

Identificación automática de artículos indexados en Latindex

Cullcay J., Ortiz J., Sumba X., Sumba F., Saquicela V.

Departamento de Ciencias de la Computación, Universidad de Cuenca, Av. 12 de Abril y Agustín Cueva, Cuenca, Ecuador, EC010201.

Autores para correspondencia: jose.cullcay@ucuenca.edu.ec, {jose.ortizv, xavier.sumba93, freddy.sumba}@ucuenca.ec, victor.saquicela@ucuenca.edu.ec

Fecha de recepción: 17 de mayo 2017 - Fecha de aceptación: 12 de Julio 2017

ABSTRACT

Identifying researchers with accepted articles in relevant indexed repositories has become increasingly important in higher education, especially in Ecuador, where Latindex is one of the most popular repositories. However, there is no automatic method to identify if an article has been indexed in that repository and currently higher-education institutes (HEI) in Ecuador have to manually recollect data about their indexed publications, providing control entities with information difficult to verify. For this reason, in this paper we present an approach to allow HEI and educational authorities to find publications that are indexed in Latindex using a set of strategies, with the aim of providing a process to identify indexed publications. Additionally, we implemented this approach as a prototype and evaluated it with a sample of publications of Ecuadorian researchers, demonstrating that the approach is both practical and useful for the mentioned case.

Keywords: Latindex, textual similarity methods, higher education, query languages.

RESUMEN

Identificar investigadores con artículos aceptados en repositorios indexados se ha vuelto más importante en la educación superior, especialmente en Ecuador, donde uno de los repositorios indexados más utilizados es Latindex. A pesar de ello, no existe un método automático para identificar si un artículo está indexado en dicho repositorio, y las instituciones de educación superior (IES) en Ecuador tienen que recolectar esos datos manualmente acerca de las publicaciones indexadas; proveyendo a las entidades de control con información difícil de verificar. Por esta razón, en este trabajo se presenta un enfoque que permite a IES y autoridades de educación encontrar publicaciones indexadas en Latindex, mediante la utilización de un conjunto de estrategias. Adicionalmente, se implementa el proceso como un prototipo que se evalúa con un conjunto de publicaciones de autores ecuatorianos, demostrando que el enfoque es tanto práctico como útil para el caso mencionado.

Palabras clave: Latindex, métodos de similitud textual, educación superior, lenguaje de consulta.

1. INTRODUCCIÓN

Los investigadores de las Instituciones de Educación Superior (IES) están obligados a publicar sus resultados para contribuir al avance de la ciencia. En el ámbito local, la información de las publicaciones indexadas de cada IES es de interés para organismos de control como el Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior (CEAACES)¹. Organismo que requiere estadísticas del número de publicaciones indexadas anualmente, puesto que sirven como parámetros para la categorización de universidades. La categoría de una universidad es importante,

¹ <http://www.ceaaces.gob.ec>

porque de ésta depende la asignación de recursos y su impacto en la producción científica del país. Por esta razón, las Instituciones de Educación Superior (IES) y sus autoridades brindan una considerable importancia a este indicador.

Los organismos reguladores de la educación en el Ecuador principalmente ponen énfasis en el número de artículos indexados en Latindex y Scopus. El sistema regional Latindex indexa revistas científicas de países latinoamericanos, España y Portugal. Este índice concentra un número importante de publicaciones relevantes que han sido creadas por investigadores ecuatorianos. Como consecuencia, tanto universidades como organismos de regulación están interesados en identificar publicaciones que aparecen en la base de datos de Latindex. La base de datos bibliográfica Scopus reúne títulos y publicaciones, que consiste en metadatos de todas las publicaciones que indexa, por lo que es factible realizar una comprobación de los artículos por IES. El caso de Latindex requiere mayor atención, debido a que dispone de un directorio de revistas (pero no de artículos), y de una página web llamada Portal de Portales Latindex (PPL)² que incluye información de algunos artículos publicados en revistas de acceso abierto que constituyen una muestra limitada del total de artículos indexados en Latindex. Adicionalmente, los motores de búsqueda online tales como Google Scholar³ o Microsoft Academic⁴ no permiten identificar si los artículos son indexados en Latindex.

