# Desambiguación del nombre de los autores en metadatos bibliográficos publicados como datos enlazados

Luis Enrique Alonso-Sierra, Yusniel Hidalgo-Delgado

Grupo de Web Semántica, Universidad de las Ciencias Informáticas, La Habana, Cuba
lealonso24@gmail.com, yhdelgado@uci.cu

**Resumen** El problema de la ambigüedad del nombre de los autores consiste en la posibilidad real de encontrar nombres escritos sintácticamente diferentes que hacen referencia a un mismo autor en una colección de metadatos bibliográficos. Esto provoca que dichos nombres cuando son publicados siguiendo los principios de los datos enlazados sean repetidos en un mismo grafo RDF. En este artículo se presenta una aproximación para dar respuesta al problema planteado basada en técnicas de agrupamiento. La propuesta utiliza elementos relacionados con los autores que están disponibles en las publicaciones científicas, tales como: coautores, afiliación, lugar de publicación y títulos de las publicaciones. Aunque la propuesta se encuentra en fase de desarrollo, se han obtenido algunos resultados alentadores. Los resultados son contrastados con otras aproximaciones similares existentes en la literatura.

**Palabras Claves:** Desambiguación, Datos Enlazados, Aprendizaje no Supervisado, Metadatos Bibliográficos

**Abstract** The problem of the ambiguity of the names of the authors is the real chance to find names written syntactically different but that refer to the same author in a collection of bibliographic metadata. Due to this, names of authors published following the linked data principles are repeated in a same RDF graph. This paper presents an approach to address this problem based in clustering techniques. The proposal uses available elements related to authors in scientific papers, such as: co-authors, affiliation, metadata sources and titles of papers. Although the proposal is under development, there have been some encouraging results. Results are contrasted with existing similar approaches in the literature.

**Keywords:** Disambiguation, Linked Data, Unsupervised Learning, Bibliographics Metadatas

## 1. Introducción

La ambigüedad en el nombre de los autores puede ser causada por la falta de un estándar único en la introducción de los metadatos bibliográficos, errores de

escritura en el registro de los nombres de los autores, entre otros. Dicho problema se manifiesta de dos formas, (1) nombres de autores diferentes sintácticamente que se refieren a la misma persona, (2) nombres sintácticamente iguales que se refieren a personas diferentes.

Un elemento a tener en cuenta con respecto al problema antes mencionado es el idioma en que están escritos los nombres de los autores. En idiomas como el Inglés, Francés y Alemán, los autores se identifican por un nombre y un apellido, lo que limita las variantes con que se pueden representar los mismos. Al contrario, en el idioma Español los autores son representados con un nombre (en ocasiones dos) y dos apellidos, lo cual aumenta las posibilidades de aparición de diferentes representaciones. También influye, en este sentido, la aparición de errores de escritura.

Actualmente la Web Semántica ha cobrado auge en diversos dominios de aplicación. Los Datos Enlazados[1] constituyen la base sobre la que se soporta su desarrollo. Estos se refieren a un conjunto de buenas prácticas para la publicación y enlazado de datos estructurados en la web. Para dicha publicación y enlazado se definen varios principios, los cuales es necesario cumplir:

- **Identificar**: Utilizar URI[1] para identificar cada recurso en la web.
- **Publicar**: Publicar cada recurso en una URI basada en HTTP[2] de modo que puedan ser fácilmente localizados y consultados.
- **Describir**: Proporcionar información detallada, útil o extra acerca de cada recurso publicado en la web.
- **Enlazar**: Enlazar los recursos publicados con otras URIs relacionadas, de forma que se potencie el descubrimiento de la información sobre la web.

El problema de la ambigüedad en el nombre de los autores introduce ruido en el proceso de publicación de los metadatos como datos enlazados. Publicar un autor (recurso) con ambigüedad en su representación implica publicar el mismo recurso varias veces, lo que trae consigo pérdida en la calidad de los recursos publicados en la web de los datos. Esto provoca que se afecte el rendimiento del entrelazado de los recursos en el proceso de publicación así como su descubrimiento por otros sistemas informáticos. También la localización de los recurso en la web se afecta debido a que un mismo recurso está publicado en diferentes URLs[3] y las relaciones que se establecen entre los autores (recursos) y otros recursos (Artículos, Webs, entre otros) pueden ser excluyentes.

