ISSN 1870-4069

Visualización de elementos de cienciometría con grafos

Pedro Bello, Meliza Contreras, Diana A. González

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, México

{pbello, mcontreras}@cs.buap.mx,
diana_gonznava@hotmail.com

Resumen. En este documento se presenta una herramienta computacional para el análisis de la cantidad de artículos de investigación generada por un conjunto de investigadores; se utilizan los conceptos básicos de la teoría de grafos para representar la cantidad de artículos generados por un investigador y como métrica principal el número de colaboraciones entre un autor y los coautores, así como el análisis de persistencia y continuidad de colaboración científica.

Palabras clave: Cienciometría, colaboración científica, grafo de colaboración.

Visualizing Scientometrics Items with Graphs

Abstract. In this work is presented a computational approach for the analysis of the number of research articles generated by a group of researchers, the basic concepts of the theory of graphs are used to represent the quantity of the items generated by a researcher and as principal metrics is considered the number of collaborations between an author and co-authors as well as analysis of persistence and continuity of scientific collaboration.

Keywords. Scientometrics, scientific collaboration, collaboration graph.

1. Introducción

Con el crecimiento de Internet en los últimos años se han venido desarrollando diversas áreas de investigación que se relacionan con el acceso a la información y el estudio del conocimiento. De esta forma tenemos áreas de interés como la Cybermetrics [1] que estudia los recursos de información en Internet, la Webometrics [14] que estudia los aspectos cuantitativos de construcción y uso de recursos de tecnologías en la web, la Informetrics [2] que maneja aspectos de recuperación de palabras, documentos y bases de datos, Bibliometrics [1] que se encarga del estudio cuantitativo de las publi-

caciones físicas y Scientometrics [3] que se encarga del estudio de la producción científica a través de métodos matemáticos y estadísticos. En la Figura 1 se muestra la relación de las diversas áreas que estudian el acceso a la información.

La Cienciometría (Scientometrics) como ciencia estudia los aspectos cuantitativos de la producción académica; surgió en Europa en 1977 con el nacimiento de la revista Scientometrics. Entre los principales temas que estudia la Cienciometría se encuentran: el crecimiento cuantitativo de la ciencia en base a la producción académica de los investigadores, el desarrollo de las áreas y subáreas, así como la productividad y creatividad de los investigadores. En este contexto se plantea una herramienta computacional que muestre de forma gráfica la producción académica de un grupo de investigadores, utilizando conceptos básicos de grafos. En primera instancia se obtiene un conjunto de datos de prueba de DBLP (Digital Bibliography & Library Project), el cual es un sitio web que posee un enorme repositorio bibliográfico de artículos relacionados con Ciencias de la Computación. El sitio está alojado en la Universidad de Trier, Alemania. Ha evolucionado desde un pequeño servidor web experimental a un popular servicio de datos abiertos para la comunidad en Ciencias de la Computación [15]. La información obtenida viene dada en formato XML (eXtensible Markup Language), que es un lenguaje estándar de marcas que posee una recomendación del World Wide Web Consortium (W3C) y que fue diseñado para almacenar y transportar datos y para ser autodescriptivo. De esta forma se procesan los archivos XML para determinar la producción académica de cada investigador.

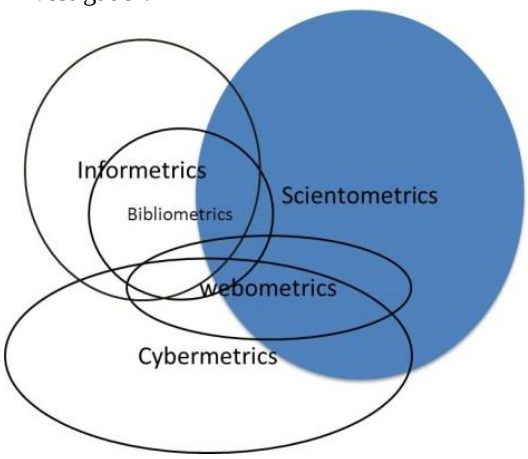

Fig. 1. Relación de las áreas de estudio de la información.