El desafío que actualmente se presenta, es encontrar el número de publicaciones indexadas en Latindex para un autor relacionado con una IES. Al momento, la solución a este problema es de forma manual, donde las universidades solicitan a sus investigadores un listado de sus artículos indexados que luego es enviada a los organismos reguladores como CEAACES. Debido a la falta de una base de artículos indexados en Latindex, los investigadores listan manualmente sus artículos para que los organismos de control realicen una verificación manual. Este enfoque, a más de ser tedioso, es muy propenso a errores y su comprobación toma tiempo.

En este trabajo se presenta un enfoque para resolver el problema planteado, que permite verificar si un artículo de un investigador ecuatoriano está indexado en Latindex o no. Para esto se ha implementado un sistema que identifica automáticamente los artículos indexados en Latindex. El trabajo parte desde la obtención de los metadatos de artículos científicos, para luego hacer una comparación contra el directorio de revistas Latindex. Posterior a esto, se presenta un método de verificación de los resultados obtenidos. Además, se presenta un método para verificar si un artículo es Latindex cuando se cuenta con metadatos limitados. Finalmente, se pone a prueba el sistema y se presentan los resultados, conclusiones y trabajos futuros.

2. TRABAJOS RELACIONADOS

Un número considerable de bases bibliográficas digitales (*Scopus, Web of Science, Google Scholar, Thompson Reuters*) mantienen información no solo de sus revistas científicas indexadas, sino también de sus artículos, lo cual facilita búsquedas y análisis bibliométricos. Sin embargo, existen excepciones como Latindex (Flores, Penkova & Román, 2010) que solamente almacena información pertinente a las revistas. La información del lugar en donde fue publicado un artículo es útil para estudios bibliométricos e incluso en muchos casos para definir la categoría de una institución, como es en el caso de Ecuador. Varios casos de estudios bibliométricos han sido realizados utilizando el índice de Latindex como en los trabajos de Compte Pujol, De Urquijo & Matilla (2016) y Morales-Morante (2015), pero la selección de publicaciones debe ser realizada manualmente, por lo que existe la necesidad de una automatización.

Por otra parte, existen varias técnicas para la detección de similitud de textos, las cuales son sintácticos, semánticos o híbridos (Gomaa & Fahmy, 2013). Otros enfoques usados es la desambiguación de entidades que pueden ser resumidos en métodos supervisados o no supervisados (Nadeau & Sekine, 2007; Sarmiento, Kehlenbeck, Oliviera & Ungar, 2009). En general, los enfoques supervisados generan un modelo de clasificación basado en diferentes atributos, el cual suele estar

² <http://www.latindex.ppl.unam.mx/index.php/index>

³ <https://scholar.google.com.ec/>

⁴ <http://academic.research.microsoft.com/>

afinado a un problema específico para el cual se entrena y suele dar buenos resultados. Sin embargo este enfoque conlleva la necesidad de disponer de datos previamente clasificados, que en algunos casos puede ser difícil de obtener. Los enfoques no supervisados son usados para agrupar entidades en base a una métrica. Estos enfoques por lo general requieren de menor intervención humana, aunque aún requieren de parametrización que debe ser afinada en función del problema. En ambos casos ya sea supervisado o no supervisado muchas veces se usa la ayuda de entidades de Wikipedia o buscadores para realizar la desambiguación (Cucerzan 2007).

Por lo tanto, facilitar la búsqueda y clasificación de publicaciones es una tarea importante para estudios bibliométricos. En este trabajo se determina si un artículo fue publicado en una revista con indexación Latindex o no y así evitar la abrumadora tarea de realizar este proceso manualmente, esto es alcanzado usando técnicas de desambiguación y estrategias de similitud.

