Authors semantic disambiguation on heterogeneous bibliographic sources

José Ortiz*, José Segarra†, Xavier Sumba‡, José Cullcay§, Mauricio Espinoza¶, and Víctor Saquicela‖
Departamento de Ciencias de la Computación, Universidad de Cuenca,
Cuenca, Ecuador.
*jose.ortizv@ucuenca.ec, †jose.segarraf@ucuenca.ec, ‡xavier.sumba93@ucuenca.ec, §jose.cullcay@ucuenca.edu.ec
¶mauricio.espinoza@ucuenca.edu.ec,‖victor.saquicela@ucuenca.edu.ec

***Abstract*—Data ambiguity from various sources remains as a complex problem that affects services provided by digital libraries. From the point of view of integration of information from different sources, the challenge of author ambiguity is one of the most important, and there are numerous methods proposed to deal with this issue using different approaches. They generally work for some scenarios but they have important limitations, specially when dealing with heterogeneous sources. In this work, we review a group of existing methods and then propose a technique that combines some of them, also incorporating a measure of distance using semantic technologies to solve the ambiguity of authors while integrating bibliographic data from various sources. This technique has been successfully tested in disambiguating Ecuadorian authors from both internal sources (institutional repositories) and external digital libraries.**

I. Introducción

En trabajos anteriores, se abordó el problema de integración de fuentes bibliográficas a través de tecnologías semánticas, tal como se describe en [1] para repositorios digitales y en [2] para librerías digitales. En este contexto han surgido varios problemas principalmente relacionados con la identificación única de los recursos bibliográficos integrados. El caso de autores es especialmente relevante en dicho escenario, debido a que estos inconvenientes dificultan tareas como reconocimiento y asignación correcta de obras de un autor, que son indispensables en actividades tales como estudios bibliométricos y descubrimiento de nuevo conocimiento. Estos problemas han sido abordados por múltiples trabajos en la comunidad científica bajo la denominación de desambiguación de nombres de autores (*Author Named Disambiguation*, AND)[3].

La ambigüedad de nombres puede ser a causa de varios factores tales como: errores ortográficos, inconsistencias al ingresar datos, variaciones del nombre o uso de iniciales. Estos factores pueden ser englobados en tres grupos: el primero es cuando varias personas comparten el mismo nombre (homónimos); el segundo grupo están todas las representaciones del nombre de un autor (sinónimos) como por ejemplo "Mauricio Espinoza" y "M. Espinoza"; y errores tipográficos u ortográficos se encuentran en el tercer grupo. Estos problemas dificultan la identificación de autores por lo que además del nombre, se puede extraer otras características que agreguen el conocimiento suficiente para determinar si dos registros hacen referencia o no la misma persona. Es deseable que estos procesos de desambiguación sean automáticos debido a que la desambiguación manual es muy costosa en cuanto a tiempo sea esta a pequeña o gran escala y peor aún, cuando se tiene casos de nombres comunes, puesto que la incertidumbre aumenta.

En el presente trabajo, se plantea un proceso para desambiguar autores entre fuentes bibliográficas digitales usando tecnologías semánticas, el cual consiste en la generación de enlaces [4] entre recursos que representan a la misma persona. Estos enlaces son generados en base a estrategias sintácticas y semánticas. Este proceso ha sido probado en el proceso de desambiguación de autores en repositorios digitales de instituciones educativas y librerías digitales.

El trabajo ha sido organizado de la siguiente manera: en la sección II se presenta los antecedentes y trabajos relacionados. En la sección III se presentan los aspectos destacados de la propuesta planteada y el aporte realizado en este campo de investigación. En la sección IV se presenta los resultados obtenidos. En la sección V se describe las conclusiones, así como posibles trabajos futuros.

II. Antecedentes y trabajos relacionados

Existe una variedad de enfoques que tratan el problema de la desambiguación de autores que particularmente difieren en función de los datos disponibles y de las fuentes de información. Un método común para llevar a cabo esta tarea es el proceso de desambiguación manual [5], pero este proceso suele ser costoso y propenso a errores cuando conlleva una gran cantidad de información. Otros intentos proponen evitar el problema de ambigüedad entre autores mediante la generación de un identificador único, como es el caso del Sistema de identificación de autores universal (Universal Author Identifier System, UAI_Sys) [6] o el uso del ORCID (Open Researcher Contributor Identification) [7]. Sin embargo, estas propuestas requerirían de la colaboración voluntaria de autores y de su adopción generalizada por parte de las fuentes bibliográficas digitales, lo cual es impráctico en escenarios reales.

Frente a las dificultades presentadas por los métodos manuales de desambiguación y la lenta adopción de medidas como identificadores universales, varias técnicas de desambiguación automáticas y semiautomáticas han sido propuestas tal como se resumen en [8][3]. Entre estas técnicas se destacan las que utilizan asistencia o entrenamiento por parte de una persona (supervisados), así como las que emplean algoritmos que no

requieren asistencia de una persona, por lo que están orientados hacia una automatización completa (no supervisados). En general, los enfoques supervisados generan un modelo de clasificación basado en diferentes atributos de los autores, el cual suele estar afinado a un problema específico para el cual se entrena y suele dar buenos resultados. Sin embargo, este enfoque conlleva a la necesidad de disponer de datos previamente clasificados, que puede ser difíciles de obtener en algunos casos. Algunos ejemplos de estos enfoques pueden utilizar algoritmos de regresión logística tal como se expone en [9], donde se calcula y asigna una métrica de relación entre publicaciones para luego formar un grafo que permite identificar publicaciones asociadas a un autor usando para esto el algoritmo de generación de comunidades (algoritmo de Blondel). Otros algoritmos de clasificación ampliamente usados son SVM, Random Forest, k-Nearest Neighbors (kNN), árboles de decisión y bayes que pueden ser encontrados en los trabajos de [10][11][12][13]. Como se afirma en [10], en general se obtiene mejores resultados con random forest en lugar de SVM. Los enfoques no supervisados son usados para agrupar varios atributos de autores usando diferentes métricas de similitud que permiten obtener distancias entre sus elementos. Entre los algoritmos de aprendizaje no supervisado están algoritmos como DBSCAN[14] y k-way spectral *clustering* [15] el cual supera a resultados con k-means. Estos enfoques por lo general requieren de menor intervención humana, aunque requieren la afinación de parámetros en función del problema.