Existen soluciones que se han desarrollado sobre las tecnologías de los datos enlazados utilizando los metadatos bibliográficos como fuente de información. DBLP[4] es uno de los sistemas desarrollados bajo dichos principios. El mismo presenta el problema de la ambigüedad en el nombre de los autores. DBLP provee un Endpoint SPARQL[2] para la consulta de los datos existentes en el sistema. Cuando intentamos obtener información relacionada con el autor cuyo nombre es **Yaima Filiberto** se realiza una consulta SPARQL como se muestra:

[1] *Identificador Universal de Recursos*
[2] *Lenguaje de Marcado de Hipertexto*
[3] *Localizador Universal de Recurso*
[4] *http://www.dblp.org*

```
1: SELECT ?o
2: WHERE {?s akt : full − name ?o.
3: FILTER regex( ?o , "Yaima Filiberto")}
```

El resultado de la consulta anterior se muestra a continuación:

**Results:**

| Result | Binding | Value |
|---|---|---|
| 1 | ?o | Yaima Filiberto |
| 2 | ?o | Yaima Filiberto Cabrera |

**Figura 1.** Resultados de la consulta SPARQL buscando a **Yaima Filiberto**

Lo anterior demuestra que los datos utilizados por DBLP poseen problemas en la representación de los nombres de los autores. Otros ejemplo que demuestra el planteamiento anterior es el nombre **Rafael Bello**. Si ejecutamos una consulta similar en el Endpoint SPARQL de DBLP con dicho nombre se obtienen resultados similares.

| Result | Binding | Value |
|---|---|---|
| 1 | ?o | Rafael Bello |
| 2 | ?o | Rafael Bello Pérez |
| 3 | ?o | Rafael Bello |

**Figura 2.** Resultados de la consulta SPARQL buscando a **Rafael Bello**

DBLP es uno de los sistemas fundamentales en el desarrollo de la Web Semántica y los Datos Enlazados. Como se puede apreciar no son pocos los nombres de los autores que poseen ambigüedad en su representación incluso en sistemas de reconocimiento mundial como DBLP. Dicho problema ha propiciado que los datos manejados por DBLP se utilicen en la valiadación de aproximaciones propuestas en la literatura para solucionar el problema de la ambigüedad en el nombre de los autores[3][4].

En el presente artículo se propone un método para la desambiguación del nombre de los autores utilizando técnicas de agrupamiento y procesamiento del

lenguaje natural. La solución está compuesta por dos procesos fundamentales, (1) pre-procesamiento de la información y generación de un vector de similitudes, luego, (2) utilizando la combinación de agrupamientos[5] se realiza el proceso de desambiguación.

La generación de los vectores de similitud está caracterizada por la utilización de la distancia de edición, esta permite cuantificar la similitud de dos cadenas de caracteres. Dicha distancia se refiere al conjunto de operaciones básicas para convertir una cadena de caracteres en otra, siendo las operaciones básicas la eliminación, adición, permutación y cambio de caracteres[6]. La combinación de agrupamientos utiliza los resultados de varios algoritmos de agrupamientos, para obtener un mejor resultado, combinando las salidas de dichos algoritmos de agrupamiento.

Este artículo está estructurado de la siguiente manera: en la sección 2 se mencionan algunos trabajos relacionados con la propuesta de solución. En la sección 3 se describe el proceso de pre-procesamiento y generación de los vectores de similitud y se presenta el modelo seleccionado para la combinación de agrupamiento. Finalmente, en la sección 4 se muestran las conclusiones del trabajo.