Se realizó una revisión del panorama general de los trabajos relacionados con el tema de estimar el trabajo de los investigadores en la producción de artículos de investigación científica. Por ejemplo, en [4] se muestra un método para el cálculo de citas por autor utilizando matrices, mientras que en [5] se realiza un estudio de la producción académica considerando varios parámetros como la relación entre coautores, co-referencias y co-citas. En [6] se realiza un trabajo para verificar si es válido el método de co-citas para medir el desempeño de los investigadores, debido a que las bases de datos de artículos muchas veces arrojan datos por autores con los mismos apellidos. En [7] se

realiza una comparación entre Web of Science (WoS) de Thomson Reuters y Scopus de Elsevier, y ambos indican que hay que tener precaución debido a que citan problemas con diferentes campos, instituciones, países y lenguajes. En [8] se presenta un estudio que tiene como objetivo examinar la asociación entre el autor y el acoplamiento bibliográfico en 18 áreas temáticas, concluyendo que no hay diferencias significativas en las diversas áreas analizadas. En [9] se realiza un estudio del desarrollo de la producción científica con los datos obtenidos de diversas bases de información en México. Finalmente, en [10] se presenta un modelo para identificar perfiles de usuarios utilizando un grafo de co-ocurrencia.

2. Scientometrics en México

En esta sección mostramos el desarrollo de la ciencia utilizando Scopus que es una base de datos de producción científica a nivel internacional que almacena principalmente: artículos científicos, libros y reportes de conferencias, ofreciendo una visión global del mundo de los resultados de la investigación en los campos de la ciencia, tecnología, medicina, ciencias sociales, artes y humanidades. Scopus ofrece herramientas inteligentes para rastrear, analizar y visualizar la investigación [16].

En la Figura 2 se muestra la producción científica por año, según los resultados se cuenta con 266,378 documentos (consulta realizada el 15 de julio 2016), de los cuales 20,386 corresponden a Ciencias de la Computación, 934 a la Benemérita Universidad Autónoma de Puebla (BUAP). En la intersección de estos conjuntos se tienen 146 artículos que corresponden a los desarrollados en Ciencias de la Computación en la BUAP.

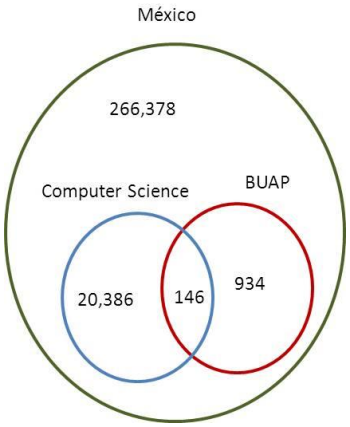

Fig. 2. Producción académica por año en México.

En la Figura 3 se indica que, de acuerdo a Scopus, la BUAP está situada en el ranking 6 de las universidades más productivas del país.

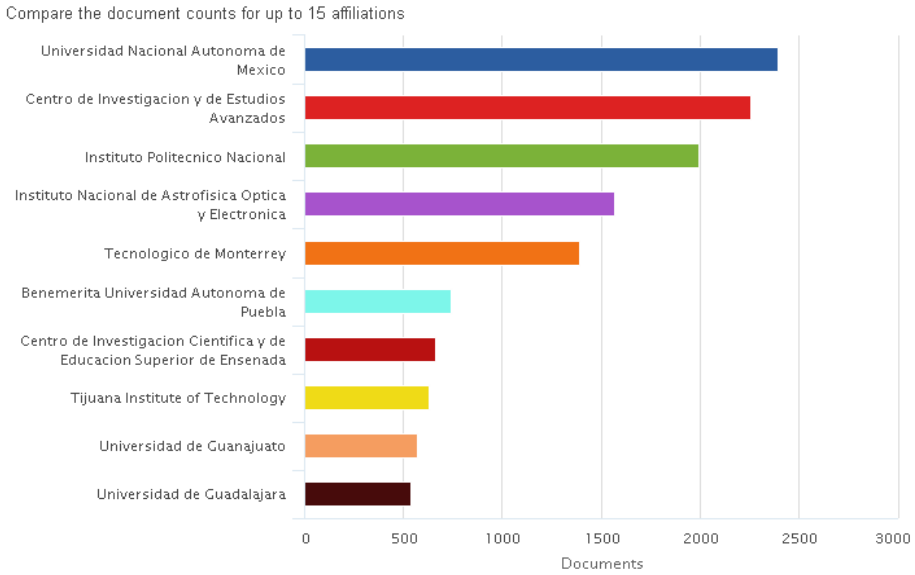

Fig. 3. Producción académica BUAP.

