada del orden de 80% o el 90% (Peña, 2002). El número de autovalores necesarios para llegar a ese umbral corresponde, en nuestro caso, al valor de *k.*

TABLA 1. Lista de notaciones usadas en los algoritmos

Notación	Significado
m	Número de términos del ROA
n	Número de objetos de aprendizaje del ROA
$M \in \mathbb{R}^{m \times 1}$	Conjunto de términos del ROA
$N_j, \ \forall\ j = 1...n$	Conjunto de términos del OA_j
tit, keyW, desc	Texto de los metadatos de título, palabras clave y descripción de cada OA
$\mathbf{A} \in \mathbb{R}^{m \times n}$	Matriz de frecuencias de términos por OA
k	Rango de truncamiento
$\mathbf{A_k} \in \mathbb{R}^{m \times n}$	Matriz de frecuencias de términos por OA truncada (LSA)
consulta	Texto de búsqueda del usuario
$\mathbf{q} \in \mathbb{R}^{1 \times m}$	Vector de frecuencias de términos de la consulta
$\mathbf{d_j} \in \mathbb{R}^{m \times 1}$	Columna j-ésima del $\mathbf{A_k}$
$\mathbf{csn} \in \mathbb{R}^{1 \times n}$	Vector de cosenos (similitud)

En este algoritmo no se incluyeron métodos de lematización, para reducir una palabra a su raíz (Bird et al., 2009), ya que se pretende definir una línea base de comparación para trabajos futuros.

En lo que tiene que ver con la generación la matriz transformada $\mathbf{A_k}$, que es la matriz índice, solo se debe ejecutar una vez, partiendo de todos los objetos de aprendizaje del repositorio. Debe ejecutarse nuevamente cuando se ingrese un nuevo objeto de aprendizaje al repositorio o se actualicen metadatos en los existentes.

En el Algoritmo 2 se parte de la matriz índice $\mathbf{A_k}$ generada con el Algoritmo 1 y del valor umbral del coseno a usar (calculado de manera empírica). El usuario debe ingresar el texto de la consulta, el cual es descompuesto en sus términos y luego se excluyen las denominadas stopwords o palabras no significativas (artículos, preposiciones, etc.). Partiendo de estos términos, se genera el vector de frecuencias de la consulta **q**. Con él se calcula el producto punto con todas y cada una de las columnas de $\mathbf{A_k}$, produciendo el vector de **c**. Cada elemento de **c** se divide entre la norma del vector de la consulta y la norma correspondiente a la columna de cada documento, generando el vector de cosenos **csn**. Este vector se ordena descendentemente y se listan sus elementos cuando están por encima del umbral definido, siendo éstos los resultados de la búsqueda en el repositorio para una consulta dada.

3.3. EXPERIMENTO

Los algoritmos fueron probados en la Federación de Repositorios de Objetos de Aprendizaje Colombia - FROAC (Universidad Nacional de Colombia, 2012) con 500 objetos de aprendizaje.

Inicialmente se ejecutó el algoritmo de generación de la matriz de consulta. Este proceso demoró alrededor de 5 minutos en un computador de escritorio, con procesador Intel® Core™ i5 y sistema operativo Windows 7. Este proceso solo se realiza una vez para todo el repositorio. Debe actualizarse cada vez que se ingrese un nuevo objeto o se modifiquen los metadatos de los existentes.

Posteriormente, con la participación de tres expertos, se realizaron búsquedas en la herramienta básica del repositorio (SQL) y haciendo uso del algoritmo con LSA. Para cada par de resultados, ellos evaluaron por consenso o mayoría simple, la precisión desde el punto de vista de contexto (con intencionalidad del usuario) y sin contexto, es decir, el objeto recuperado tiene que ver con el tema sin considerar lo esperado por el experto (ver Tabla 2). Los temas buscados fueron en las disciplinas de programación e inteligencia artificial, que son las áreas de desempeño de los expertos.