3. PROCESO DE IDENTIFICACIÓN DE ARTÍCULOS INDEXADOS EN LATINDEX

El objetivo de este trabajo es identificar publicaciones indexadas en Latindex. Para realizar esto, se necesita obtener metadatos acerca de las publicaciones de autores y buscar una correspondencia con las revistas indexadas en Latindex. El proceso inicia con la recopilación de datos de artículos académicos publicados, sobre los cuales se identificará si fueron indexados en Latindex. Adicionalmente, el proceso para identificar publicaciones indexadas en Latindex requiere pasos intermedios que permitan mejorar la comparación, pues los nombres de las revistas no necesariamente son los mismos. Luego de esto se puede realizar una comparación de nombres de revistas para identificar si alguna de ellas está indexada en Latindex.

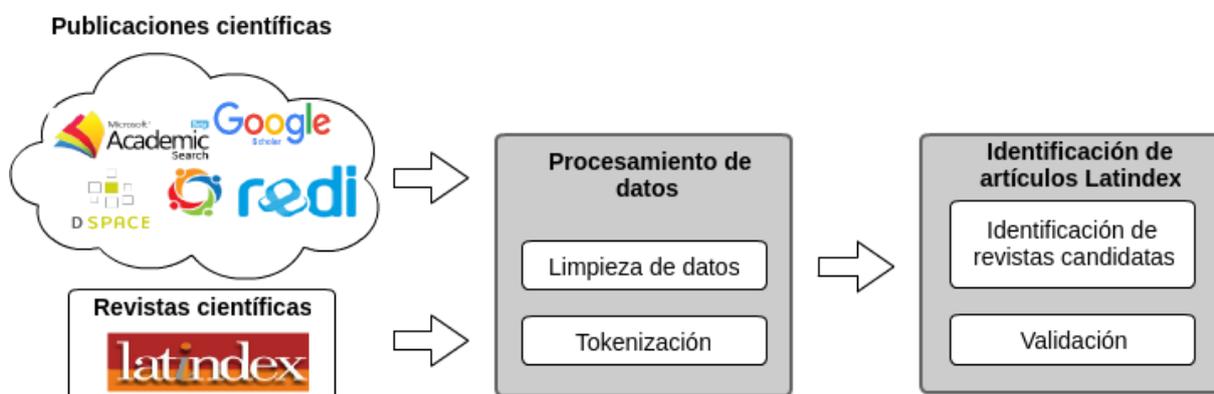


Figura 1. Proceso de detección.

En la Figura 1 se presenta de forma general el proceso de detección y comprobación planteado. En una primera etapa, la información de publicaciones y revistas indexadas es extraída desde las fuentes bibliográficas. Esta información es procesada con el fin de mejorar su calidad antes de comenzar con el proceso de detección. La comparación entre publicaciones y revistas se realiza a través de varias estrategias, esto genera una lista de revistas candidatas para una publicación. Finalmente, se aplica una verificación de los resultados parciales obtenidos a través de búsquedas en la Web. A continuación, se detalla cada una de las etapas del proceso.

3.1. Publicaciones científicas

En esta etapa, se realiza una extracción de artículos académicos (por investigador). Esta información ha sido recopilada por el proyecto REDI (Repositorio Ecuatoriano de Investigadores)⁵ (Sumba, Sumba,

⁵ <http://redi.cedia.org.ec/>

Tello, Baculima, Espinoza *et al.*, 2016) y se encuentra disponible en formato RDF⁶, la misma que puede ser consultada mediante sentencias SPARQL⁷. Los datos extraídos son el título del artículo, resumen, palabras clave (opcional), año de publicación (opcional) y el ISSN (o nombre de la revista) en caso de estar disponibles. Esta información permite caracterizar a los artículos y es utilizada dentro del proceso de detección de correspondencias a revistas Latindex. Cabe mencionar que este proceso también puede ser utilizado sobre otros indexadores que presenten el mismo problema.

