Clasificación de autores para un proceso de recomendación integrado a un metabuscador científico

Cantero, A.[1], Kuna, H.[1], Rey, M.[1]

[1] Departamento de Informática. Facultad de Ciencias Exactas, Químicas y Naturales.
Universidad Nacional de Misiones
{canteroalejandro, hdkuna}@gmail.com

Resumen. La búsqueda de producción científica en la web se ha convertido en un desafío, tanto por volumen, variedad y velocidad de actualización. Se requiere de herramientas que ayuden al usuario a obtener resultados relevantes ante la ejecución de una consulta. Dentro de estas herramientas, este equipo ha desarrollado un metabuscador específico para el área de ciencias de la computación. En su evolución se pretende incorporar recomendaciones de autores para cada consulta de sus usuarios. La generación de tales recomendaciones requiere de un método que cuente con la capacidad de clasificar a los autores para definir su inclusión y posición en un listado de sugerencias para el usuario final. En este trabajo se presenta un método que cumple con este objetivo, habiendo sido evaluado y habiendo obtenido resultados que permiten plantear su inclusión en un posterior desarrollo del sistema recomendador.

Palabras clave: indicadores bibliométricos, datos científicos, autores científicos, esquema de clasificación, sistemas de recomendación.

1 Introducción

La recuperación desde la web de la producción científica es, probablemente, uno de los desafíos más importantes de la actividad científica en la actualidad. El gran volumen de datos disponibles, su heterogeneidad y su velocidad constante de actualización hacen que sea necesario contar con herramientas que faciliten el acceso a información relevante ante una consulta particular. Como respuesta a esta situación, este equipo de trabajo en particular ha desarrollado un Sistema de Recuperación de Información (SRI), específicamente un metabuscador de dominio específico, para la recuperación de publicaciones y datos científicos pertenecientes a las ciencias de la computación.

Este tipo de herramientas no son estancas, sino que constantemente están en evolución para proveer resultados de mayor relevancia y calidad a sus usuarios sin que esto perjudique su rendimiento general. Ejemplos de esto son los algoritmos de ranking [1], así como la generación automática de expansiones de consulta [2].

En este contexto, una de las mejoras propuestas para el SRI [3], es la creación de un método que permita la recomendación de autores relacionados con la consulta de un usuario, a fin de reducir los tiempos de espera mientras el metabuscador ejecuta las consultas sobre fuentes externas, generando asi una respuesta complementaria de potencial interés para el usuario.

Con este objetivo se identificaron tres tareas principales a resolver previamente: la recopilación sistemática y automatizada de los datos primarios [4], la necesidad de elaborar una estructura que sea capaz de contener de forma organizada los metadatos recolectados [5], y la clasificación de autores. Esta última, tiene por objetivo describir y sintetizar la trayectoria que tiene cada autor respecto a un área de investigación, evaluando aspectos como la actividad, la vigencia, la especificidad dentro de una temática específica y su impacto en la comunidad científica entre otras cuestiones.

El objetivo del presente trabajo es definir un método de clasificación para un sistema de recomendación de autores científicos, considerando métricas e índices utilizados en el ámbito académico para la evaluación de los autores y su producción.

El resto del artículo se estructura de la siguiente manera, en la sección 2 se mencionan los antecedentes de la línea de investigación, en la sección 3 se presenta el método de clasificación, en la sección 4 se detalla la evaluación y validación del método propuesto, en la sección 5 se mencionan las conclusiones a las que se ha arribado y en la sección 6 se lista la bibliografía consultada.

2 Antecedentes

2.1 Un Meta-buscador para las ciencias de la computación

Un producto software debe evolucionar con el tiempo, incorporar nuevas funcionalidades y optimizar sus procesos a fin de brindar una mejor experiencia a sus usuarios. En este sentido, en la actualidad el metabuscador desarrollado para la recuperación de documentos científicos del área de Ciencias de la Computación, pretende integrar diferentes procesos de explotación de información que sean de utilidad para la extracción de conocimiento a partir de los datos que se procesan a partir de las búsquedas de los usuarios [6]. El objetivo de la incorporación de estos procesos es optimizar el funcionamiento general del sistema, tanto en lo que respecta a sus procesos internos como a aquellos componentes que tienen interacción con el usuario, en una búsqueda de mejorar su experiencia con la herramienta. En lo que respecta al presente trabajo, la integración de un sistema de recomendación de autores se concibe como una solución que aportaría al usuario un conjunto de datos acotado complementario a los resultados de la búsqueda.