Enfoques modernos utilizan técnicas combinadas como en [16] que emplea *fingerprints* y *clustering*, obteniendo representaciones de los contenidos de los documentos en forma de hash, para luego comparar de forma rápida y agrupar mediante clustering. En el trabajo de [14] se usa LASVM (una variante de SVM) para calcular la distancia entre las publicaciones y luego agrupa las publicaciones por autor usando DBSCAN. Sin embargo, estas técnicas aún requieren de intervención manual para casos concretos como cuando una obra es asignada mediante *clustering* a varias personas.

Adicionalmente, existen enfoques que varían en función a la información a la que pueden acceder en su proceso para determinar la identidad de un autor. En estos enfoques, por lo general, se toma información provista por las mismas publicaciones tales como nombres de autores, keywords, coautores [17]. Otras emplean información adicional que comúnmente es tomada de fuentes de la Web, como [18] que usa Wikipedia como base de conocimientos y [19] que usa la estructura de los enlaces de páginas Web y un método de *clustering* para desambiguar nombres de personas.

Aunque hasta el momento existen varios métodos para la desambiguación de autores, es evidente que muchos de los enfoques dependen de las situaciones particulares tal como se concluye en [3]. La mayoría de trabajos en AND se enfocan en librerías digitales; sin embargo, en el presente trabajo con el fin demostrar la flexibilidad del presente enfoque, se trata el problema tanto sobre librerías digitales como en repositorios institucionales. Para esto, se emplea una combinación de métricas tanto sintácticas como semánticas que son empleadas sobre campos comunes que disponen los recursos bibliográficos disponibles en estas fuentes.

III. Proceso de desambiguación semántica de autores

El problema de la desambiguación de autores ha sido extensamente estudiado por gran parte de la comunidad científica y bibliográfica. Las bases digitales que almacenan e indexan contenido científico tales como DBLP[1], Scopus[2], Google Schoolar[3], entre otras, han recibido especial atención por su importancia, concentrando varios de estos esfuerzos que abogan en mejorar la calidad de sus contenidos. Sin embargo, otras fuentes de información como los repositorios institucionales de tipo académico han quedado rezagadas en la aplicación de estas nuevas técnicas, debido principalmente a sus particularidades (estándares, formatos, etc.). Este problema se evidencia especialmente en la falta de mecanismos de desambiguación de autores fuera del ámbito de las bases digitales.

El presente trabajo propone una solución integral al problema de desambiguación de autores para fuentes bibliográficas. Solución que hereda las principales ventajas de los planeamientos existentes en el ámbito de bases digitales y que pretende expandir su utilidad hacia otras fuentes de información. Para esto, se utilizan los principios de datos enlazados que permiten crear una capa de abstracción sobre las fuentes originales simplificando el problema. En la figura 1 se ilustra un ejemplo de la problemática de desambiguación de autores bajo esta perspectiva, donde las fuentes de información han pasado por una etapa de conversión a RDF[4] (*RDF-zation*) que estandariza la información de fuentes heterogéneas. De esta manera la desambiguación se reduce a un problema de enlazado de recursos en la Web.

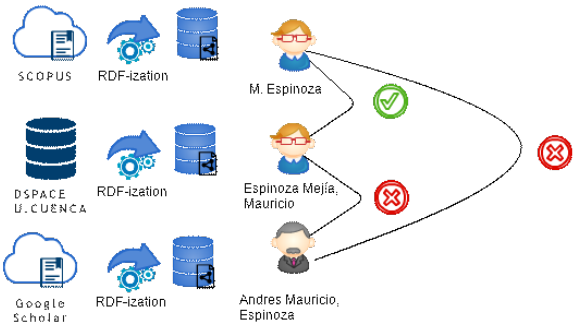

Figura 1. Problema de desambiguación

III-A. Proceso de desambiguación de autores

Como se mencionó anteriormente, la propuesta presentada parte del supuesto que todas las fuentes se encuentren disponibles como datos enlazados (*Linked Data*) y que puedan

[1]http://dblp.uni-trier.de/
[2]https://www.scopus.com/
[3]https://scholar.google.com
[4]Resource Description Framework

ser accedidos a través de los estándares de la Web Semántica (RDF/SPARQL). Adicionalmente, es deseable que la información haya sido generada usando ontologías estándar para representar información bibliográfica como: foaf[5], bibo[6], etc. Si bien los algoritmos propuestos dentro de la desambiguación son independientes de estos formatos, su utilización es recomendable debido a que fomentan la estandarización y facilitan la integración a nivel Web de los recursos.

En la figura 2 se presenta el proceso seguido, en el cual se definen tres etapas principales: la primera etapa es la caracterización de los autores que tiene por objetivo la obtención de las características más importantes de los autores que se van a desambiguar, esto a través de la recopilación de su afiliación, metadatos de sus publicaciones y coautores; la segunda etapa es el análisis semántico que se encarga de descubrir información relevante a partir de la información recabada, detectando tópicos o áreas de interés y filtrando información de poca utilidad; finalmente, en la etapa de evaluación se comparan semánticamente todos los autores (registros ambiguos) en base a varias métricas, permitiendo definir si efectivamente se tratan del mismo individuo.