3.2. *Revistas científicas*

En esta etapa se obtiene un listado de todas las revistas indexadas en Latindex a través del servicio web⁸ que ofrecen para búsquedas. El listado de revistas se obtiene en formato JSON y contiene información acerca del nombre de la revista, su ISSN, año inicial de publicación, editorial, etc. Esta información es la que se utiliza para realizar una comparación con la lista de publicaciones. Existen dos observaciones necesarias para la identificación de artículos indexados en Latindex que hicieron preferible la estrategia de descargar la información del listado de las revistas. La primera es que si bien está disponible un servicio de búsqueda de Latindex para consultar nombres de revistas indexadas, este proceso es difícil de automatizar debido a que los nombres de las revistas no necesariamente son idénticos o muy parecidos. Un caso común es la revista *Maskana*, la cual aparece en la base de datos de artículos de REDI como “*Maskana. Revista Científica*”, mientras que la misma aparece como “*Maskana*” en el directorio de Latindex. Esto hace que al buscar ese nombre en el directorio de Latindex, no se obtengan resultados y que la única forma de identificación sea manual; esto puede ser cierto para muchos buscadores de artículos. La segunda es que el servicio web de búsquedas de Latindex está sujeto a las limitaciones de cualquier servicio web, es decir que se depende de la confiabilidad de la red, la disponibilidad del servicio (pueden existir pérdidas de servicio), las posibles limitaciones de tráfico y consultas, etc. Por estas razones se decidió descargar los datos completos y realizar una comparación local de nombres de revistas.

3.3. *Procesamiento de datos*

En esta etapa se realiza una limpieza de los datos obtenidos, en especial de los nombres de las revistas. Una de las ocurrencias más comunes a limpiar es la aparición del formato de publicación de la revista como parte del nombre. Por ejemplo, la revista “*Iberoamericana*” aparece como “*Iberoamericana (Online)*” en el directorio de Latindex, dificultando así la comparación entre nombres de revistas. Por eso se aplica una limpieza de palabras vacías (*stopwords*) a los nombres de revistas. Entre las palabras vacías que se eliminan se encuentran: “Print”, “En línea”, “Impresa”, “Online”, “Internet”, “Imprimé”, “On-line”, “Impresso”, “En ligne”. Algunas de ellas (como Online e Internet) deben estar antes de un paréntesis cerrado “)” para considerarse como palabras vacías, pues cuando no están entre paréntesis pueden ser parte del nombre de una revista. El proceso de limpieza de datos depende mucho del estado de los datos en la fuente.

Luego se realiza una tokenización de los nombres de las revistas, tanto las que se extrajeron del repositorio de investigadores, como las que provienen de Latindex. Esto se justifica en que, como se mencionó anteriormente, existen casos en los que el nombre está mezclado con otras palabras que dificultan su comparación. Un ejemplo es la revista Deusto, que aparece en el directorio Latindex como “*Deusto, estudios cooperativos (Ed. impresa)*”, en donde se puede eliminar la parte entre paréntesis en el paso anterior, pero aún el nombre no está claro al aparecer como “*Deusto, estudios cooperativos*”. Esta revista puede ser la misma que “*Deustos*” o que otras con nombre similar, como la “*Revista Estudios Cooperativos*” de Uruguay. Por esta razón se divide cada nombre de la revista en *tokens*, los cuales representarán los posibles nombres de las revistas. En todos los casos se prioriza el primer *token* para las comparaciones, pues se observa que generalmente el nombre de la revista va al principio del título completo.

⁶ Resource Description Framework

⁷ SPARQL Protocol and RDF Query Language

⁸ <http://latindex.org/latindex/Solar/Busqueda>

3.4. Identificación de revistas candidatas

La identificación de publicaciones Latindex ha sido implementada en un proceso que consta de dos fases: La primera comprende la comparación directa del ISSN que proviene de los metadatos de los artículos con los registros del catálogo de Latindex. La segunda fase consiste en comparar el nombre de la revista en el artículo con las revistas Latindex (tokenizadas) y encontrar una lista de revistas candidatas en base a comparaciones sintácticas. Posteriormente se procede a la verificación del emparejamiento realizado entre revistas mediante una búsqueda de términos exactos en la Web con los datos del artículo y los de la potencial revista Latindex.

