Modelo de distribución de conocimiento tácito a través de redes sociales para la toma de decisiones.

(November 2018)

José Martínez, Jorge Rodas.

Abstract— En el proceso de investigación, necesitamos analizar grandes cantidades de información, otras investigaciones, libros, manuales etc. Cuando gran cantidad de información es obtenida de previas investigaciones, las fuentes de información son de las más recientes o provienen de investigadores con gran conocimiento en el campo, esto nos garantiza que la calidad de nuestra investigación mejorará al estar muy bien fundamentada. Esto nos lleva al hecho de que las publicaciones de artículos hechas en ciertos temas sean un poco redundantes en las referencias o bases de estudio, las cuales a lo largo del tiempo van cambiando abriendo paso a nuevas publicaciones con mayor importancia en el conocimiento, manteniendo algunas otras y desechando aquellas que se han vuelto obsoletas en el tema. Esto probablemente causa que grupos de personas que trabajan en el mismo tema han estado apoyándose mutuamente y sin darse cuenta han formado una red de investigadores en una manera tácita. Con esta investigación vamos a probar la existencia de estas redes de conocimiento y presentarlas en una manera gráfica para visualizarlas y analizar su distribución y así facilitar la toma de decisiones.

Index Terms— knowledge networks, Decision making, social network

1 Introduccion

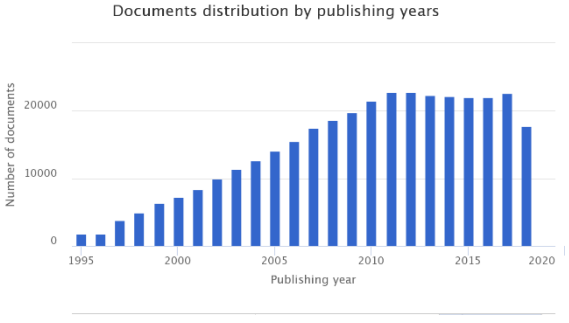
Durante el desarrollo de proyectos es necesario comenzar una investigación acerca de otras investigaciones en el mismo tema en donde encontremos los avances más novedosos y nuevos en la materia a lo cual dentro de las tesis doctorales llaman "estado del arte" o SoA.

Para poder llegar a resultados satisfactorios y de calidad en la obtención de información necesaria podría tomarnos mucho tiempo, semanas, meses o inclusive años dependiendo el tema y el alcance de la investigación. En ese tiempo los investigadores seleccionan un número determinado de fuentes de conocimiento como muestra, de las cuales obtendrán un listado de artículos que formaran su listado de referencias o base de conocimiento. Pero qué pasaría si esa selección no fuera la más óptima o si pudiera ser más precisa de tal manera que nos ayude a alcanzar mejores objetivos. Si pudiéramos ver de una manera gráfica que elementos forman parte del SoA, cuales son los de mayor relevancia y cuales pudieran excluirse fácilmente por no dar gran contribución en un tema específico.

La existencia de relaciones dentro de los artículos científicos por medio de referencias y sus autores ha creado una red de conexiones entre ellos que si pudiéramos visualizarla nos mostraría datos de gran interés para el investigador, como la formación de grupos de trabajo a través del tiempo. A estos grupos interconectados vamos a llamarlos redes de conocimiento.

2 Grades Cantidades de Conocimiento

De acuerdo a datos obtenidos del portal SciELO.org [1] (Scientific Electronic Library Online por sus siglas en ingles), una de las hemerotecas virtuales más importantes en américa latina. El número de publicación de documentos científicos ya sean en html o pdf ha ido creciendo de manera significativa como muestra en la figura 1.



1. Archivos registrados en SciELO.org por año.

Algunos de los problemas a los cuales se enfrentan los investigadores durante la realización de sus proyectos son:

- *Ing. Jose de Jesus Martinez Silva. Maestria en Cómputo aplicado. Universidad Autonoma de Ciudad Juarez. E-mail: jesus_mtz_s@live.com.*
- *Dr. Jorge Rodas Osollo Universidad Autonoma de Ciudad Juarez. E-mail: jorge.rodas@uacj.mx*

DERECHOS DE AUTOR Y DERECHOS CONEXOS, año 1, No. 1, Agosto - diciembre 2017, es una Publicación semestral editada por el Clúster de Inteligencia Artificial (Ai Cluster), Bulevar Tomás Fernández 8945, Parques Industriales, 32470 Cd Juárez, Chih. Tel. (656) 2027902, https://aicluster.org/ Editor responsable: Jorge Rodas Osollo. Reserva de Derechos al Uso Exclusivo No. 00-2018-00000000-000, ISSN: 0000-0000, ambos otorgados por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este Número, Ai Cluster, Victor Hernández, Bulevar Tomás Fernández 8945, Parques Industriales, 32470 Cd Juárez, Chih., 1 de enero de 2018.

los relacionados con su tema de investigación, y este problema con el paso del tiempo se vuelva cada vez más complicado, ya que el número de publicaciones va en aumento constantemente.