En las siguientes secciones se presentan a más detalle cada una de las etapas presentadas, especificando los procesos internos que realizan y las consideraciones de desarrollo. Con el fin de facilitar la comprensión del proceso, se ejemplifica cada una de estas etapas con un ejemplo real de ambigüedad entre autores encontrados en los repositorios institucionales de las universidades del Ecuador. Específicamente se tomará al autor "Mauricio Espinoza Mejía", docente e investigador de la Universidad de Cuenca que ha realizado varias colaboraciones con otras instituciones y del cual se conocen varias representaciones ambiguas dentro de los repositorios y bases digitales tal como se ejemplifica en la figura 1.

III-B. Extracción y caracterización de autores

Para extraer la información de los autores se utilizan los *SPARQL Endpoints* de cada una de las fuentes de datos que van a ser tratadas. Así, mediante consultas SPARQL simples y usando los modelos ontológicos definidos dentro de los repositorios de datos enlazados se obtiene una lista de autores. Una vez obtenido dicha lista se completa la información de cada autor, obteniendo los metadatos de sus documentos asociados, afiliación y coautores.

La lista de autores se obtiene consultando todas las instancias o entidades de la clase persona (*foaf:Person*) dentro de los repositorios. Esta información se consigue ejecutando el código SPARQL presentado en el segmento de código 1.

```
SELECT ?uri ?fname ?lname {
    ?uri a <http://xmlns.com/foaf/0.1/Person>.
    ?uri <http://xmlns.com/foaf/0.1/fistName> ?fname.
    ?uri <http://xmlns.com/foaf/0.1/lastName> ?lname.
}
```

Listing 1. Consulta para selección de autores.

[5]http://xmlns.com/foaf/spec/
[6]http://bibliontology.com/

Esta consulta proporciona los nombres registrados en el repositorio y una URI que sirve de identificador del recurso. En la tabla I se presenta un extracto de los resultados obtenidos sobre repositorio reales.

Cuadro I
AUTORES DEL REPOSITORIO

URI	Nombre	Repositorio
SCOPUS:author/author_id/ 57193429229	M.Espinoza	Scopus
CEDIA:contribuyente/ESPINOZA__MAURICIO	Espinoza, Mauricio	CEDIA
UDC:contribuyente/ESPINOZA_MEJIA__JORGE_MAURICIO	Espinoza Mejía, Jorge Mauricio	Universidad de Cuenca

Por otro lado, los metadatos de los documentos asociados a cada uno de los autores y la información de coautores se obtienen mediante la consulta SPARQL presentada en el segmento de código 2. Esta consulta debe ejecutarse para cada uno de los elementos de la lista de autores. En la tabla II se presenta un ejemplo de los resultados obtenidos para el recurso "Espinoza, Mauricio" del repositorio Universidad de Cuenca.

```
SELECT ?af ?title ?abtract ?subject ?cafn ?caln {
    <%AuthURI%> <http://xmlns.com/foaf/0.1/Organization> ?af.
    ?d <http://purl.org/dc/terms/creator> <%AuthorURI%>.
    ?d <http://purl.org/ontology/bibo/abstract> ?abstract.
    ?d <http://purl.org/dc/terms/title> ?title.
    ?d <http://purl.org/dc/terms/subject> ?subject.
    ?coauthoruri <http://purl.org/dc/terms/creator> ?d.
    ?coauthoruri <http://xmlns.com/foaf/0.1/firstName> ?cafn.
    ?coauthoruri <http://xmlns.com/foaf/0.1/lastName> ?caln.
    FILTER (str(?coauthoruri) != '%AuthURI%').
}
```

Listing 2. Consulta para selección de coautores.

Cuadro II
COAUTORES

Título	*Abstract*	*Subjects*	Coautores
RDF-ization of DICOM medical images towards linked health ...	This paper proposes a novel strategy for semantifying DICOM...	LINKED HEALTH DATA, SEMANTIC WEB...	Andrés Tello, Saquicela Víctor, ...
Plataforma para la búsqueda por contenido visual y semántico de ...	Este trabajo describe una plataforma que permite automatizar...	ONTOLOGÍAS MÉDICAS, SEGMENTACIÓN, ..	Lizandro Solano, Patricia Gonzalez, Andres Tello, ...

Como resultado de esta etapa se obtiene una lista de todos los autores disponibles en todos los repositorios digitales analizados. Cada autor además posee información de su contexto dentro del repositorio, que consiste en sus coautores, metadatos de sus documentos (título, *abstract*, *subjects*) y su afiliación. Si bien esta información sirve para caracterizar inicialmente a los autores, es necesaria una etapa adicional de pre procesamiento de esta información que permitirá disminuir los datos disponibles, tomando una muestra más pequeña y

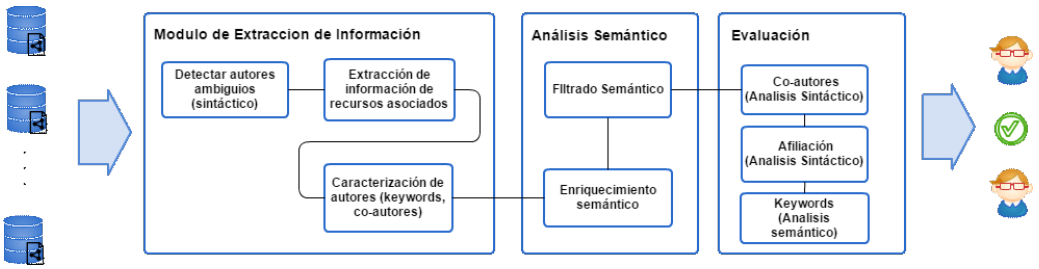

Figura 2. Etapas de desambiguación de autores

representativa de las características del autor. En la siguiente sección se describe el proceso realizado mediante el análisis y filtrado semántico.