La primera fase identifica si un artículo está indexado en Latindex mediante el número ISSN. Para ello, se obtiene el ISSN de entre los metadatos del artículo (en caso de que se disponga de esta propiedad) y luego se busca en el directorio de revistas de Latindex. Si se encuentra una revista en Latindex que contenga ese código ISSN, entonces el artículo original se marca inmediatamente como indexado en Latindex. Un ejemplo es el caso del ISSN 1390-4450 perteneciente a la Revista de la Facultad de Ciencias Médicas de la Universidad de Cuenca, la cual aparece en los metadatos de un artículo y que también está presente en la base de datos de Latindex, por lo que se coloca el artículo como candidato a ser indexado en Latindex.

La segunda fase involucra la comparación entre los *tokens* provenientes del nombre de las revistas, para identificar si alguno de esos nombres es similar a una revista Latindex. Para comparar se prioriza el primer *token*, pues se observa que generalmente el nombre de la revista va al principio. Por ejemplo, se comparan el nombre de revista “*Maskana. Revista Científica*” que viene del repositorio de artículos, y se encuentra que su primer token “Maskana” coincide con la revista “*Maskana*” del directorio Latindex. Para realizar la comparación sintáctica de una manera flexible, es decir, que tolere pequeñas diferencias en los nombres, pero sin perder efectividad, se utiliza un promedio de las métricas del Coseno, Levenshtein y Jacaré (Cohen, Ravikumar & Ginberg, 2003). Estas métricas permiten medir diferentes aspectos de similitud entre las cadenas de texto a considerar, donde el valor 0 significa que las cadenas no tienen ninguna similitud, mientras que un valor de 1 significa que son idénticas. Entonces se calcula el coeficiente de similitud entre las revistas de los artículos de REDI con los nombres obtenidos desde Latindex, mediante la métrica seleccionada. Si el coeficiente es superior a un límite preestablecido, la revista comparada es seleccionada. Según las pruebas empíricas realizadas, un límite apropiado para el coeficiente de similaridad es 0.85. Por ejemplo, uno de los artículos extraídos incluye el nombre de la publicación como “Revista de la Facultad de Ciencias Médicas”, la cual tiene como coincidencia el nombre de la “Revista de la Facultad de Ciencias Médicas de la Plata” de Latindex con un coeficiente de 0.88. Todas las coincidencias (revistas seleccionadas) son almacenadas en un listado para ser enviado al siguiente paso.

3.5. Validación

Si bien la identificación de revistas candidatas es una herramienta útil para reducir el campo de búsqueda, sus resultados requieren de un refinamiento posterior para comprobar su validez. Esto debido al alto nivel de ruido o errores de ambigüedad que pueden producirse en las comparaciones sintácticas del nombre de la revista. Por ejemplo, es común encontrar revistas con nombre genérico como ‘Revista de Medicina’ o ‘Revista científica’. Debido a estos casos de ambigüedad no se puede establecer de forma directa la relación de una publicación con una revista candidata. Es por esta razón que las coincidencias entre nombres de revistas pasan a un proceso de verificación para comprobar que el artículo pertenezca realmente a una revista Latindex. La estrategia de verificación planteada en este trabajo se basa en la observación de que la mayoría de artículos indexados en Latindex poseen fichas de información o accesos libres de texto completo en la Web. Y que en dichas páginas es común encontrar los metadatos de los artículos (título, resumen, palabras clave) junto con información de la revista en la que fue publicado (nombre de la revista y/o ISSN).