## 2. Trabajos relacionados

El problema de la desambiguación del nombre de los autores en metadatos bibliográficos ha sido tratado de diversas formas. En[7] y [3] se modela el problema de forma probabilística, donde se calcula la probabilidad de que dos nombres de autores se refieran a la misma persona. Este tipo de solución tiene en cuenta elementos disponibles en los metadatos bibliográficos (co-autores, la afiliación, los lugares de publicación y los nombres de los autores). Para la modelación del problema de forma probabilística es necesario conocer el dominio de las variables del problema y el comportamiento de los datos.

Por otro lado, [8] y [9] plantean el problema como un modelo de clasificación supervisada, donde se verifica si dos autores se refieren a la misma persona. En este sentido, es necesario poseer datos que puedan ser utilizados para entrenar el modelo obtenido, situación poco aplicable en escenarios reales.

También[10] y [11] proponen solucionar el problema utilizando técnicas de agrupamiento. A diferencia de las anteriores, no se necesita conocer el comportamiento de los datos. La principal limitante de este tipo de aproximación es la estimación correcta del número de autores que están presentes en los datos analizados.

Entre las limitaciones de las aproximaciones existentes en la literatura se pueden mencionar: (1) ninguna está orientada a las particularidades del idioma español y (2) no se realiza un pre-procesamiento de la información con el objetivo de eliminar datos con ruidos e inconsistentes.

# 3. Propuesta de solución

La propuesta de solución comprende dos procesos fundamentales, (1) identificación de las relaciones existentes entre los nombres de los autores aplicando una distancia de edición y (2) representación de las relaciones identificadas en un vector de similitudes. Dicho vector está compuesto por la similitud que existe entre los elementos del contexto de los autores (co-autores, afiliación, lugares de publicación y títulos de las publicaciones). Seguidamente, utilizando dichos vectores se realiza el proceso de combinación de agrupamientos.

## 3.1. Representación de los vectores de similitud

La calidad de los metadatos es un elemento importante en el problema tratado. Realizar tareas de pre-procesamiento elevaría la calidad de los mismos. Las tareas llevadas a cabo en este sentido fueron: (1) conversión de todos los elementos que representaban cadenas de caracteres a minúscula, (2) eliminación de las tildes de las palabras que las presentaban y (3) eliminación de caracteres extraños.

## 3.2. Homogenización de las afiliaciones

Otro proceso realizado en el pre-procesamiento de la información fue la homogenización de las afiliaciones de los autores debido a que es uno de los elementos de mayor peso en el proceso de desambiguación[12]. Por ejemplo, las afiliaciones *"Universidad de las Ciencias Informáticas"* y *"Facultad 3, Universidad de las Ciencias Informáticas"* sintácticamente representan dos instituciones diferentes, cuando en realidad se refieren a una sola. Dicha homogenización se realizó utilizando la distancia de edición. Tomando como referencia el valor de dicha distancia se agruparon todas las afiliaciones cuya distancia fuera inferior a un determinado valor (umbral) que permitiera considerar que dichas afiliaciones representan variantes de la misma institución. El valor del umbral puede ser calculado de dos formas: (1) a través de un umbral absoluto y (2) a través de un umbral relativo. Debido a que la utilización de un umbral absoluto no es eficaz en la formación de grupos, se propone un umbral relativo.

$$\beta = \alpha * \text{mín}(|A|, |B|) \tag{1}$$

Donde A y B representan la cantidad de palabras de las afiliaciones.

Para comparar las afiliaciones estas se tomaron como una lista de palabras y se eliminaron los puntos. En muchas ocasiones las afiliaciones no tienen las palabras en un mismo orden y la similitud que resultaría de aplicar la distancia de edición sería mayor que la que resultaría intuitivamente. Seguidamente, se determinan qué palabras son lo suficientemente parecidas, limitando el valor de la distancia de edición a un valor menor o igual a 1, lo cual condiciona que los errores que puedan aparecer sean solamente errores de escritura. Luego se calcula la razón que existe entre las palabras que coinciden en las dos afiliaciones

y todas las palabras. Finalmente se compara el valor calculado con el umbral, si es mayor, entonces las afiliaciones comparadas son agrupadas, luego los autores determinan si las afiliaciones agrupadas en un mismo conjunto se refieren a la misma afiliación.