2.2 Sistemas de Recomendación

Los Sistemas de Recomendación (SR) son técnicas y herramientas software que sugieren ítems que podrían ser relevantes para un usuario particular. Existen tres categorías generales para los métodos que son utilizados actualmente [7] , estos están clasificados de acuerdo a la forma en que son obtenidas las recomendaciones.

Los basados en contenido recomiendan ítems en función de valoraciones de ítems similares que el usuario haya realizado previamente [8]. Los SR colaborativos elaboran sus recomendaciones en base a perfiles de usuario con similitudes que hayan valorado a un elemento previamente [9]. Y los métodos híbridos se utilizan para proveer recomendaciones usando conceptos de ambos enfoques previos, superando así algunas de sus limitaciones.

2.3 Clasificación de autores para un recomendador

Un criterio típico de recomendación es la devolución de un conjunto de documentos en el cual todos sus elementos tienen un atributo con un valor que coincide total o parcialmente con el interés de un usuario. Sin embargo, únicamente retornar un conjunto desordenado de autores que mostraron alguna coincidencia con los intereses del usuario podría no ser un criterio efectivo. Por ello se consideró el desarrollo de un método capaz de organizar y clasificar a los autores considerando los aspectos más relevantes para la comunidad científica.

En una tarea de relevamiento bibliográfico se encontraron estudios enfocados en la evaluación del rendimiento de los autores en un nivel institucional [10] y sobre los factores que afectan su desempeño [11, 12]. Se observó que plataformas científicas como Microsoft Academic Search [13] (MAS) pone a disponibilidad de sus usuarios una sección de ranking de autores clasificados por "TOP" el cual permite discriminar por área de investigación, índice o puntaje y una ventana de tiempo. El equipo de investigación detrás de la plataforma AMiner[14] ha trabajado en el desarrollo de métricas que permiten evaluar a los autores por dimensiones o aspectos relevantes y en base a ello permite a sus usuarios una amplia gama de filtros para que estos puedan encontrar algún autor experto o influyente dentro de un área de investigación. Algunos autores han estudiado y analizado la trayectoria de los autores desde una dimensión con el fin de determinar el grado de utilidad de las métricas o del estado del arte de un área específica. Otros optaron por desarrollar modelos predictivos para detección de potenciales autores emergentes haciendo uso de reconocidas y nuevas métricas bibliométricos. Sin embargo, a pesar de los trabajos hallados no se encontró una taxonomía estandarizada de la trayectoria de los científicos y por ello se propone una clasificación de autores que considere los aspectos más relevantes de la trayectoria según las dimensiones más relevantes para la comunidad y que sea capaz de adaptarse a las necesidades de un proceso de recomendación.

3 Desarrollo del método de clasificación de autores

3.1 Cuestiones relativas a los datos

La sistematización de la recuperación de datos primarios para el método de clasificación de autores se estructuró a partir del trabajo realizado previamente sobre el metabuscador. Se utilizaron los procedimientos de Extracción, Transformación y Carga (ETL por su sigla en inglés) de los datos de las entidades involucradas en los procesos del metabuscador [4]. Estos datos fueron almacenados siguiendo los lineamientos planteados para la definición de los perfiles de las entidades con las que opera el SRI. Puntualmente, para este trabajo son de especial importancia los perfiles de los autores generados a partir de las fuentes de consulta habitual del metabuscador [5].

3.2 Métricas y aspectos evaluables de autores

Como se describió en la sección 2.2 para el logro de los objetivos del SR es necesario definir qué criterios se utilizarán para la evaluación de los autores. En [15] se describen

todos los posibles aspectos evaluables de un autor científico detallando para cada caso las técnicas disponibles y más utilizadas para su cuantificación. Considerando [14, 15] se elaboró la Tabla 1 que resume los aspectos que resultan más relevantes para la evaluación de la trayectoria.