La verificación propuesta se realiza de forma automática mediante una búsqueda Web de varios términos: título de la publicación, resumen, nombre de la revista candidata, ISSN de la revista candidata. Específicamente se utilizó el buscador Web Bing⁹ debido a que posee una API gratuita y permite realizar

⁹ <https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/>

búsquedas exactas de términos. Si se obtienen resultados a través de los términos buscados, se considera que el artículo ha sido efectivamente publicado en una revista Latindex. Si no es el caso se reducen ciertos términos de búsqueda (ISSN y *abstract*) para realizar comprobaciones menos estrictas. Finalmente, si no se puede establecer una relación entre la publicación y la revista, el proceso termina. Para ilustrar este caso, se toma el ejemplo de un artículo publicado en la revista “*Estoa*”, que tiene una similitud sintáctica con la revista “*Estoa: Revista de la Facultad de Arquitectura y Urbanismo de la Universidad de Cuenca*” luego de comparar los tokens, pues se compara el nombre “*Estoa*” con el primer token “*Estoa*” antes del símbolo “:”. Para el caso del artículo “Validación de un test de ureasa para diagnóstico de helicobacter pylori en comparación con el clotest y en referencia a la histología”, se encuentra una revista candidata de Latindex con el nombre “Revista de la Facultad de Ciencias Médicas de la Universidad de Cuenca” que tiene el ISSN “1390-4450”, se valida mediante la API de Bing. La validación indica que existen resultados en la Web que contienen el título del artículo, su ISSN y el nombre de la revista candidata de Latindex, lo que permite concluir que el artículo está indexado en Latindex.

Artículos con revista no identificada

En algunos casos las publicaciones científicas extraídas carecen de la información referente a la revista que publica el artículo científico como el nombre o el ISSN, lo cual dificulta la identificación de pertenencia al catálogo Latindex. Para estos casos se ejecuta dos procedimientos adicionales con el objetivo de conocer la revista científica en la que fue publicado el artículo, los cuales se listan a continuación.

- √ *Extracción de keywords:* En este proceso se extraen las palabras clave (*keywords*) de la publicación, y el listado de las revistas Latindex con sus temas y subtemas. El objetivo de este proceso es caracterizar las temáticas o áreas de conocimiento tanto de las revistas como de la publicación. Un ejemplo es el artículo “Prevalencia de portadores nasales de *Staphylococcus aureus* en el personal del Hospital Vicente Corral Moscoso y Hospital Militar, patrón de sensibilidad antimicrobiana”, la cual incluye las palabras clave “prevalencia”, “senos paranasales”, “*staphylococcus aureus*”, “resistente”, “meticilina”, “infecciones bacterianas”, “microbiología”, “pruebas de sensibilidad”.
- √ *Selección de revistas Latindex similares al artículo científico:* Este proceso busca reducir el campo de búsqueda de revistas relacionadas con el artículo a través de una comparación semántica entre las palabras clave del artículo y los temas de las revistas. Esta comparación se realiza utilizando la distancia semántica *Normalized Wikipedia Distance* (Schaefer, Heinert & Gottron, 2014), la cual utiliza la base de conocimiento de Wikipedia para determinar la cercanía entre dos términos. Mediante observaciones de los datos se determinó que un valor de media armónica menor a 0.75 de las distancias entre revistas y la publicación proporciona buenos resultados. Para ilustrar el proceso, se considera el artículo mencionado en el punto anterior, donde se toman las palabras claves del mismo y se las compara con los temas y subtemas de las revistas Latindex, encontrándose “Revista de la Facultad de Ciencias Médicas de la Universidad de Cuenca” entre ellas, debido a que su temática es “Ciencias Médicas” y “Medicina”.
- √ *Detección de revista mediante búsquedas en la web:* Una vez que se tiene un conjunto de posibles revistas Latindex cuya temática sea similar a la tratada en el artículo, se verifica que su año de inicio de publicación no sea mayor que el año de publicación del artículo y se procede a realizar una verificación entre las revistas Latindex candidatas y el artículo, de forma similar como se menciona en secciones anteriores, utilizando búsquedas en la Web. De esta forma se pueden encontrar artículos indexados en Latindex mediante sus metadatos básicos (título, resumen y *keywords*), sin contar con el nombre de la revista donde fue publicada. En el caso del artículo mencionado anteriormente, se utilizó su título, el resumen, el nombre de las revistas candidatas Latindex y sus ISSN, encontrándose una coincidencia con “Revista de la Facultad de Ciencias Médicas de la Universidad de Cuenca”, por lo que se marca a ese artículo como indexado en Latindex.