### 3.3. Relación entre los nombres de los autores

Uno de los elementos de importancia en la desambiguación es la similitud que existe entre los nombres de los autores[12]. En la propuesta de solución se desarrolló una función de similitud para comparar dichos elementos. La misma tiene como principal componente la distancia de edición y su objetivo principal es establecer relaciones entre los nombres de los autores, además, cuantificar dichas relaciones.

**3.3.1. Función de similitud** La función de similitud se define como un conjunto de comparaciones entre las sub-cadenas que componen los nombres de los autores. Dichas comparaciones permiten calcular la distancia de edición entre las mismas tolerando también errores de escritura. Para sub-cadenas de longitud siete se tolera un error de escritura y para sub-cadenas de longitud catorce se toleran dos errores de escritura. A continuación se muestra una tabla con el sistema de puntuación empleado para realizar la comparación entre las sub-cadenas:

| **Valor devuelto** | **Interpretación** |
|---|---|
| 0 | La coincidencia es total |
| 1 | Hay un cambio de edición entre las dos sub-cadenas |
| 2 | Hay dos cambios de edición entre las sub-cadenas y la menor de ellas tiene un longitud menor que catorce |
| $\infty$ | La disimilitud es demasiado grande |

**Cuadro 1.** Sistema de puntuación para la comparación de sub-cadenas

Otro elemento que se tuvo en cuenta en la función de similitud entre los nombres fueron las partículas de los mismos. Las partículas se refieren a palabras cortas que no representan nombres: *del, la, el, entre otras.* Estas palabras fueron eliminadas en la comparación de los nombres.

También se detectaron y procesaron los apellidos compuestos. En el contexto del artículo los apellidos compuestos se refieren a errores en los mismos, por ejemplo: **Pedro GarcíaSierra**. Para solucionar este problema se llevó a cabo el siguiente proceso:

Teniendo, por un lado, una palabra **a** que forma parte de las palabras del nombre y, por otro lado, un conjunto de palabras **C** cuya unión puede haber formado la primera. De ser cierto lo anterior, podemos afirmar que $\mathbf{a} \in \mathbf{C}$, por tanto se concluye que probablemente el conjunto de palabras **C** ha sido

compuesta por la unión de dos elementos del nombre (por ejemplo, la unión de los dos apellidos). A continuación se muestra una tabla con el sistema de puntuación empleado para cuantificar la identificación de apellidos compuestos:

| **Valor devuelto** | **Interpretación** |
|---|---|
| 1-4 | Es un posible apellido compuesto pero se habrían producido errores en su escritura |
| 0 | Tanto el comienzo como el final del apellido del autor se corresponden con palabras del nombre del otro autor |
| $\infty$ | No existen las suficientes coincidencias |

**Cuadro 2.** Sistema de puntuación para la identificación de los apellidos compuestos

La función de similitud entre los nombres de los autores tiene dos objetivos fundamentales. Primero, determinar si dos nombres de autores son parecidos o no y segundo cuantificar la similitud entre los mismos. A los resultados de la **comparación de sub-cadenas** y **comprobación de apellidos compuestos** los llamaremos **disimilitud**.

Para cada apellido del autor **A** puntuaremos las coincidencias con el nombre del autor **B**. Los siguientes casos se evalúan en orden. Una vez se cumplen las pre-condiciones de uno se aplican los consecuentes y se obvian los demás.

| **Interpretación** | **Operación a realizar** |
|---|---|
| El apellido coincide con alguno de los apellidos del otro autor | $similitud = similitud + 30 - disimilitud * 10$ |
| El apellido coincide con alguna de las palabras del nombres del otro autor | $similitud = similitud + 10 * disimilitud * 5$ |
| Comprobar si el apellido puede ser compuesto | $similitud = similitud + 5 - disimilitud * 10$ |
| No era ninguno de los casos anteriores | Se termina la comparación y el valor devuelto es 0 |