En la Figura 4 se muestra la producción de los documentos reportados por área de conocimiento en México. Se observa que el 24% corresponde a producción de artículos en medicina y solo el 7.7% corresponde al área de Ciencias de la Computación.

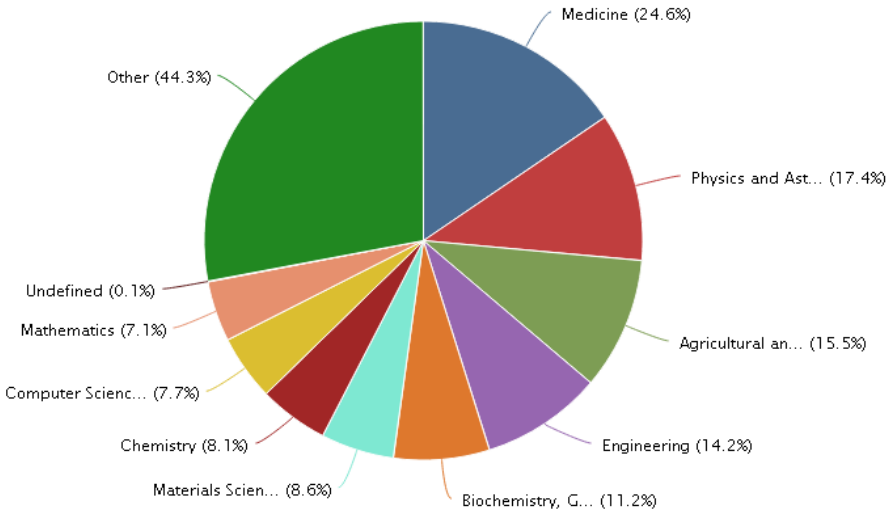

Fig. 4. Producción científica por área.

En la Figura 5 se muestra la producción científica por autor en México, en el área de Ciencias de la Computación.

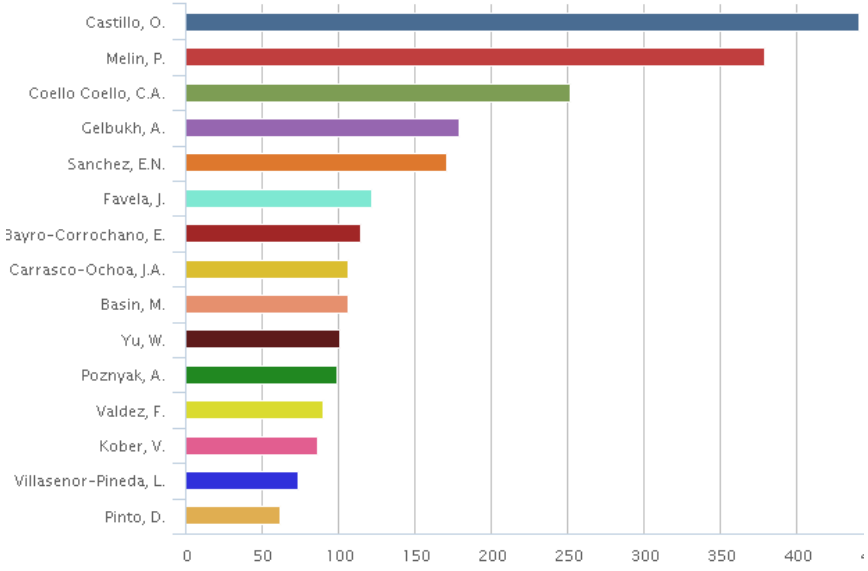

Fig. 5. Producción científica por autor en Ciencias de la Computación (México).