Algoritmo 1 Generación de la matriz índice

Entrada: Metadatos de OAs del repositorio
Salida: $\mathbf{A_K} \in \mathbb{R}^{m \times n}$

```
1: M ← 0
2: N ← 0
3: j ← 0
4: para todo OA hacer
5:         leer tit,keyw,desc
6:         concat ← tit + keyw + desc
7:         S ← tokenizer (concat)
8:         S ← eliminaStopWords (S)
9:         N_j ← S
10:        M ← M ∪ S
11:        j ← j + 1
12: fin
13: A ← GeneraMatrizFrec ( M,N )
14: A_k ← LSA ( A )
15: devolver A_k
```

Con respecto al algoritmo de búsqueda LSA, se definió un umbral de coseno de 0.05, por encima del cual se mostraba los resultados de la búsqueda, valor determinado de manera empírica.

Algoritmo 2 Consulta de OAs del repositorio

Entrada: $\mathbf{A_k} \in R^{m \times n}$, *consulta, umbralCos*
Salida: **csn** $\in R^{1 \times n}$

1: S ← ***tokenizer*** (*concat*)
2: cons ← ***eliminaStopWords*** (cons)
3: **q** ← ***generaVectorFrec*** (cons)
4: ***para*** *j* ← 1 *hasta n* ***hacer***
5: $c_j \leftarrow \langle \mathbf{q}^T, \mathbf{d_j} \rangle$
6: $csn_j \leftarrow \frac{c_j}{\|q\|\|d_j\|}$
7: ***fin***
8: **csn** ← *ordena* (**csn,** *´descendente´*)
9: ***para*** *j* ← *hasta n* ***hacer***
10: ***si*** $csn_j > umbralCos$ ***entonces***
11: ***imprimir*** csn_j
12: ***fin***
13: ***fin***

4. RESULTADOS Y ANÁLISIS

Un primer elemento a considerar es el tiempo de generación de la matriz índice de consulta. Para 500 objetos, el tiempo que se tardó la generación puede considerarse significativo. Esto se debe a que la descomposición en valores singulares tiene complejidad algorítmica del orden $m^2n + n^3$ (Golub & Van Loan, 2012). Un factor a favor de su uso, es que solo hay que hacerlo una vez para todo el repositorio. Solo se debe volver a generar la matriz, cuando se agreguen nuevos objetos de aprendizaje al repositorio o haya una modificación en los existentes, en los metadatos que tienen que ver con la consulta. Después de generada la matriz, las consultas son ágiles.

Los experimentos llevados a cabo con los criterios antes señalados muestran que cuando las búsquedas se realizan sin contexto, la búsqueda básica SQL y con LSA, con una y dos palabras, su precisión es casi equivalente. En el momento que se realiza con tres palabras, se da una mejora sustancial en la precisión de la búsqueda con LSA respecto a la búsqueda básica (Figura 4).

En lo que tiene que ver con la búsqueda con intencionalidad (con contexto), a partir del uso de dos palabras, el algoritmo de búsqueda con LSA produce mejores resultados (Figura 5).

En ambos casos para LSA, al aumentar el número de palabras de la búsqueda, hay una mejora en la precisión respecto a la búsqueda básica. Esto es congruente con la literatura, ya que la búsqueda semántica realizada a través de la matriz transformada $\boldsymbol{A_{K'}}$ incluye ya las relaciones latentes entre todos los términos y documentos del repositorio, lo que favorece una mayor similitud en la búsqueda, cuando el número de términos crece. En cambio, la búsqueda básica se fundamenta solo en la coincidencia individual de las palabras de búsqueda con cada uno de los documentos, sin tener en cuenta el contexto en el que se encuentra dicho término dentro del documento en particular y el repositorio en general.

TABLA 2. Precisión en búsquedas en el repositorio FROAC con una, dos y tres palabras

Palabras de búsqueda	Básica SQL		con LSA	
	Sin contexto	Con contexto	Sin contexto	Con contexto
Agentes	0.44	0.17	0.95	0.32
Algoritmo	0.96	0.00	0.98	0.12
Artificial	1.00	1.00	1.00	0.00
Inteligencia	1.00	1.00	1.00	1.00
Lenguaje	0.60	0.10	0.55	0.36
Programación	0.94	0.00	1.00	0.00
Algoritmos lenguaje	0.55	0.06	0.62	0.09
Artificial agentes	0.51	0.49	0.93	0.91
Inteligencia agentes	0.52	0.52	0.86	0.89
Inteligencia artificial	0.67	0.33	0.56	0.67
Programación algoritmo	0.62	0.60	0.77	0.77
Programación lenguaje	0.07	0.07	0.29	0.29
Inteligencia artificial agentes	0.54	0.51	1.00	0.92
Programación algoritmos lenguaje	0.52	0.04	0.63	0.16

Otro elemento importante que se pudo apreciar del uso de la búsqueda con LSA, es que los resultados se presentan ranqueados, es decir, aparecen en primer lugar los objetos de aprendizaje con mayor similitud. Esto no ocurre con la búsqueda básica.