**Cuadro 3.** Sistema de puntuación para la comparación entre los apellidos

Luego de encontrar una correspondencia para cada uno de los apellidos del autor, es necesario establecer una correspondencia entre los elementos restantes del nombre. Estos pueden ser o palabras del nombre o iniciales. En caso de ser palabras del nombre la puntuación es similar a la anterior (la diferencia principal es que no se suelen unir los nombres, por lo tanto no evaluamos esta posibilidad). En caso de ser iniciales se realizan las operaciones que se ilustran a continuación:

| Interpretación | Operación a realizar |
|---|---|
| La inicial corresponde con la inicial de alguno de los nombres a los que todavía no se le ha encontrado una correspondencia | $similitud = similitud + 20$ |
| La inicial corresponde con la inicial de algún apellido del otro autor a los que no se le encontró correspondencia | $similitud = similitud + 5$ |

**Cuadro 4.** Sistema de puntuación para la comparación entre las iniciales

Luego de determinar el valor de la función de similitud de acuerdo a lo mostrado anteriormente, este se compara con un umbral determinado de forma similar al utilizado en la homogenización de las afiliaciones. Si dicho valor es mayor que el umbral utilizando entonces dichos nombres posiblemente representen representaciones diferentes del mismo autor.

### 3.4. Relaciones de similitud

Luego de determinar la similitud entre los nombres de los autores se establecen relaciones entre los mismos utilizando los elementos disponibles para este proceso.

La utilización de los co-autores constituye un elemento importante en el proceso de desambiguación[12]. Para el establecimiento de la similitud entre los co-autores de dos autores se siguió el proceso descrito en **Algoritmo 1**.

**Algoritmo 1** Similitud entre los co-autores

**Entrada:** : Lista de co-autores del autor A con longitud $L1$, Lista de co-autores del autor B con longitud $L2$.
**Salida:** Similitud entre los co-autores.

**Inicializar:** Hacer similitud mayor *sim_may* y suma similitud *sum_sim* variable con valor 0
**para** $i = 1$ hasta $L1$ **hacer**
  **para** $j = 1$ hasta $L2$ **hacer**
    Calcular *similitud* entre el co-autor i-ésimo y el co-autor j-ésimo
    **si** $similitud > sim\_may$ **entonces**
      $sim\_may = similitud$
    **fin si**
  **fin para**
  $sum\_sim+ = sim\_may$
**fin para**
**devolver** $sum\_sim/L1$

Un elemento importante en la validación de la solución propuesta es el análicis de la eficiencia de los algoritmos presentados. La eficiencia se puede difinir como el uso óptimo de los recursos que el algoritmo utilice, en este caso, el tiempo de ejecución, teniendo en cuenta que dicho algoritmo llega a los objetivos

propuestos. Existen diferentes formas de evaluar la eficiencia de los algoritmos una de ellas es a través de la complejidad del mismo haciendo uso de la notación asintótica.

Para el **Algoritmo 1**, la mayor complejidad temporal radica en la aparición de dos ciclos anidados (líneas 2 y 3), siendo el resto de las sentencias del algoritmo de complejidad constante $O(1)$. Lo anterior produce una complejidad temporal cuadrática $O(n^2)$ para el algoritmo.

Otra relación de similitud que se establece es entre las afiliaciones de los autores, siendo las afiliaciones un elemento de importancia en el proceso de desambiguación[12]. Debido al proceso de normalización realizado con anterioridad la similitud que se establece entre las afiliaciones de los autores, es determinar si estas coinciden o no.

También se estableció una relación entre los lugares de publicación. Para el establecimiento de dicha relación se siguió un procedimiento similar al utilizado en la relación entre los co-autores. Primeramente se verifica cuántos lugares de publicación coinciden entre los dos autores comparados y finalmente se divide el número de coincidencias entre la cantidad de lugares de publicación menor de los dos autores.