III-C. Análisis semántico

El objetivo de esta etapa consiste en emplear tecnologías semánticas para mejorar la descripción de los datos que caracterizan a un autor, para esto se dispone de dos etapas: En la primera se extrae información a partir de los metadatos de los documentos (título, *abstract*, *subjects*) mediante el reconocimiento de entidades y usando una base de conocimiento como Dbpedia. En segundo lugar se filtran los conceptos o entidades con el objetivo de conservar únicamente aquellos términos que representen mejor el área de trabajo del autor. Mediante este procedimiento se busca que un autor sea representado a partir de un número reducido y representativo de las palabras clave obtenidas a partir de sus documentos.

III-C1. Detección de entidades: Para convertir la descripción de los metadatos de un documento tales como título, palabras clave y *abstract* en una cantidad manejable y representativa de información, se emplearon técnicas de minería de textos para la identificación de entidades (*Named Entity Recognition* - NER)[20]. Las técnicas NER permiten reconocer diferentes tipos de entidades como: localizaciones, personas y conceptos que son referenciadas dentro de un documento o segmento de texto. En este caso en particular se utilizó la herramienta Dbpedia Spotlight[7] para la aplicación de esta técnica sobre los metadatos de los documentos, la cual emplea una extensa base de conocimiento como es Dbpedia[8] [21] para la detección de entidades. Mediante la aplicación de esta herramienta se puede descubrir una diversa variedad de entidades dentro de los textos creados por un autor, esta entidades están modelados como conceptos dentro de los vocabularios ontológicos.

En la tabla III se presenta un ejemplo de las entidades descubiertas usando Dbpedia Spotlight para el documento "Plataforma para la búsqueda por contenido visual y semántico de imágenes médicas", del autor "Mauricio Espinoza". Los documentos en español (como es este caso) son traducidos al inglés usando un Servicio Web de traducción antes de ser analizados con Spotlight. Esto por cuanto el desarrollo de las técnicas NER y de la base de conocimiento (Dbpedia) en sí tienen un mayor desempeño en su versión en inglés con respecto a otros idiomas y por tanto se obtienen mejores resultados.

[7]dbpedia-spotlight.org
[8]http://dbpedia.org/

Cuadro III
ENTIDADES DESCUBIERTAS

Segmento de texto	Concepto detectado
semantic	http://dbpedia.org/resource/Semantic_Web
DICOM	http://dbpedia.org/resource/DICOM
medical imaging	http://dbpedia.org/resource/Medical_imaging
ontologies	http://dbpedia.org/resource/Ontology_(information_science)

III-C2. Filtrado semántico: En la mayoría de documentos las palabras clave tomadas de los metadatos y las entidades extraídas a partir de los *abstract* representan las áreas de interés de un autor; sin embargo, existen otros casos donde más bien pueden llegar a producir errores e inconsistencias. Por ejemplo, muchas de las palabras clave ingresadas en los metadatos de documentos incluyen referencias a localizaciones e instituciones como "Provincia del Azuay" u "Hospital Regional Vicente Corral", etc. También es común que se incluyan categorizaciones propias de la universidad como: Tesis de pregrado, Tesis de maestría, etc. Estas referencias no ayudan a distinguir entre autores, sino que al contrario pueden introducir ruido al proceso de comparación. Por otro lado, las entidades reconocidas mediante NER también son susceptibles a errores, en especial cuando los textos son cortos. Un ejemplo de ambigüedad introducido por el proceso NER es la definición de las siglas, así en ciertos casos "NGD" que puede tomar el significado de "*Normalized Google Distance*" (Contexto informático) cuando en realidad puede referirse a "*Non-Good Delivery*" (Contexto de manipulación de barras de oro). Es por todo esto, que se implementó una actividad de filtrado semántico de las palabras clave, que está pensada en mejorar la calidad de las palabras clave que representan un autor.

La primera parte del proceso de filtrado utiliza una lista de palabras vacías (*stopwords*), la cual se creó tras un análisis de las palabras clave usadas en los metadatos de los documentos. Esta lista identifica términos comunes para referirse a localizaciones e instituciones como: "Cantón", "Provincia", "Hospital", etc. Cuando uno de estos términos es encontrado, se desecha toda la palabra clave del proceso de desambiguación. Por ejemplo la palabra clave "Cantón Cuenca - Azuay" es

ignorada puesto que contiene la palabra vacía "Cantón". Adicionalmente, los entidades descubiertas con Dbpedia Spotlight que se identifiquen como una localización geográfica (clase *Dbpedia:Place*) también son ignoradas, lo que se consigue consultado el *SPARQL Endpoint* de Dbpedia. Las localizaciones son evitadas en el proceso de desambigüación debido a que pueden causar problemas principalmente entre autores ambigüos que comparten una localización (ciudad, País, etc.). Considerando las localizaciones en muchos casos se asociaban dos autores no por su área de trabajo sino por la región en la que realizaban sus trabajos. En resumen, esta primera fase de filtrado permite eliminar referencias inútiles del proceso de desambiguación y que de lo contrario pueden introducir ruido al proceso.

La segunda y última parte del proceso de filtrado semántico consiste en eliminar las palabras clave que tengan menor relevancia semántica para un autor. Esto se consigue evaluando la similitud semántica entre cada una de las palabras clave con respecto a las demás del conjunto, permitiendo determinar qué tan relacionadas están las palabras clave entre sí. Aquellas palabras que presenten menor relación semántica con respecto a las demás serán consideradas como ruido, con lo cual se eliminan los posibles conceptos detectados de forma errónea o palabras clave que no aporten a la identificación del área de trabajo de un autor.