Los índices bibliométricos son conocidos y utilizados por toda la comunidad científica en general. Dado que cada indicador puede presentar ventajas y desventajas en su utilización, se volvió frecuente en la comunidad, la adopción de más de un indicador en la evaluación de los autores [16]. Siguiendo esta línea de pensamiento, en

Tabla 1. Aspectos evaluables de la trayectoria de los científicos.

Aspecto	Definición	Técnica
Actividad	Evaluación de la producción científica, su circulación, dispersión y el impacto que ha producido en la comunidad.	El recuento de las publicaciones considerando el espacio de tiempo transcurrido entre las mismas.
Diversidad	Evaluación del grado de especificidad dentro de un campo, debido que un autor puede estar involucrado en diferentes campos de investigación.	Determinar la especificidad de un autor respecto a un área temática obteniendo una proporción entre la cantidad de artículos dedicados a un área específica con respecto al total de su producción.
Vigencia	Relativo a la vida media de un artículo considerando su citación a través del tiempo.	El recuento de las citaciones obtenidas en las publicaciones utilizando ventanas de tiempo.

este trabajo se propone el uso de una combinación de índices estratégicamente organizados que permitan evaluar más de un aspecto relevante. De esta manera, con el fin de desarrollar un conjunto de métricas primarias y derivadas de los índices bibliométricos más conocidos para la evaluación de autores se propone utilizar los siguientes índices:

Índice de Impacto Promedio: a fin de generar un índice que permita reflejar un impacto unificado de la producción de un autor se procedió al análisis de los índices H, G e i10 [17–19] de una muestra de 73[1] autores. De este análisis se observó la correlación que existe entre estos índices en reflejar la trayectoria de los autores, esto se debe a que todos ellos se basan en la cantidad de citas que han obtenido las publicaciones del autor.

Observando esta situación se propone un índice que promedie los anteriores obteniendo uno simplificado para la evaluación del autor definida por la ecuación (1).

$$IIP(a_n) = \frac{(indiceH(a_n) + indiceG(a_n) + indiceI10(a_n)}{3} \quad \textbf{(1)}$$

Índice de Longevidad: este índice se propone a fin de representar la vida académica del autor. Se define este índice (2) como la diferencia entre el último año de publicación y el año de su primera publicación.

[1] Muestra compuesta por índices bibliométricos de autores dedicados al campo de minería de datos. Fuente: elaboración propia con datos de MAS, query: "*datamining*".

$$IL(a_n) = (\text{año } del \text{ último } paper) - (\text{año } del \ primer \ paper) \quad \textbf{(2)}$$

Tasa de Publicaciones por Ventana de Tiempo: a fin de reflejar el desempeño de un autor dentro de una ventana de tiempo, se define este índice (3) como la relación entre el número de trabajos de un autor publicados entre los años de inicio y fin que determinan la ventana de tiempo.

$$\text{Sean } i \in (añoInicio,..,añoFin) \; ; \; p \in publicaciones \; ; \quad \textbf{(3)}$$

$$P(a) = \{p_i \; es \; publicación \; del \; autor \; a\} \; ; \; TPpVT\,(a_n) = \frac{\text{count}(\text{P}(a_n))}{añoFin - añoInicio}$$

Índice de Diversidad Relativo: el índice de diversidad (ID) definido en [14] intenta reflejar el grado de participación que tiene un autor en relación a diversos campos de investigación dentro de una ciencia o área de conocimiento. En este índice los valores más altos corresponden a aquellos autores que tienen un alto grado de participación en varios campos. Con el fin de evaluar la diversidad de un autor según un contexto determinado, se propone un índice que calcule el grado de diversidad de un autor con respecto a otros autores (4). Para ello se calcula la distancia que existe entre el mayor índice de diversidad de una colección con respecto a un determinado autor, de esta manera el más cercano y el más lejano estarán próximos a 1 y 0 respectivamente.

$$IDR(a_n) = \frac{ID(a_n)}{max(ID(a_1), ID(a_2), \ldots, ID(a_z))} \quad \textbf{(4)}$$

3.3 Clasificación de Autores

Dado que no se encontró en la literatura una clasificación estandarizada de los autores en función de su trayectoria científica, se propuso la definición de un conjunto de clases que se consideraron representativas de las diferentes variedades de autores. A estas clases se vincularon indicadores que permiten evaluar distintos aspectos del autor en función de la especificidad de cada clase definida. Y se estableció para cada índice un umbral que permita determinar si un autor cumple o no todas las condiciones para pertenecer a una clase.