5. CONCLUSIONES

El LSA es un buen método para usarse en sistemas de recuperación de información, en entornos de repositorios de objetos de aprendizaje, en donde se manejen algunas

decenas de miles de objetos de aprendizaje. Esta cifra debido al alto costo computacional, especialmente en tiempo, del cálculo de la matriz índice término-documento que como se mencionó arriba (Sección 4), tiene una complejidad algorítmica alta. Su posibilidad de realizar búsquedas semánticas de información, aumentará las posibilidades de acceso de los usuarios a los recursos educativos relevantes presentes en los repositorios.

Figura 4. Precisión en las búsquedas del repositorio (sin contexto)

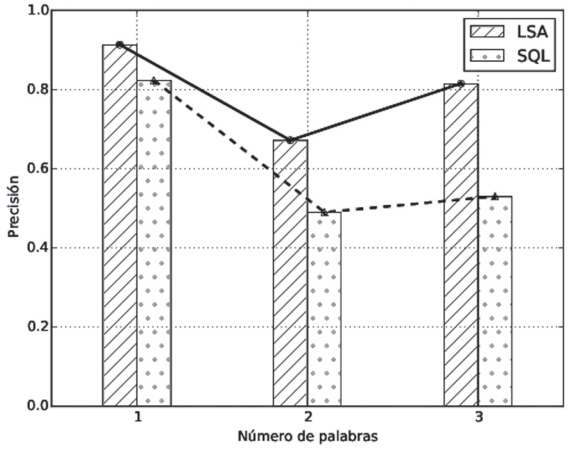
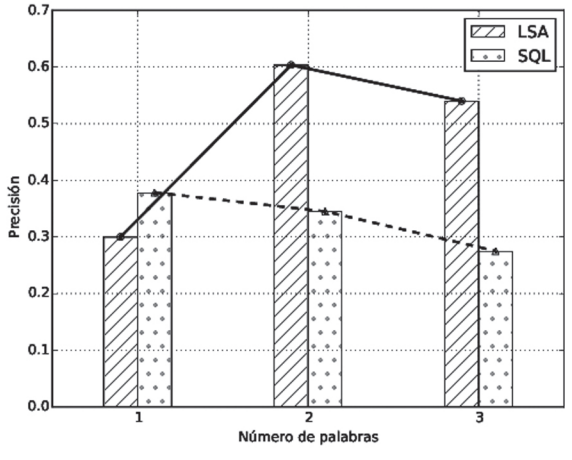


Figura 5. Precisión en las búsquedas del repositorio (con contexto)



El algoritmo puede mejorarse con la incorporación métodos de lematización, lo que permitiría aumentar la precisión desde el punto de vista de la semántica, ya que las variaciones de las palabras como las que están en plural, formas verbales, entre otras, quedarían incluidas como términos únicos y las búsquedas quedarían menos influidas por la exactitud en la escritura de las palabras. A su vez, disminuiría el número de términos, con el consecuente ahorro de tiempo de cálculo de la matriz índice término-documento.

Otro aspecto de mejora en el sistema propuesto, tiene que ver con el uso de algoritmos rápidos de descomposición en valores singulares de matrices, que permitirá su implementación en repositorios que contengan un gran volumen de objetos de aprendizaje. En este trabajo se experimentó con el contenido textual de tres metadatos. La implementación aquí presentada se puede extender a búsquedas de texto completo de objetos de aprendizaje textuales, si se tiene acceso al contenido textual completo del objeto de aprendizaje.

6. REFERENCIAS BIBLIOGRÁFICAS

[1] Alharbi, A., Henskens, F. & Hannaford, M. (2012). A Domain-Based Learning Object Search Engine to Support Self-Regulated Learning. International Journal of Computer and Information Technology 1(1), 83-93.

[2] Ali, N., Guéhénueuc, Y.-G., & Antoniol, G. (2013). Trustrace: Mining Software Repositories to Improve the Accuracy of Requirement Traceability Links. IEEE Transactions on Software Engineering, 39(5), 725–741.