Para la implementación de este filtrado se empleó la medida de relación semántica NWD (*Normalized Wikipedia Distance*) [22], debido a que presentó mejores resultados en el proceso de comparación semánticamente de palabras clave. NWD es una métrica simple que evalúa la distancia semántica entre dos cadenas de texto, mediante operaciones de búsqueda (*Full-text*). NWD es una adaptación de NGD (*Normalized Google Distance*) [23] que opera sobre Wikipedia como base de conocimiento en lugar del motor de búsqueda de Google. La métrica NWD ofrece una gran flexibilidad, puesto que no requiere de vocabularios fijos ni información previamente estructurada para su utilización como lo hacen la mayoría de métricas disponibles en el estado del arte. La fórmula 1 es utilizada para evaluar NWD, la misma que fue implementada usando la API de búsqueda de Wikipedia[9].

$$NWD(x,y) = \frac{\max(\log f(x), log f(y)) - \log f(x,y)}{\log N - \min(\log f(x), \log f(y))} \quad (1)$$

donde, $f(t)$: Número de artículos que contienen el término t, N: Número total de artículos de Wikipedia, $f(t_1, t_2)$: Número de artículos que contiene los términos t_1 y t_2 al mismo tiempo.

La evaluación de la similitud entre las palabras clave se define mediante la sumatoria de la distancia de una palabra clave con respecto a las demás palabras del conjunto. Es decir, si un autor posee las palabras clave p_1, p_2,..., p_n la relevancia $r(i)$ de cada palabra clave se estima mediante la fórmula 2. Nótese que al tratarse de distancias semánticas las palabras clave que obtengan menor valor $r(i)$ se consideran más relevantes para un autor.

[9]https://www.mediawiki.org/wiki/API:Search

$$r(i) = \sum_{j=0}^{j<n} NWD(p_i, p_j), j \neq i \quad (2)$$

El criterio seguido para definir el número de palabras clave que deban sobrepasar el filtro se definió mediante una regla práctica basada en las observaciones realizadas sobre los datos. Se definió que para un conjunto N de palabras clave se debería seleccionar $\lfloor 2{,}5 * ln(N) \rfloor$ palabras más relevantes para ser usadas por la siguiente etapa. Esta regla ofrece un crecimiento amortiguado del número de palabras clave a ser usados, de manera que autores con pocas palabras clave no las pierdan debido al filtrado semántico y al mismo tiempo que autores con demasiadas palabras clave limiten el número de estas.

En la figura 3 se presenta gráficamente el filtrado de las palabras clave de el autor "Mauricio Espinoza" del repositorio de la Universidad de Cuenca. En esta figura se muestra como las palabras clave más relacionadas (*semantic web*, *Open Data*, ..) con su área de trabajo se agrupan en el centro y las menos relacionadas (*Software GIS*, *big data*, ..) son excluidas usando el umbral establecido.

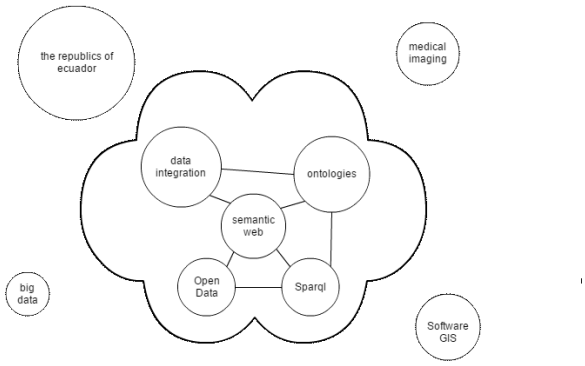

Figura 3. Filtrado semántico de palabras clave

III-D. Evaluación

Como se presenta en la figura 4, la comparación final entre los autores y su desambiguación se realiza en dos etapas. En la primera etapa, se determinan autores candidatos (usando un método de *blocking* [24]) que podrían tratarse del mismo individuo a través de la detección de nombres similares. En la segunda etapa, se realiza una comparación semántica entre los candidatos con la información obtenida luego del proceso de análisis semántico y usando las características obtenidas: afiliación, coautores y palabras clave. Finalmente, los candidatos que se encuentren dentro del umbral de distancia semántica establecido son considerados equivalentes y se crean enlaces para estos. A continuación se detallan estas dos etapas y cada una de las métricas que utilizan.

III-D1. Autores candidatos: Existe una gran variedad de algoritmos para evaluar la similitud sintáctica entre dos cadenas de texto, sin embargo, los algoritmos basados en *tokens* son los más utilizados para la detección de nombres similares. Estos algoritmos representan los nombres como conjuntos de

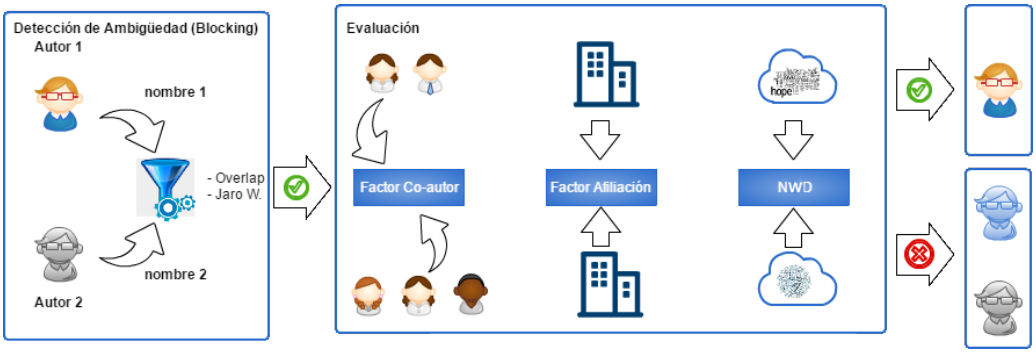

Figura 4. Comparación entre autores

palabras (*tokens*) y aplican operaciones de conjuntos para determinar su similitud. Ejemplos típicos de estos algoritmos usados en la comparación de nombres son Jaccard y Overlap, tal como se trata en [11]. No obstante, para cubrir de manera más general este problema es necesario agregar más características a estas métricas con el fin de hacerlas más flexibles a las particularidades de los nombres de personas como: iniciales, abreviaturas y errores de escritura [17].