Se propone un esquema de clasificación compuesto por 7 (siete) clases que sintetizan la trayectoria de los autores y los aspectos considerados de interés dentro de la comunidad científica, organizados en las siguientes agrupaciones:

Autores Nuevos: Son aquellos autores que su tiempo de actividad en la labor científica es reducida y que han tenido un impacto considerable con respecto a sus colegas con igual tiempo de actividad. Restricciones aplicadas:

(a) Índice de longevidad entre 1 y 5 años.

(b) La Cantidad de Citas deberá ser igual o superior al percentil 30 de la Cantidad de Citas de todos los autores que cumplen con la condición (a).

Autores Top: Son aquellos autores que han mostrado un impacto por encima de la media de sus colegas, destacándose la comunidad científica. Restricciones:

(c) IIP mayor o igual al 150% del valor del desvío estándar obtenidos a partir de la serie de IIP de todos los autores.

Autores Expertos: Son aquellos que han trabajado en proyectos de un campo determinado y son referentes en los mismos debido al impacto de sus trabajos. Restricciones:

(d) El IDR del autor deberá estar por debajo del 45%. Excluyendo del cálculo los valores que superen el Límite Superior considerados valores anómalos.
(e) El IIP deberá ser superior al percentil 40 del IIP de todos los autores excluyendo los valores que superen el Límite Superior por ser considerados anómalos.

Autores de Moda: Son aquellos autores que han realizado trabajos nuevos o innovadores en un campo de conocimiento que le han provisto de un impacto considerable en la comunidad en los últimos años. Restricciones:

(f) El autor no podrá ser moda si cumple con las restricciones de un autor Experto.
(g) El *TPpVT* de los últimos 5 años deberá ser mayor al valor de toda su carrera.
(h) El IIP del autor deberá ser mayor al primer cuartil (1Q) del IIP de todos los autores que cumplen con la condición (g).

Autores con Trayectoria: Son aquellos autores que han desarrollado una carrera como investigadores y han realizado contribuciones con impacto dentro de las ciencias de la comunidad científica. Restricciones:

(i) IIP deberá ser mayor al segundo cuartil (2Q) del IIP de todos los autores.

Autores One Hit Wonder: Son aquellos autores que han logrado un impacto en la comunidad por en un momento acotado, pero no se ha observado continuidad en su labor científica, al menos desde lo que respecta a publicaciones de su autoría. Restricciones:

- El Índice de longevidad deberá ser igual a 1 año.
- El IIP deberá ser mayor a la mediana o 2Q del IIP de todos los autores con longevidad menor o igual a 5 años.

Autores Sin Clasificación: Estos autores no corresponden con ninguna de las agrupaciones anteriores, razón por la cual se decidió agruparlas en una clase que permitan estudios posteriores.

3.4 Algoritmo propuesto para la clasificación de autores

Aunque preliminarmente se puede llegar a la conclusión de que el proceso de clasificación puede consistir en la aplicación directa de los axiomas sobre un conjunto de autores, realizarlo sin una estrategia que permita el cálculo de varios índices y la comprobación de todas las condiciones para cada una de las clases se incurriría en un uso ineficiente del proceso de cómputo. Si a esto se añade que existen clases con mayor especificidad en la evaluación de los autores, criterio reflejado en el número de índices utilizados y los umbrales de aceptación, ocurre que algunas clases son más específicas que otras que evalúan aspectos más generales. Esta situación hace que un autor puede estar en los umbrales de aceptación de más de una clase ya que puede cumplir con las condiciones necesarias, sin embargo, ante esta situación se priorizará evaluar y clasificar al autor en función de aquellas clases más específicas que coincidan con su perfil.