Los sistemas como Scopus permiten realizar un análisis de la producción de documentos científicos. Sin embargo, es posible obtener más conocimiento si, por ejemplo, se desea determinar con que personas se relaciona un investigador determinado. Es justo aquí el punto de estudio del trabajo presentado en este artículo, al proponer una herramienta computacional que aplique diversas métricas para el análisis complementario de la producción académica en México.

3. Metodología

La propuesta para la visualización de elementos de Cienciometría mediante grafos se desarrolló en base a la siguiente metodología de trabajo:

1. Extracción de la información en una base de información de la publicación de artículos.
2. Revisión de la información y el formato XML utilizado.
3. Desarrollo de un programa en PHP que utiliza la librería Vis.js para la representación visual del grafo.
4. Visualización del grafo de colaboración entre autores.

En la Figura 6 se presenta el modelo de la propuesta de la herramienta computacional para determinar la colaboración entre autores en el desarrollo de artículos de investigación. En primera instancia se toman los datos de una base de datos de información de artículos; en este caso se utilizó DBLP que genera archivos en el formato XML, poste-

riormente se creó una aplicación en PHP [17] utilizando una librería grafica denominada Vis.js [18] para la representación del grafo de colaboración, el cual es visualizado a través de un navegador web.

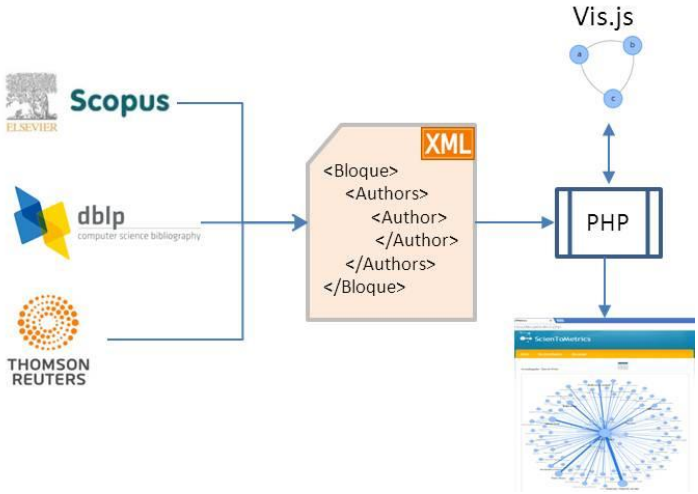

Fig. 6. Modelo de la herramienta computacional propuesta.

3.1. Características cualitativas y cuantitativas

Para medir el desempeño de los investigadores en la producción científica existen diferentes opiniones, por ejemplo [11] propone diferenciar entre indicadores cuantitativos y cualitativos. Las características cualitativas [12] son difíciles de medir mientras que las características cuantitativas utilizan métodos estadísticos. En este trabajo se utilizan medidas cuantitativas básicas.

- Cantidad de publicaciones en revistas científicas y memorias en extenso.
- Cantidad de publicaciones respecto a los coautores.

4. Herramienta computacional propuesta

La propuesta de esta herramienta computacional para el análisis y representación de la colaboración entre autores de artículos de investigación, se basa en la creación de un grafo que acumula la cantidad de artículos producidos por un determinado investigador.

Sea G = (V, E) un grafo sin dirección con un conjunto V de vértices (o nodos) y un conjunto E de aristas (o arcos). El grafo generado tiene las siguientes características:

- Expansión: el nodo central del autor principal acumula la cantidad de artículos en los que ha participado como autor o coautor:

$$v_i = \sum_{n=1}^{\infty} \text{i}, \quad donde\ i\ es\ autor\ o\ coautor. \tag{1}$$

- Amplitud: el arco de cada conexión con los coautores representa la cantidad de artículos en los que ambos han participado. Sea e_{ik} la existencia de una publicación del autor i con el coautor k:

$$Amplitud = w(e_{ik}), donde\ w\ es\ el\ peso\ de\ la\ arista\ (e_{ik}). \tag{2}$$

- Tasa de participación: para grafos relacionados con la colaboración entre los investigadores se calcula el promedio de las participaciones en los artículos:

$$t_{ik} = \frac{v_i}{w_{ik}}. \tag{3}$$

- Relación: en el grafo de colaboración se muestra la relación entre los n autores, donde se diferencian mediante los nodos expandidos, los investigadores que más participan en colaboración con los demás autores:

$$r_i = \sum_{k=1}^{n}(w_{ik}), para\ i = 1..n. \tag{4}$$

Fig. 7. Grafo de colaboración del autor y coautores.