En el presente trabajo se plantea la utilización de una métrica híbrida que utiliza los enfoques de *tokens*, similitud sintáctica de texto, iniciales y abreviaturas. Específicamente se propone usar una versión adaptada de *overlap* para la comparación de nombres, la cual se complementa con la métrica de "Jaro-Winkler" para la generación de *matchs* flexibles entre *tokens*. Este propuesta es una adaptación del trabajo presentado en [17], que incorpora una métrica de similitud sintáctica a la comparación de los nombres. En la fórmula 3 se presenta esta métrica de forma más formal.

$$NameSim(N_1, N_2) = \frac{MJW(N_1, N_2) + p * MAI(N_1, N_2)}{min(NT(N_1), NT(N_2))} \quad (3)$$

donde, MJW: Número *matchs* entre las palabras (*tokens*) de los nombres N_1 y N_2 usando la métrica Jaro-Winkler con un umbral de $0{,}95$.
MAI: Número de *matchs* entre las palabras (*tokens*) de los nombres N_1 y N_2, tomando iniciales y abreviaturas (ignorando las palabras usadas en los *matchs* de MJW).
$NT(N)$: Número total de *tokens* en N.
p: Penalización para *matchs* con iniciales y abreviaturas ($p = 0{,}95$).

Para la detección de la lista de autores candidatos a ser desambiguados se aplica esta métrica. Los pares de autores que sobrepasen un umbral de similitud del nombre de $0{,}9$ se agregan a la lista de candidatos a desambiguar. En la tabla IV se presenta un extracto de la lista de nombres de autores comparados con 'Mauricio Espinoza'. Esta lista presenta los casos más comunes encontrados en la comparación de nombres como: utilización de iniciales, errores de escritura y nombres incompletos.

III-D2. Comparación semántica: La comparación de candidatos se realiza en tres partes: afiliación, coautores y palabras clave. Para cada una se han definido métricas que aportan conocimiento de la relación entre los candidatos. Al finalizar la evaluación de cada una estas se genera un índice de similitud, el cual se utiliza para determinar si se trata o no del mismo autor. A continuación se explica cada una de las partes de esta comparación.

Cuadro IV
AUTORES CANDIDATOS

Nombre Fuente 1	Nombre Fuente 2	Similitud
Mauricio Espinoza	Mauricio Espinoza B	1
Mauricio Espinoza	Andrés Espinoza	0.5
Mauricio Espinoza	Jorge Mauricio Espinoza Mejía	1
Mauricio Espinoza	M Espinoza	0.95
Mauricio Espinoza	Mauricio Espinosa	0.949

El factor de afiliación FA se define como: si dos autores candidatos comparten una misma afiliación (han publicado bajo la misma institución) el factor toma el valor de $0{,}9$, caso contrario se asigna el valor de 1. Esta condición se basa en la suposición que si encontramos autores candidatos con nombres parecidos y que trabajen para la misma institución es más probable que se traten de la misma persona.

El factor de coautores FC por su parte se activa cuando el nombre de al menos un coautor de los dos candidatos es compartido, en este caso FC pasa a ser $0{,}8$, caso contrario se mantiene en 1. Esta regla se basa en el principio que si dos candidatos comparten coautores es muy probable que se trate del mismo individuo. Hay que destacar que para evaluar la similitud de los nombre de los coautores se utiliza la misma métrica definida para descubrir a los autores candidatos como se explica en la subsección anterior. Este tipo de suposiciones es muy común en los algoritmos de desambiguación usados en bases digitales porque mejora notablemente la precisión.

En el presente trabajo se propone estimar la distancia semántica (DS) entre dos autores candidatos comparando sus palabras clave a través de NWD. Para esto ha definido como métrica el promedio de las distancias semánticas entre las palabras clave de los autores. Este índice pretende reflejar cual es la distancia semántica entre los temas de interés de los candidatos. Donde una distancia menor significa que tratan temáticas parecidas y una distancia mayor implica temáticas

distintas. En la fórmula 4 se presenta formalmente la definición de la distancia entre dos autores candidatos con conjuntos de palabras clave A y B respectivamente.

$$DS(A,B) = \frac{\sum_{i}^{i<N(A)} \sum_{j}^{j<N(B)} NWD(A_i, B_j)}{N(A) * N(B)} \quad (4)$$

$N(X)$: Número de palabras clave del conjunto X.
X_i: La i-ésima palabra clave del conjunto X.

Finalmente, la distancia total DT entre dos autores candidatos se define como $DT = FA * FC * DS$. Este valor resume toda la información de dos autores y la cercanía semántica entre sí. A partir de este índice se aplica un filtrado simple con un umbral de 0,7 (obtenido experimentalmente). De manera que todos los pares de autores candidatos con un valor DT menor a 0,7 son considerados el mismo individuo y por tanto enlazados. Para esto se recomienda la utilización del vocabulario ontológico OWL[10] que define equivalencias entre dos recursos a través del concepto *owl:sameAs*. Estos enlaces deben ser registrados en los repositorios de datos enlazados para su posterior utilización y pos procesamiento.

IV. Aplicación y resultados

La propuesta de desambiguación expuesta en el presente trabajo ha sido probada exitosamente en el contexto académico ecuatoriano. Específicamente se ha trabajado en dos ámbitos: desambiguación de autores dentro de los repositorios institucionales de las universidades ecuatorianas e identificación de investigadores ecuatorianos sobre bases de datos digitales externas. La naturaleza de estas actividades ha permitido probar la aplicabilidad de la propuesta sobre varias fuentes de datos tanto internas (Ecuador) como externas.