Por último, se estableció una relación de similitud entre los títulos de las publicaciones de los autores. Teniendo en cuenta que los títulos de las publicaciones son un conjunto de palabras separadas por espacios, se sigue un procedimiento similar al establecido para la normalización de las afiliaciones. Se comparan cada una de las palabras del título del autor **A** con las del autor **B**, utilizando la distancia de edición. Luego, se seleccionan las palabras que coinciden en los dos títulos y finalmente se encuentra la razón entre la cantidad de palabras coincidentes en los dos títulos y la cantidad de palabras del menor de los títulos.

### 3.5. Combinación de Agrupamientos

La combinación de agrupamientos es el principal elemento dentro de la propuesta de solución. Las relaciones establecidas en las secciones anteriores son utilizadas para generar un vector de similitudes. Dicho vector es utilizado como elemento de agrupamiento en esta sección.

En la literatura se han propuesto diversos algoritmos para realizar dicho proceso[5], [13]. En la propuesta de solución se utiliza un método basado en la co-ocurrencia[5]. Partiendo de una matriz de co-ocurrencia donde se representan la cantidad de veces que un objeto ha sido colocado en el mismo agrupamiento luego de ejecutar **N** métodos de agrupamientos, dicho número se compara con el umbral fijo 0.5 y todos aquellos objetos que en la matriz de co-ocurrencia sea mayor que dicho umbral entonces son colocados en un mismo agrupamiento. El procedimiento es descrito en **Algoritmo 2** .

De acuerdo con[5] la complejidad temporal del algoritmo mostrado es de $O(n^2)$.

La propuesta de solución descrita en este artículo tiene como principal resultado el diseño de un algoritmo para la desambiguación del nombre de los autores en metadatos bibliográficos. El mismo se describe en **Algoritmo 3**.

**Algoritmo 2** Combinación de agrupamientos

**Entrada:** $n$ objetos, $\alpha$ combinación de agrupamientos conformada por $m$ particiones del conjunto de objetos.
**Salida:** Partición de consenso.

**Inicializar:** Hacer una matriz de co-asociación *co_assoc* nula de dimensiones $n$ *x* $n$.
Inicializar *co_assoc*
**para** $i = 1$ hasta $m$ **hacer**
  Correr el i-ésimo método de agrupamiento y producir la partición $\alpha_i$.
  **para** $i = 0$ hasta $m$ **hacer**
    **para** $j = 0$ hasta $m$ **hacer**
      **si** objeto i-ésimo y j-ésimo pertenecen al mismo agrupamiento en $\alpha_i$ **entonces**
        $co\_assoc(i, j) = co\_assoc(i, j) + 1/m$
      **fin si**
    **fin para**
  **fin para**
**fin para**
**para** $i = 0$ hasta $m$ **hacer**
  **para** $j = 0$ hasta $m$ **hacer**
    **si** $co\_assoc(i, j) > 0{,}5$ **entonces**
      Unir los objetos i-ésimo y j-ésimo en el mismo agrupamiento. Si los objetos pertenecen a dos agrupamientos diferentes se unen en uno solo.
    **fin si**
  **fin para**
**fin para**
Cada objeto no incluido en ningún agrupamiento forma un agrupamiento que solo lo contenga a él.

**Algoritmo 3** Algoritmo de desambiguación

**Entrada:** Lista de objetos a desambiguar con longitud $N$, *umbral* de comparación.
**Salida:** Lista de Autores desambiguados.

**Inicializar:** Hacer una lista de autores parecido *aut_par* vacía, Lista resultado *result_list* de autores desambiguados.
**para** $i = 0$ hasta $N$ **hacer**
  $aut\_par.add(objeto\ i - esimo)$
  **para** $i = i + 1$ hasta $N$ **hacer**
    Calcular *similitud* entre el objeto i-ésimo y j-ésimo.
    **si** $similitud > umbral$ **entonces**
      $aut\_par.add(objeto\ j - esimo)$
      Calcular *sim_vec* entre los elementos restantes de los datos de entrada presentes en *aut_par*.
    **fin si**
  **fin para**
  Combinación de agrupamientos().
  Adicionar un objeto de cada agrupamiento formado a la lista de autores desambiguados
**fin para**
**devolver** *result_list*