[3] Astudillo, G. J. (2011). Análisis del estado del arte de los objetos de aprendizaje. Revisión de su definición y sus posibilidades. Technical report, Universidad Nacional de La Plata.

[4] Aswani Kumar, C., Radvansky, M., & Annapurna, J. (2012). Analysis of a vector space model, latent semantic indexing and formal concept analysis for information retrieval. Cybernetics and Information Technologies, 12(1), 34–48.

[5] Atkinson, J., Gonzalez, A., Munoz, M., & Astudillo, H. (2013). Web metadata extraction and semantic indexing for learning objects extraction. In Recent Trends in in Applied Artificial Intelligence, volume 7906 LNAI (pp. 31–140). Springer-Verlag Berlin Heidelberg.

[6] Baeza-Yates, R. & Ribeiro-Neto, B. (2011). Modern Information retrieval - the concepts and technology behind search. Essex: Addison Wesley.

[7] Barak, M. & Ziv, S. (2013). Wandering: A Web-based platform for the creation of location-based interactive learning objects. Computers & Education, 62, 159–170.

[8] Becerra, C., Astudillo, H., & Mendoza, M. (2012). Improving Learning Objects Recommendation Processes by Using Domain Description Models. LACLO, 3(1).

[9] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. Sebastopol: O'Reilly Media.

[10] Borg, M., Runeson, P., & Ardo, A. (2014). Recovering from a decade: a systematic mapping of information retrieval approaches to software traceability. Empirical Software Engineering, 19(6), 1565–1616.

[11] Brut, M. M., Sedes, F., & Dumitrescu, S. D. (2011). A semantic-oriented approach for organizing and developing annotation for E-learning. IEEE Transactions on Learning Technologies, 4(3), 239–248.

[12] Capobianco, G., De Lucia, A., Oliveto, R., Panichella, A., & Panichella, S. (2013). Improving IR-based traceability recovery via noun-based indexing of software artifacts. Software-Evolution and Process, 25(7), 743–762.

[13] De Lucia, A., Di Penta, M., Oliveto, R., Panichella, A., & Panichella, S. (2013). Labeling source code with information retrieval methods: an empirical study. Empirical Software Engineering, 1–38.

[14] Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391.

[15] Díaz M, L. G. (2007). Estadística Multivariada: Inferencia y Métodos. Bogotá: Universidad Nacional de Colombia.

[16] Dit, B., Revelle, M., Gethers, M., & Poshyvanyk, D. (2013). Feature location in source code: a taxonomy and survey. Journal of Software-Evolution and Process, 25(1), 53–95.

[17] Gil, A. B., De la Prieta, F., & López, V. F. (2010). Hybrid Multiagent System for Automatic Object Learning Classification. In E. S. Corchado Rodriguez (Ed.), Hybrid Artificial Intelligency Systems (pp. 61–68). San Sebastián, Spain: Springer.

[18] Golub, G. H. & Van Loan, C. F. (2012). Matrix computations, volume 3. JHU Press.

[19] Gracia, J.-M. (2002). Álgebra Lineal tras los buscadores de Internet. Technical report.

[20] IEEE (2002). Standard for Learning Object Metadata. Technical report, Institute of Electrical and Electronics Engineers, New York.

[21] Ismail, A. & Joy, M. (2011). Semantic Searches for Extracting Similarities in a Content Management System. In International Conference on Semantic Technology and Information Retrieval, number June, (pp. 113–118)., Putrajaya (Malaysia). IEEE.

[22] Jain, V. & Singh, M. (2013). Ontology Based Information Retrieval in Semantic Web: A Survey. International Journal of Information Technology and Computer Science, 10, 62–69.

[23] Laender, A. H. F., Gonçalves, M. A., Cota, R. G., Ferreira, A. A., Santos, R. L. T., & Silva, A. J. C. (2008). Keeping a Digital Library Clean : New Solutions to Old Problems. In Eighth ACM symposium on Document engineering, (pp. 257–262)., New York. ACM.

[24] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2-3), 259–284.

[25] López Guzmán, C. (2005). Los Repositorios de Objetos de Aprendizaje como soporte a un entorno e-learning. PhD thesis, Universidad de Salamanca.