En la Figura 7 se muestra el grafo generado con las publicaciones de la Dra. Mireya Tovar. Como se puede notar, los nodos tienen diferentes tamaños debido a la operación de expansión (1). Además, los arcos de conexión tienen diferente amplitud (2) debido a que existe mayor colaboración entre el autor principal y sus coautores, por ejemplo, la cantidad de artículos entre Mireya Tovar y David Pinto es de 24 como se indica en

la arista correspondiente. En el sistema desarrollado, al acercarse al arco correspondiente se muestra la cantidad de artículos de colaboración. En la Figura 8 se indica el grafo generado de relación entre un conjunto de investigadores; sobresalen los investigadores con mayor productividad (3) y los arcos correspondientes son más amplios ya que representan la mayor colaboración entre pares de investigadores.

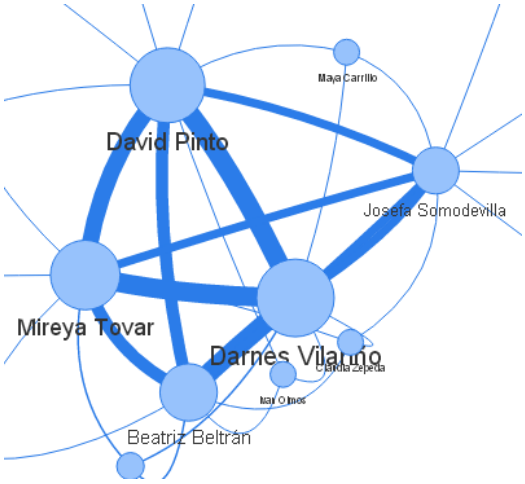

Fig. 8. Grafo que representa la relación entre un grupo de investigadores.

Los grafos de las Figuras 7 y 8 son generados con la herramienta desarrollada. La interfaz principal se muestra en la Figura 9, las opciones que se tiene son: generar el grafo por investigador o por grupo.

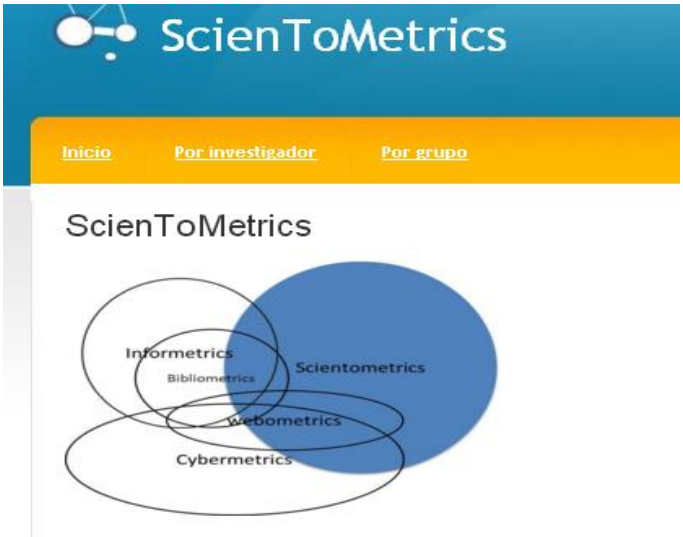

Fig. 9. Vista principal del sistema desarrollado.

En la Figura 10 se muestra el grafo generado de la relación de trabajo entre autores de artículos (4); en el sistema se cuenta con un icono en el lado superior derecho que da como resultado los valores mostrados en la Tabla 1, los cuales corresponden a la cantidad de artículos en colaboración del autor seleccionado.