Con motivo de dar solución a estas cuestiones se propone un algoritmo de clasificación de autores a partir de un árbol de decisión booleano ad-hoc que dado una secuencia específica en la evaluación de los datos disponibles permita una clasificación eficiente.

```
Variables: autor_i: i-ésimo autor de una colección.
autores_sin_clasificar: vector cuyos elementos corresponden a
autores no clasificados.
IIP_nuevos: vector con los IIP de autores con menos de 5 años de
actividad académica.
IIP_todos: vector con los IIP de todos los autores.
IIP_expertos_LLS: vector con los IIP de los autores por debajo
del límite superior respecto a todos los autores.
IIP_moda: vector con los IIP de los autores que tienen un
promedio de publicaciones mayores en los últimos 5 años.
CC_nuevos: vector con la cantidad de citas de los autores con menos
de 5 años de actividad académica.
Por cada autor_i en autores_sin_clasificar:
  si calcLongevidad(autor_i) <= 1
   and IIP > calc2Cuartil(IIP_nuevos)
     clasificarComoOneHitWonder(autor_i)
  sino calcLogevidad(autor_i) <= 5
   and calcPercentil30(CC_nuevos)>= cantidadDeCitas(autor_i)
     clasificarComoNuevo(autor_i)
  sino (calcIndiceDeDiversidad(autor_i)
   and calcIIP(autor_i) < calcLimiteSuperior(IIP_todos)
   and calcDiversidadRelativa(autor_i) < 0.45
   and calcIPP(autor_i) >= calcPercentil40(IIP_expertos_LLS))
     clasificarComoExperto(autor_i)
  sino calcPromPubs5anios(autor_i) >= calcPromPubsTotalAños(autor_i)
   and calc1Cuartil(IIP_moda) < calcIPP(autor_i)
     clasificarComoModa(autor_i)
  sino calcIPP(autor_i) >= 1.5 * calcDesvioEstandar(IIP_todos)
    clasificarComoTop(autor_i)
  sino calcIPP(autor_i) > calc2Cuartil(IIP_todos)
    clasificarComoTrayectoria(autor_i)
  sino
    clasificarComoSinClasificacion(autor_i)
fin
```

4 Experimentación y Resultados

En la literatura del área se pudo observar que los métodos de clasificación son evaluados en líneas generales según dos dimensiones: el esquema de clasificación en sí mismo y las decisiones de clasificación [20].

En [21] se plantea que cuando un esquema de clasificación tiene el menor grado de incertidumbre posible, más efectivas serán sus decisiones de clasificación. Esto se logra cuando se tiene el mayor conocimiento acerca del dominio en el que se está desarrollando. Por tal motivo se realizó una evaluación del esquema analizando el grado de incertidumbre para cada dimensión:

- Objeto: Las propiedades de los autores para generar la clasificación están basados en índices bibliométricos de amplio uso y pueden ser generados a partir del conocimiento de las publicaciones y cantidad de citas de cualquier autor.
- Clase: Debido que la clasificación es propuesta en este trabajo, se tiene conocimiento de todas las restricciones definidas por cada clase, por lo que resulta claro y preciso.
- Esquema: En esta dimensión se presenta un grado de incertidumbre medio, debido a que no existe un estándar único establecido que permita una definición ampliamente aceptada por la comunidad. El grado de precisión en la definición de las clases se sustenta en un análisis subjetivo relacionado a la búsqueda de representar en las clases el interés que se podría tener por un autor en el contexto de aplicación del método de clasificación.

Para la evaluación de las decisiones de clasificación se siguió el enfoque propuesto en [21] que propone evaluar la bondad de ajuste de cada clase al evaluar autores. En este caso, el proceso de evaluación consistió en la ejecución de una serie de algoritmos de clasificación automática sobre un conjunto de datos de autores previamente clasificados utilizando el método planteado en el presente trabajo. El objetivo en esta tarea consistió en verificar si un método automático podría ser entrenado en base a la clasificación generada y luego evaluar la efectividad de los resultados. En caso de obtener valores de efectividad elevados, se podría postular que las clases y el algoritmo de clasificación ad-hoc desarrollados son consistentes.