[26] Martin, D. I. & Berry, M. W. (2007). Mathematical Foundations Behind Latent Semantic Analysis. In Handbook of Latent Semantic Analysis chapter 2, (pp. 35–55). New York: RoutLedge.

[27] Mihalcea, R. & Radev, D. (2011). Graph-Based Natural Language Processing and Information Retrieval. New York: Cambridge University Press.

[28] Ministerio_de_Educación_Nacional, C. (2012). Recursos Educativos Digitales Abiertos - Colombia. Bogotá: MEN.

[29] Moore, M. G. (2013). Handbook of Distance Education. New York.

[30] Nanba, H. & Okumura, M. (2005). Automatic Detection of Survey Articles. In Advanced Technology for Digital Libraries (pp. 391–401). Vienna, Austria: Springer.

[31] Novelli, A. D. P. & Parente De Oliveira, J. M. (2012). Simple Method for Ontology Automatic Extraction from Documents. International Journal of Advanced Computer Science and Applications, 3(12), 44–51.

[32] Oliphant, T. E. (2006). A Guide to NumPy. Trelgol Publishing.

[33] Peña, D. (2002). Análisis de datos multivariados. Madrid: McGraw Hill.

[34] Plaza Morales, L. (2011). Uso de Grafos Semánticos en la Generación Automática de Resúmenes y Estudio de su Aplicación en Distintos Dominios: Biomedicina , Periodismo y Turismo. PhD thesis, Universidad Complutense de Madrid.

[35] Poshyvanyk, D., Guéhéneuc, Y.-G., Marcus, A., Antoniol, G., & Rajlich, V. (2007). Feature Location Using Probabilistic Ranking of Methods Based on Execution Scenarios and Information Retrieval. IEEE Transactions on Software Engineering, 33(6), 420–432.

[36] Rodríguez M, P. A., Isaza, G., & Duque Méndez, N. D. (2012). Búsqueda personalizada en Repositorios de Objetos de Aprendizaje a partir del perfil del estudiante. Avances Investigación en Ingeniería, 9(1), 73–83.

[37] Sabitha, A. S. & Mehrotra, D. (2013). A push strategy for delivering of Learning Objects using meta data based association analysis (FP-Tree). In 2013 International Conference on Computer Communication and Informatics, (pp. 1–4). IEEE.

[38] Sánchez-Alonso, S., Ovelar, R., & Sicilia, M.-Á. (2007). Estándares de e-learning. In A. Landeta Etxeberría (Ed.), Buenas Prácticas de e-learning.

[39] Saum, R. R. (2007). An Abridged History of Learning Objects. In Learning Objects for Instruction: Design and Evaluation. Idea Group Inc (IGI).
[40] Smine, B., Faiz, R., & Desclés, J.-P. (2012). Extracting relevant learning objects using a semantic annotation method. In International Conference on Education and e-Learning Innovations, (pp. 1–6). IEEE.
[41] Stathopoulos, S. & Kalamboukis, T. (2015). Applying latent semantic analysis to large-scale medical image databases. Computerized Medical Imaging and Graphics, 39, 27–34.
[42] Tabares, V., Duque Méndez, N. D., Moreno, J., & Ovalle, D. (2014). FROAC - Federación de Objetos de Aprendizaje Colombia. In Novena Conferencia Latinoamericana de Objetos y Tecnologías de Aprendizaje, Manizales. Universidad Nacional.
[43] Tabares Morales, V., Duque Méndez, N. D., & Moreno, J. (2011). Evaluación experimental de la calidad en la recuperación de objetos de aprendizaje desde repositorios remotos. In Congreso de Ambientes Virtuales Adaptativos CAVA 2011, Bogotá.
[44] Universidad Nacional de Colombia, G. (2012). Federación de Repositorios de Objetos de Aprendizaje Colombia.
[45] Van de Sompel, H., Chute, R., & Hochstenbach, P. (2008). The aDORe federation architecture: digital repositories at scale. International Journal on Digital Libraries, 9(2), 83–100.
[46] Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013). ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. Decision Support Systems, 54, 1442–1451.
[47] Zamani, S., Lee, S. P., Shokripour, R., & Anvik, J. (2014). A noun-based approach to feature location using time-aware term-weighting. Information and Software Technology, 56(8), 991–1011.