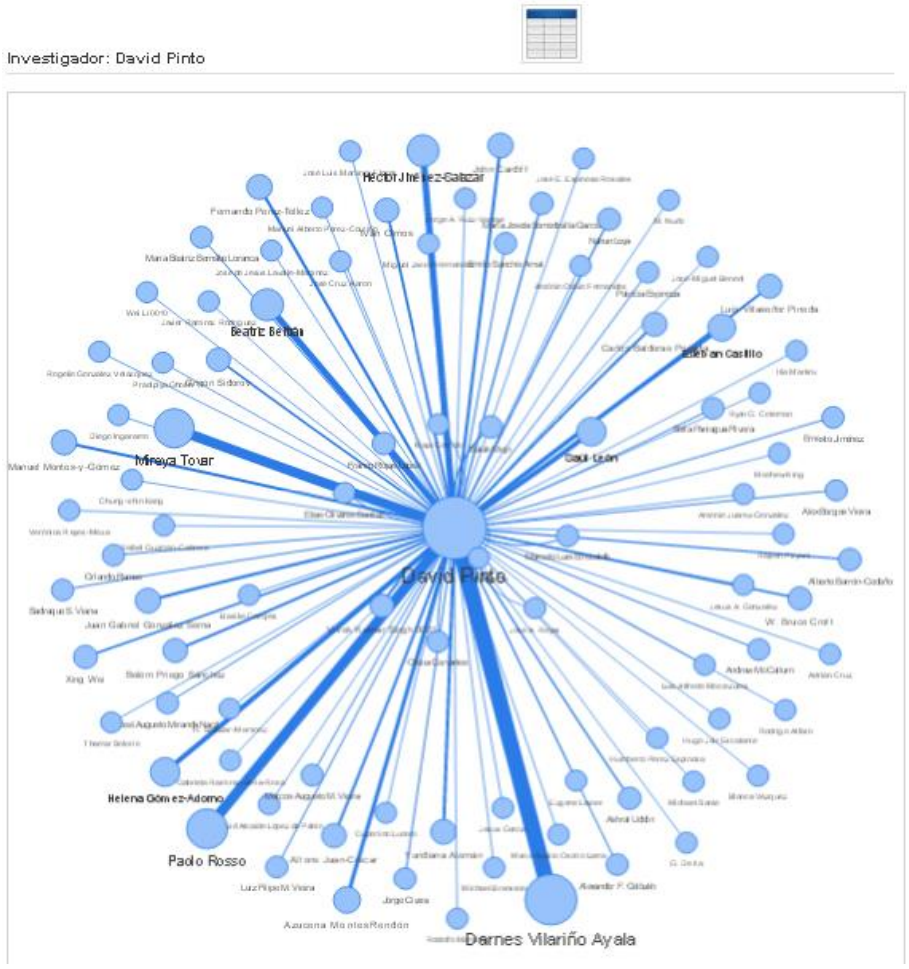

Fig. 10. Grafo de colaboración del Dr. David Pinto a través del sistema Scientometrics.

Tabla 1. Matriz generada para el grafo de relaciones entre investigadores.

	Pinto	Tovar	Beltran	Vilariño	Castro	Somodevilla
Pinto	0	24	15	39	0	3
Tovar	24	0	19	29	1	3
Beltran	15	19	0	18	1	2
Vilariño	39	29	18	0	1	5
Somodevilla	3	3	2	5	0	0

4.1. Estructura de la aplicación

La aplicación propuesta utiliza los grafos como estructura de datos y los archivos xml donde se encuentra almacenada la información de las publicaciones de los investigadores. Se utiliza además PHP como lenguaje de programación y un visor web para la presentación de los resultados. PHP es un lenguaje del lado del servidor, por lo que requiere cualquier servidor web para su ejecución. La principal ventaja de utilizar esta estructura del sistema es que permite ofrecer el servicio a través del servidor a muchas personas con conexión a Internet. Además, facilita la comunicación y extracción de la información para alimentar el sistema. PHP ofrece un conjunto de tecnologías adicionales que permiten la presentación de resultados de forma más atractiva, e interactuar con otros lenguajes de programación y con gestores de bases de datos.

4.2. Pruebas de la aplicación

El sistema desarrollado permite generar el grafo de colaboración de los investigadores. Las pruebas generadas corresponden a leer un archivo xml de DBLP, y se procesa en un programa PHP. Los datos extraídos se muestran en la Tabla 1, donde se indica los datos de un grupo representativo de investigadores de Ciencias de la Computación y se verificó que los resultados obtenidos corresponden a la cantidad de artículos reportados en la base de información correspondiente.