Finalmente se determina la complejidad temporal del **Algoritmo 3**. Este algoritmo posee dos ciclos anidados $O(n^2)$ (líneas 2 y 4), mientras que en el primer ciclo anidado existe una sentencia de complejidad $O(n^2)$ (línea 12), el resto de las sentencias son de complejidad contante $O(1)$. Como resultado, la complejidad temporal global del algoritmo es de $O(n^3)$, siendo una complejidad temporal aceptable con respecto a los estándares de diseño.

## 4. Conclusiones

En este artículo se propone un algoritmo para la desambiguación del nombre de los autores en metadatos bibliográficos. El algoritmo propuesto utiliza dos técnicas fundamentales, (1) la distancia de edición y (2) la técnica de combinación de agrupamientos. Entre sus bondades se destaca que se logra realizar un pre-procesamiento de los datos que sirven de entrada al algoritmo, mejorando considerablemente los resultados del mismo. La propuesta no es sensible al multilingüismo, lo que se logra mediante la utilización de la distancia de edición en la comparación de cadenas. Por último, la propuesta utiliza la técnica de combinación de agrupamientos, lo que constituye un elemento novedoso de la propuesta debido a que no se ha reportado su utilización en este tipo de problemas.

## Referencias

1. Berners-Lee, T.: Linked Data Design Issues, http://www.w3.org/DesignIssues/LinkedData.html.(2006)
2. Domingue, J., Fensel, D., Hendler, J.A.: Handbook of Semantic Web Technologies. Springer Heidelberg Dordrecht London New York (2011)
3. Li, S., Cong, G., Miao, C.: Author name disambiguation using a new categorical distribution similarity. In: Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases. pp 569-584. Springer-Verlag Press. Berlin, Heidelberg (2012)
4. Cheng, Y., Chen, Z., Wang, J., Agrawal, A., Choudhary, A.: Bootstrapping Active Name Disambiguation with Crowdsourcing. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. pp. 1213-1216.(2013)
5. Fred, A.: Finding Consistent Clusters in Data Partitions. Multiple Classifier Systems. Springer Berlin Heidelberg (2001).
6. Marzal, A., Vidal, E.: Computation of Normalized Edit Distance and Applications. vol 5, num 19, pp. 926-932. IEEE Press. (1993)
7. J.Pricilla.: An Efficient Framework for Name Disambiguition In Digital Library. vol 2. num 4. pp. 1097-1105. ISSN 2319-7242. (2013)
8. Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., Pinheiro, D.: A boosted-trees method for name disambiguation. Scientometrics. vol 93. num 2. pp. 391-411 (2012)
9. Ferreira, A., Machado, T.M, Gonalves, M.A.: Improving Author Name Disambiguation with User Relevance Feedback. Journal of Information and Data Management, vol 3. num 3. pp. 332. (2012)
10. Bernardi, R., Le, Dieu-Thu.: Metadata enrichment via topic models for author name disambiguation. In: Proceedings of the 2009 International Conference on Advanced language Technologies for Digital Libraries. vol 3. num 3. pp. 92-113. Springer-Verlag Press Berlin, Heidelberg (2011)

11. Veloso, A., Ferreira, A.A., Gonccalves, M.A., Laender, A., Meira,Jr., W.: Cost-effective on-demand associative author name disambiguation. vol 48. num 4. pp. 680-697. (2012)
12. Torvik, V.I., Weeber, M., Swanson, D.R., Smalheiser, N.R.: A probabilistic similarity metric for Medline records: A model for author name disambiguation. vol 56. num 2. pp. 140-158. (2005)
13. Dimitriadou, E., Weingessel, A., Hornik, K.: Voting-Merging: An Ensemble Method for Clustering. Artificial Neural Networks ICANN 2001. Springer Berlin Heidelberg Berlin (2001)