4. EVALUACIÓN Y RESULTADOS

La evaluación de la propuesta presentada se elaboró en base a los artículos generados por investigadores de la Universidad de Cuenca. Esta información fue obtenida a través del proyecto REDI y consta de un grupo de 452 artículos de investigadores de esta institución. Para evaluar la efectividad de la propuesta se tomó un subconjunto aleatorio de 45 artículos que representan cerca del 10% de los artículos. En la Tabla 1 se presenta un resumen de las publicaciones utilizadas junto con la información con la que se disponía para realizar la detección.

Tabla 1. Subconjunto de artículos.

Metadatos disponibles	Nro. Publicaciones
ISSN	15
Nombre de revista	25
Título, resumen y palabras clave	5
Total: 45	

Este subconjunto fue sometido tanto a la asignación manual de un experto como al proceso de detección automático. La información obtenida de las dos fuentes fue contrastada tomando a la asignación del experto como *Gold Standard*. En la Tabla 2 se presenta un resumen de la comparación realizada. Donde las columnas ‘Latindex’ y ‘No Latindex’ representan el número de publicaciones que fueron identificadas por el experto como pertenecientes a revistas de Latindex y las que no respectivamente. Por otro lado, las filas ‘Positivos’ y ‘Negativos’ representan los resultados obtenidos por el proceso automático, donde se asigna como positivo si una publicación se detectó como Latindex y negativo en caso contrario.

Tabla 2. Resultados.

	Latindex	No Latindex
Positivos	18	1
Negativos	2	24
Total:	20	25

Los resultados obtenidos permiten calcular una precisión de 0.95 y una exhaustividad de 0.9, y por tanto un *f-score* total de 0.92. La precisión calculada demuestra que el proceso planteado tiene una alta confiabilidad al seleccionar publicaciones que realmente forman parte de Latindex, lo que se puede atribuir a la etapa de validación mediante búsquedas en la Web. Esto se evidencia en el bajo número de falsos positivos encontrados en el proceso de evaluación de los resultados. Además, de que se pudo comprobar que las asignaciones incorrectas se producen por inconsistencias en los metadatos provenientes de las fuentes y no por errores del proceso en sí.

Por otro lado, el valor obtenido de exhaustividad refleja que un pequeño número de publicaciones fueron incorrectamente identificadas como no pertenecientes a Latindex. Este inconveniente fue relacionado con particularidades en los metadatos de ciertas publicaciones y revistas. Por ejemplo, algunas publicaciones tienen registrado dentro de sus metadatos el ISSN de la versión online de sus revistas mientras que el directorio de Latindex cuenta con los ISSN de las versiones impresas. Este tipo de detalles dificultan en gran medida la automatización de la identificación y validación para estas revistas.

Otra causa de errores de falsos negativos fue relacionada con las publicaciones que tienen revistas sin identificar. Principalmente debido a que la comparación semántica entre las palabras clave de las publicaciones y los temas de las revistas puede ser imprecisa cuando las revistas cubren áreas diversas o muy generales. Por ejemplo, la revista ‘Maskana’ tiene entre sus temas de interés ‘Ciencias de la Ingeniería, Ciencias Médicas, Artes y Humanidades, Ciencias Exactas y Naturales, Ciencias Sociales’. Esto hace que sea muy difícil que publicaciones sobre temáticas puntuales como ‘Ingeniería de Software’ o ‘Web Semántica’ sean relacionadas con dicha revista. Sin embargo, este tipo de

publicaciones existen dentro de la revista Maskana y son ignoradas dentro del proceso por la falta de similitud semántica con los temas registrados en los metadatos de la revista.

5. CONCLUSIONES

La identificación y clasificación de publicaciones científicas es un requerimiento recurrente dentro de organizaciones y centros de investigación. Este problema es de especial interés en el contexto ecuatoriano, donde organismos rectores de la educación superior como el CEAACES utilizan esta información para procesos de acreditación y categorización de las IES. En este contexto surge la necesidad de desarrollar mecanismo automático para la detección de artículos dentro de sistemas de indexación como Latindex.