Siguiendo este enfoque se conformó una base de datos de autores extraídos de las plataformas MAS y AMiner mediante un proceso automatizado de ETL desarrollado para el metabuscador. El método de recolección de los datos se inició utilizando como punto de partida los autores de los artículos científicos que han sido recuperados por las consultas de los usuarios, los datos recuperados están conformados por los índices bibliométricos, publicaciones y coautores disponibles en la plataforma de origen. Posteriormente se procedió a la evaluación de la calidad de los datos recuperados con el fin de determinar el grado de correctitud, completitud y consistencia mediante técnicas de muestreo aleatorio y el "estándar de oro". De esta manera el dataset quedo conformado por un total de 2164 perfiles de autores siguiendo la estructura definida en [5]. Posteriormente, se seleccionó una serie de métodos de clasificación, algunos de la familia TDIDT[2] como son: C4.5, Random Forest, Simple CART y Rule Induction (RI), en forma conjunta con una red neuronal de tipo perceptrón multicapa y un clasificador bayesiano implementados en la herramienta WEKA. Los parámetros de los métodos empleados fueron establecidos con los valores por defecto que plantea la herramienta, siempre que el dataset no requiriera algún tipo de ajuste por las características del mismo. Los resultados obtenidos del proceso de entrenamiento-testeo (proporción 70/30 del dataset) de los métodos mencionados se detallan en la tabla 2, en el cual se observan resultados con una eficiencia entre el 81.82% al 95.22% de efectividad en la clasificación y una tasa de error por debajo del 20% para el peor de los casos. En los casos de baja efectividad, se observó que la clase que presenta el mayor problema de precisión es *Moda* debido que el algoritmo confunde a la clasificación de los autores con la clase *Trayectoria*, esto se debe probablemente a lo descrito previamente relativo al solapamiento entre las clases más específicas y las generales.

[2] TDIDT sigla para "Top Down Induction of Decision Trees".

Tabla 2. Resultados algoritmos de clasificación automática.

Criterios de Evaluación	C4.5	RF	Simple CART	RI	RNA-MC	Bayes.
Accurancy	95.22%	95.22%	94.3%	92.45%	81.97%	81.82%
Clasification Eror	4.78%	4.78%	5.7%	7.55%	18.03%	18.18%
Weighted Mean Precision	95.28%	95.23%	94.56%	78.74%	74.95%	83.28%
Weighted Mean Recall	94.38%	94.08%	88.92%	79.63%	74.82%	76.61%

Para el recall o exhaustividad se observa que las clases *Moda* y *OHW* presentan el menor porcentaje de efectividad con respecto a las clases Top y Nuevo respectivamente, esto se debe probablemente a que estas clases son similares en ciertos aspectos, sin embargo, las restricciones establecidas por clase definen su pertenencia a una u otra.

5 Conclusiones y Líneas futuras de investigación

Se ha generado un método de clasificación para autores científicos en el marco de un SRI del área de ciencias de la computación. El método presentado incluye la definición de una serie de clases que representan diferentes estadios en la carrera de un investigador, siendo contemplados diversos factores como el impacto generado, su especialización y su trayectoria. El desarrollo presentado ha sido validado en términos de su definición y pertinencia, siendo que sobre un dataset en el que fue aplicada la clasificación se probaron algoritmos de clasificación automática, obteniendo resultados de alta efectividad y baja tasa de errores.

El método se ajusta a las necesidades de un SR de autores para un metabuscador de tipo científico del área de ciencias de la computación. Su implementación permitirá presentar perfiles de autores de relevancia para los usuarios a partir de sus consultas.

Como líneas futuras se proponen, inicialmente, la implementación del sistema de recomendación para el metabuscador. Además, la transición del método desarrollado a uno que integre técnicas de clasificación automática a fin de generalizar y optimizar los procesos involucrados, evaluando la factibilidad de incluir en el análisis una mayor cantidad de indicadores.