El primer escenario de aplicación se elaboró en el contexto de la integración de los repositorios digitales del Ecuador [11]. Donde se integró un conjunto de veinte y un repositorios institucionales de las universidades del Ecuador. El proceso de desambiguación se aplicó sobre aproximadamente sobre 145000 registros de estudiantes, docentes e investigadores obteniendo 1960 enlaces entre autores ambiguos. De esta forma se mejoró la calidad de la información contenida dentro de los repositorios desde una perspectiva interna. Por otro lado, un escenario de aplicación externo se implementó en el proyecto "Repositorio ecuatoriano de investigadores" (REDI)[12]. Donde se identificaron a los investigadores ecuatorianos y se desambiguaron sus perfiles dentro de las bases digitales externas: Scopus, Microsoft Academic y Google Schoolar.

En la tabla V se presenta un extracto de los enlaces descubiertos a través de la aplicación de la propuesta presentada. La información corresponde tanto a la desambiguación de autores interna (repositorios del Ecuador) como los enlaces encontrados a fuentes externas. Específicamente se presenta los resultados del ejemplo descrito en este documento, en este caso del autor Mauricio Espinoza.

[10]https://www.w3.org/OWL/
[11]http://fedquest.cedia.org.ec/
[12]redi.cedia.org.ec

La tabla presenta pares de autores que han sido considerados como equivalentes mediante el proceso de desambigüación presentado en este trabajo. La información que se presenta consta de la fuente de la información, un identificador del recurso (URI) y el nombre del autor como se registra en la fuente (Nombre). Nótese que los resultados presentados abordan cuatro casos de ambigüedad presentes en las fuentes. Primero, autores ambiguos dentro de una misma fuente (fila 1). Segundo, autores ambiguos entre fuentes internas (Repositorios ecuatorianos) en la fila 7. Tercero, autores ambiguos entre fuentes internas y externas, presentes en las filas 2, 3, 4, 5 y 6. Finalmente, autores ambiguos entre fuentes externas (fila 8).

V. Conclusión y trabajos futuros

La desambiguación de autores es una problemática común en todos los sistemas de información bibliográfica y que por su importancia ha recibido especial atención de la comunidad de investigadores. Sin embargo, la mayor parte de los esfuerzos investigativos se han centrado en solucionar los problemas de ambigüedad en bases digitales de producción científica dejando en segundo plano a otros sistemas como los repositorios digitales. En este contexto surge la necesidad de atraer la investigación a nuevas fuentes de información, adaptando las técnicas existentes de desambiguación a este nuevo entorno.

En este trabajo se presenta un nuevo proceso de desambiguación semántica de autores que busca abordar este problema de forma más integral, considerando la heterogeneidad de las fuentes de información. Proceso que trasparenta las particularidades de los sistemas de información y adapta técnicas semánticas de vanguardia como: reconocimiento de entidades, métricas de similitud semántica y bases de conocimiento ontológicos. Adicionalmente, el enfoque de desambiguación presentado se enmarca en los principios de datos enlazados y Web Semántica que amplía su ámbito de aplicación a la Web.

Hay que destacar que los resultados obtenidos están limitados a ciertas suposiciones, que podrían afectar al rendimiento del algoritmo, tales como: las publicaciones o recursos bibliográficos están correctamente asignadas a sus autores, por lo que no se considera el problema de reasignación de obras. También, se considera, que es poco probable que autores con nombres similares trabajen en temáticas similares.

El trabajo futuro se centrará en el mejoramiento del proceso de desambiguación planteado mediante la explotación de estructuras ontológicas que utilizan las bases de conocimiento (jerarquías, clasificaciones de conceptos, etc). Estas mejoras estarán orientadas a cubrir nuevas y más complejas fuentes de información, así como refinar los resultados obtenidos. Finalmente, se propone expandir la utilidad de los algoritmos desarrollados en este trabajo para atacar otros problemas comunes en los sistemas bibliográficos y de manejo de autores como catalogación automática de documentos e identificación de redes de colaboración entre autores.

Cuadro V
DESAMBIGUACIÓN PARA EL AUTOR MAURICIO ESPINOZA

Fuente1	URI 1	Nombre 1	Fuente 2	URI 2	Nombre 2
U. Cuenca	UDC:/ESPINOZA_MEJIA __JORGE_ MAURICIO	Jorge Mauricio Espinoza Mejía	U. Cuenca	UDC:/contribuyente/ ESPINOZA__MAURICIO	Mauricio Espinoza
CEDIA	CEDIA:/contribuyente/ ESPINOZA__MAURICIO	Mauricio Espinoza	Scopus	Scopus:/author/author_id/ 57193429229	M. Espinoza
U. Cuenca	UDC:/ESPINOZA_MEJIA __JORGE_ MAURICIO	Jorge Mauricio Espinoza Mejía	Scopus	Scopus:/author/author_id/ 57193429229	M. Espinoza
U. Cuenca	UDC:/ ESPINOZA_MEJIA __JORGE_ MAURICIO	Jorge Mauricio Espinoza Mejía	GoogleScholar	GoogleScholar:/author/ Mauricio_Espinoza	Mauricio Espinoza Mejia
U. Cuenca	UDC:/ESPINOZA_MEJIA __JORGE_ MAURICIO	Jorge Mauricio Espinoza Mejía	Academics	Academic:/detail/ 2273716818	Mauricio Espinoza Mejia
CEDIA	CEDIA:/contribuyente/ ESPINOZA__MAURICIO	Mauricio Espinoza	GoogleScholar	GoogleScholar:/author/ Mauricio_Espinoza	Mauricio Espinoza Mejía
CEDIA	CEDIA:/contribuyente/ ESPINOZA__MAURICIO	Mauricio Espinoza	U. Cuenca	UDC:/ESPINOZA_MEJIA __JORGE_ MAURICIO	Jorge Mauricio Espinoza Mejía
GoogleScholar	GoogleScholar:/author/ Mauricio_Espinoza	Mauricio Espinoza Mejía	Academics	Academic:/detail/ 2273716818	Mauricio Espinoza Mejia

AGRADECIMIENTOS

Al Departamento de Ciencias de la Computación de la Universidad de Cuenca. Adicionalmente, al Consorcio Ecuatoriano para el Desarrollo de Internet Avanzado (RED-CEDIA), por el financiamiento brindado a esta investigación, mediante el proyecto "Repositorio Semántico de Investigadores del Ecuador" y al grupo de trabajo de Repositorios Digitales.