Dada la gran cantidad de información que se utiliza en este tipo de sistemas, es conveniente aplicar una prueba de rendimiento (performance testing) [13] para determinar qué tan rápido responde el sistema con archivos muy extensos y con grupos de investigadores muy amplio.

5. Conclusiones

Se desarrolló una propuesta inicial de un sistema Scientometrics a través de la recopilación de información de una base de información de documentos de investigación en Ciencias de la Computación como DBLP. Dicha propuesta está acompañada de un prototipo con el fin de mostrar que el m4odelo es viable. En este trabajo también se expone que existen sistemas donde generar de forma automática algunos estadísticos que permiten analizar la situación actual de la investigación en México. Sin embargo, es necesario aplicar otro tipo de estrategias de medición, por ejemplo, mostrar la relación de trabajos que existe entre investigadores, indicar cuantos productos se generan en la investigación, indicar la importancia que tiene un investigador respecto a otros, entre otros factores.

El prototipo mostrado sin duda puede ser mejorado, agregando estadísticos que puedan medir de forma cualitativa la investigación y más estadísticos para medir de forma cuantitativa la cantidad de artículos desarrollados por año. Otro de los factores que se pueden medir y determinar gráficamente en este prototipo es conocer el o los colaboradores de un grupo de investigación que participan poco y así integrarlos al trabajo en colaboración con los demás investigadores.

Referencias

1. Thelwall, M., Tsou, A., Weingart, S., Holmberg, K., Haustein, S.: Tweeting links to academic articles. Cybermetrics: International Journal of Scientometrics Informetrics and Bibliometrics, 17 (1), 1–8 (2013)
2. Bar-Ilan, J.: Citations to the Introduction to informetrics. Scientometrics, 82(3), 495–506 (2010)
3. Amara, N., Landry, R: Counting citations in the field of business and management: Why use Google Scholar rather than the Web of Science. Scientometrics, 93 (3), 613–625 (2012)
4. Pinski, G.: Citation based measures of research interactivity. Scientometrics, 2 (4), 257–263 (1980)
5. Krauze, T.K, Mcginnis, R.: A matrix analysis of scientific specialties and Careers in science. Scientometrics, 1 (5-6), 419–444 (1979)
6. Garfield, E.: Is citation analysis a legitimate evaluation tool? Scientometrics, 1 (4), 359–375 (1979)
7. Mongeon, P: The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106 (1), 213–228 (2016)
8. Gazni, A., Didegah, F: The relationship between authors' bibliographic coupling and citation exchange: analyzing disciplinary differences. Scientometrics, 107 (1), 609–626 (2016)
9. Ashraf, A., Singh, V. K., Pinto, D., Olmos, I.: Scientometric mapping of computer science research in Mexico. Scientometrics, 105 (1), 97–114 (2015)
10. Espinoza, P., Vilariño, D.; Pinto, D., Somodevilla, J., Tovar M.: Metodología basada en grafos para la identificación de perfiles de usuario. Research in Computing Science, 97, 127–139 (2015)
11. Fehnert, B., Kosagowsky, A., Measuring User Experience - Complementing Qualitative and Quantitative Assessment. In: MobileHCI '08: Proceedings of the 10th international conference on Human computer interaction with mobile devices and services, ACM, 383–386 (2008)
12. Ayala, S, Cuaya, Medina, M, Muñoz, A.: Representación con Restricciones de Medidas Cualitativas: Aplicación a un Problema de Scheduling. LANMR (2006)
13. Somerville, I: Software engineering. Addison Wesley (2011)
14. Webometrics. `https://en.wikipedia.org/wiki/Webometrics` (2016), Accedido el 26 de Abril de 2016
15. DBLP. `http://dblp.uni-trier.de/faq/What+is+dblp.html` (2016), Accedido el 26 de Abril de 2016
16. SCOPUS. `https://www.elsevier.com/solutions/scopus` (2016), Accedido el 6 de Abril de 2016
17. Php. `https://secure.php.net/` (2016). Accedido el 26 de Abril de 2016
18. Visjs. `http://visjs.org/` (2016). Accedido el 26 de Abril de 2016