6 Bibliografía

1. Kuna, H.D., Rey, M., Martini, E., Solonezen, L., Sueldo, R.: Generación de un algoritmo de ranking para documentos científicos del área de las Ciencias de la Computación. Presented at the XVIII Congreso Argentino de Ciencias de la Computación , Mar del Plata, Argentina October (2013).
2. Rey, M., Kuna, H.D., Martini, E., Podkowa, L., Pautsch, J.G.A., Zamudio, E.: Generación de un método de expansión de consultas basado en ontologías para un sistema de recuperación de información. Presented at the XX Congreso Argentino de Ciencias de la Computación , La Matanza, Argentina (2014).
3. Rey, M., Kuna, H.D., Martini, E., Canteros, A., Cantero, A., Rambo, A., Biale, C.O.: Propuesta de esquemas de perfiles para la recuperación de datos científicos para un sistema

de recuperación de información del área de Ciencias de la Computación. Presented at the XXII Congreso Argentino de Ciencias de la Computación , Junín, Argentina (2016).
4. Rey, M., Kuna, H.D., Rambo, A., Canteros, A., Cantero, A., Martini, E., Corrales, N., Rauber, F.: Propuesta de procesos complementarios para un sistema de recuperación de información. Presented at the XIX Workshop de Investigadores en Ciencias de la Computación , Buenos Aires August 23 (2017).
5. Kuna, H., Rey, M., Zamudio, E., Olivas, J.A., Rambo, A., Cantero, A., Canteros, A., Martini, E., Biale, C.O.: An entity profile schema for data integration in an academic metasearch engine. In: International Conference on Artificial Intelligence. , Las Vegas, Nevada (2017).
6. Kuna, H.D., Rey, M., Zamudio, E., Canteros, A., Cantero, A., Rambo, A.R., Martini, E., Pautsch, J.G.A., Biale, C.O., Krujoski, S., Rauber, F.: Diseño y construcción de procesos de explotación de información para el área de ciencias de la computación. Presented at the XX Workshop de Investigadores en Ciencias de la Computación, WICC 2018 (2018).
7. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering. 17, 734–749 (2005).
8. Mooney, R.J., Roy, L.: Content-based Book Recommending Using Learning for Text Categorization. In: Proceedings of the Fifth ACM Conference on Digital Libraries. pp. 195–204. ACM, New York, NY, USA (2000).
9. Rich, E.: Stereotypes and User Modeling. In: Kobsa, A. and Wahlster, W. (eds.) User Models in Dialog Systems. pp. 35–51. Springer Berlin Heidelberg (1989).
10. Carpenter, M., Gibb, F., Harris, M., Irvine, J., Martin, B., Narin, F.: Bibliometric profiles for British academic institutions: An experiment to develop research output indicators. Scientometrics. 14, 213–233 (1988).
11. McGrail, M.R., Rickard, C.M., Jones, R.: Publish or perish: a systematic review of interventions to increase academic publication rates. Higher Education Research & Development. 25, 19–35 (2006).
12. Costas, R., van Leeuwen, T.N., Bordons, M.: A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. J. Am. Soc. Inf. Sci. 61, 1564–1581 (2010).
13. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. (Paul), Wang, K.: An Overview of Microsoft Academic Service (MAS) and Applications. In: Proceedings of the 24th International Conference on World Wide Web. pp. 243–246. ACM, New York, USA (2015).
14. Tang, J.: AMiner: Mining Deep Knowledge from Big Scholar Data. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 373–373. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2016).
15. Velasco, B., Bouza, J.M.E., Pinilla, J.M., Román, J.A.S.: La utilización de los indicadores bibliométricos para evaluar la actividad investigadora. Aula abierta. 40, 75–84 (2012).
16. Costas, R., Bordons, M.: Is g-index better than h-index? An exploratory study at the individual level. Scientometrics. 77, 267–288 (2008).
17. Hirsch, J.E.: An index to quantify an individual's scientific research output. PNAS. 102, 16569–16572 (2005).
18. Egghe, L.: Theory and practise of the g-index. Scientometrics. 69, 131–152 (2006).
19. Aguillo, I.F.: Is Google Scholar useful for bibliometrics? A webometric analysis. Scientometrics. 91, 343–351 (2012).
20. Ranganathan, S.R.: Prolegomena to library classification. Madras Library Association, Madras (1937).
21. Bedford, D.: Evaluating classification schema and classification decisions. Bulletin of the American Society for Information Science and Technology. 39, 13–21 (2013).