REFERENCIAS

[1] J. Segarra, J. Ortiz, M. Espinoza, and V. Saquicela, "Integration of digital repositories through federated queries using semantic technologies," in *Computing Conference (CLEI), 2016 XLII Latin American*. IEEE, 2016, pp. 1–9.

[2] X. Sumba, F. Sumba, A. Tello, F. Baculima, M. Espinoza, and V. Saquicela, "Detecting Similar Areas of Knowledge Using Semantic and Data Mining Technologies," *Electronic Notes in Theoretical Computer Science*, vol. 329, pp. 149–167, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1571066116301165

[3] N. R. Smalheiser and V. I. Torvik, "Author name disambiguation," *Annual Rev. Info. Sci & Technol.*, vol. 43, no. 1, pp. 1–43, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1002/aris.2009.1440430113

[4] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data – The Story So Far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[5] C. L. S. MLS, E. D. J. MLS, and A. L. M. MLS, "When A. Rose Is Not A. Rose," *Medical Reference Services Quarterly*, vol. 22, no. 4, pp. 1–11, 2003, pMID: 14711044. [Online]. Available: http://dx.doi.org/10.1300/J115v22n04_01

[6] D. A. Dervos, N. Samaras, G. Evangelidis, J. P. Hyvärinen, and Y. Asmanidis, "The universal author identifier system (uai_sys)," 2007.

[7] L. L. Haak, M. Fenner, L. D. Paglione, E. Pentz, and H. Ratner, "Orcid: a system to uniquely identify researchers," *Learned Publishing*, vol. 25, pp. 259–264, 2012.

[8] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender, "A brief survey of automatic methods for author name disambiguation." *SIGMOD Record*, vol. 41, no. 2, pp. 15–26, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/sigmod/sigmod41.html#FerreiraGL12

[9] T. Gurney, E. Horlings, and P. Van Den Besselaar, "Author disambiguation using multi-aspect similarity indicators," *Scientometrics*, vol. 91, no. 2, pp. 435–449, 2012.

[10] P. Treeratpituk and C. L. Giles, "Disambiguating authors in academic publications using random forests," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2009, pp. 39–48.

[11] T. Huynh, K. Hoang, T. Do, and D. Huynh, "Vietnamese author name disambiguation for integrating publications from heterogeneous sources," in *Asian Conference on Intelligent Information and Database Systems*. Springer, 2013, pp. 226–235.

[12] K.-H. Yang, H.-T. Peng, J.-Y. Jiang, H.-M. Lee, and J.-M. Ho, "Author name disambiguation for citations using topic and web correlation," *Research and advanced technology for digital libraries*, pp. 185–196, 2008.

[13] H. Han, W. Xu, H. Zha, and C. L. Giles, "A hierarchical naive bayes mixture model for name disambiguation in author citations," in *Proceedings of the 2005 ACM symposium on Applied computing*. ACM, 2005, pp. 1065–1069.

[14] J. Huang, S. Ertekin, and C. L. Giles, "Efficient name disambiguation for large-scale databases," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2006, pp. 536–544.

[15] C. L. Giles, H. Zha, and H. Han, "Name disambiguation in author citations using a k-way spectral clustering method," in *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*. IEEE, 2005, pp. 334–343.

[16] H. Han, C. Yao, Y. Fu, Y. Yu, Y. Zhang, and S. Xu, "Semantic fingerprints-based author name disambiguation in chinese documents," *Scientometrics*, pp. 1–18.

[17] M. Shoaib, A. Daud, and M. Khiyal, "Improving similarity measures for publications with special focus on author name disambiguation." *Arabian Journal for Science & Engineering (Springer Science & Business Media BV)*, vol. 40, no. 6, 2015.

[18] R. C. Bunescu and M. Pasca, "Using encyclopedic knowledge for named entity disambiguation." in *Eacl*, vol. 6, 2006, pp. 9–16.

[19] R. Bekkerman and A. McCallum, "Disambiguating web appearances of people in a social network," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 463–470.

[20] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, January 2007, publisher: John Benjamins Publishing Company. [Online]. Available: http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002

[21] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia spotlight: shedding light on the web of documents," in *Proceedings of the 7th international conference on semantic systems*. ACM, 2011, pp. 1–8.

[22] C. Schaefer, D. Hienert, and T. Gottron, "Normalized relevance distance-a stable metric for computing semantic relatedness over reference corpora," in *Proceedings of the Twenty-first European Conference on Artificial Intelligence*. IOS Press, 2014, pp. 789–794.

[23] R. Cilibrasi and P. M. B. Vitányi, "The google similarity distance," *CoRR*, vol. abs/cs/0412098, 2004. [Online]. Available: http://arxiv.org/abs/cs/0412098

[24] M. Bilenko, "Adaptive blocking: Learning to scale up record linkage," in *In Proceedings of the 6th IEEE International Conference on Data Mining (ICDM-2006*, 2006, pp. 87–96.