En el presente trabajo se presentó un proceso automático de detección de publicaciones indexadas por Latindex. Este proceso combina métricas de comparación tanto sintácticas como semánticas sobre los metadatos de las publicaciones y el catálogo de revistas de Latindex. Además, se presenta una estrategia de validación de los resultados obtenidos basada en búsquedas Web que mejora la efectividad obtenida.

Finalmente, la propuesta presentada fue aplicada y evaluada con información perteneciente a la producción científica de la Universidad de Cuenca. El proceso de evaluación se realizó mediante un *Gold Standard* elaborado por un experto, lo que permitió determinar la efectividad real de la propuesta presentada. Los resultados obtenidos en cuanto a precisión y exhaustividad demuestran la validez del método planteado e incentivan a su aplicación en un contexto más amplio.

En cuanto al trabajo futuro este se centrará en mejorar la visibilidad de los resultados obtenidos a través de la aplicación de tecnologías de Datos Enlazados (*Linked Data*). Además, se pretende incorporar esta información dentro del proyecto REDI para facilitar su consumo y explotación por parte de la comunidad de investigadores. Por otro lado, también se mejoraran los algoritmos utilizados en esta propuesta en especial para casos donde los metadatos disponibles de los artículos sean limitados.

AGRADECIMIENTOS

Este trabajo fue realizado en el Departamento de Ciencias de la Computación de la Universidad de Cuenca y ha sido patrocinado por RED-CEDIA¹⁰ como parte del proyecto “Repositorio Ecuatoriano De Investigadores” (REDI) y del grupo de trabajo de repositorios digitales.

REFERENCIAS

- Cohen, W., Ravikumar, P., & Fienberg, S. (2003). *A comparison of string distance metrics for name-matching tasks*. International Joint Conference on Artificial Intelligence, pp. 73-78.
- Compte Pujol, M., De Urquijo, B., & Matilla, K. (2016). La investigación en marcas de territorio y diplomacia pública en España. Un estudio bibliométrico de las revistas científicas españolas especializadas en Comunicación indexadas en Latindex (1980-2016). *Anales de Documentación*, 19(2), pp. 1-53.
- Cucerzan, S. (2007). *Large-scale named entity disambiguation based on Wikipedia data*. Disponible en: <https://www.semanticscholar.org/paper/Large-Scale-Named-Entity-Disambiguation-Based-on-W-Cucerzan/1c909ac1c331c0c246a88da047cbdcca9ec9b7e7>

¹⁰ <https://www.cedia.edu.ec>

- Flores, A. M., Penkova, S., & Román, A. (2010). *Once años de Latindex: una experiencia al servicio de las publicaciones científicas iberoamericanas*. Disponible en: <https://digital.csic.es/handle/10261/22942>.
- Gomaa, W., & Fahmy, A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13), 13-18.
- Morales-Morante, L. F. (2015). *Production and impact of Peruvian social science journals in the Latinex catalogue*. *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*. Disponible en: https://www.academia.edu/29726710/Production_and_impact_of_Peruvian_social_science_journals_in_the_Latinex_catalogue
- Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1), pp. 3-26.
- Sarmiento, L., Kehlenbeck, A., Oliveira, E., & Ungar, L. (2009). *An approach to Web-Scale named-entity disambiguation*. In: Perner, P. (Ed.). *Machine learning and data mining in pattern recognition*. Lecture notes in computer science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 689-703.
- Schaefer, C., Hienert, D., & Gottron, T. (2014). *Normalized relevance distance-a stable metric for computing semantic relatedness over reference corpora*. In: Proc. of the 21st European Conference on Artificial Intelligence (pp. 789-794). IOS Press.
- Sumba, X., Sumba, F., Tello, A., Baculima, F., Espinoza, M., & Saquicela, V. (2016). Detecting similar areas of knowledge using semantic and data mining technologies. *Electronic Notes in Theoretical Computer Science*, 329, 149-167.