Sin embargo estas grandes cantidades de conocimiento no deben de verse como un problema, al contrario, gracias a estas grandes cantidades de artículos podemos obtener importantes datos de comportamiento, de ausencia o de explotación de temas entre muchas otras tendencias de interés estadístico.

3 Bases y Fuentes de Conocimiento

Existen múltiples fuentes de bibliotecas virtuales de acceso abierto en web, de los cuales se puede acceder a los artículos en distintos formatos descargándolos directamente de la web que proveedor del servicio, pero existe una iniciativa en la cual se centrara el desarrollo de este proyecto: OAI-MPH Open archives Initiative Protocol for Metadata Harvesting, el cual es un protocolo usado para la extracción de metadatos de repositorios públicos de información de artículos científicos, estos metadatos son compartidos en archivos XML usando este servicio web, de esta manera se podrá acceder de una manera más amplia a las bases de datos publicas sin descargar los documentos enteros de las publicaciones.

Para la realización de pruebas utilizamos la base de datos de la web de aminer.org en su versión V4 [2], esta web publica sets de datos de documentos científicos recopilados a su vez de repositorios públicos con el propósito de su uso en investigaciones. El set de datos utilizado fue: "*DBLP-Citation-network V4: 1, 511,035 papers and 2, 084,019 citation relationships (2011-01-08)*" en donde los datos están estructurados en texto plano divididos por indicadores de propiedades.

Tabla 1.

DBLP-Citation-network V4: 1,511,035		
# de Papers	**# de Citas**	**Promedio de año de publicacion**
1,511,035	2,084,019	2001.7

4 Redes de Conocimiento

Con el alto crecimiento de publicaciones científicas y su correcta documentación, en la actualidad encontramos millones de artículos disponibles en librerías electrónicas con acceso al público en general, los cuales individualmente cada uno referencian una lista de artículos que representan el conocimiento base usado durante su desarrollo.

De esa manera se han ido creando enlaces entre todas las publicaciones realizadas, algunos autores ganando mucho más conexiones gracias la importancia o novedad del tema y algunos otros quedando un poco pobres de enlaces, esto es a lo que llamamos redes de conocimiento.

Utilizando estas redes de conocimiento y filtrándolas por mente los artículos con mayor relevancia en cuando a número de referencias de otros artículos hacia él, tal y como se muestra en la tabla 2.

Tabla 2.Artículos con mayor número de referencias en la base de datos de aminer.org usada de prueba.

Titulo	#Referencias
C4.5: Programs for Machine Learning	360
Modern Information Retrieval	230

Pero si visualizáramos esta red de conocimiento resultante no solo obtendríamos un ranking ordenado por número de referencias, si no que podríamos ver que artículos se encuentran cerca o más lejos de la nube con mayor interés o donde se está desarrollando con mayor fuerza el tema y así poder discriminar aquellos artículos que a pesar de su número de relaciones no se encuentra tan cerca de aquellos que si son parte primordial de la red. Asi mismo podremos ver que con el paso del tiempo el desarrollo de estas redes va tomando nevos horizontes para crear unevos temas de investigacion o expandir los actuales.

5 Herramienta

Para poder demostrar la existencia de las redes de conocimiento y al mismo tiempo poder ofrecer una utilería de búsqueda de artículos científicos usando las redes de conocimiento para la toma de decisiones, hemos desarrollado una herramienta la cual contiene la base de datos de aminer.org que mencionamos anteriormente. El desarrollo del sistema es web utilizando Python 3.6 como el lenguaje que se encargara de la parte del proceso de los datos y la parte grafica será mostrada usando la librería de Python plotly.py, la cual nos ayudara a mostrar las redes de conocimiento por medio de grafos 3D, manipulacion del mismo, zoom y la posibilidad de ver el nombre del nodo al situar el puntero sobre el. Esta ultima librería aparte de encargarse de los graficos 3D, tambien nos ayuda a ordenar los nodos en el espacio de tal manera que puedan ser visualizados de la mejor forma y no ocupen el mismo espacio varios nodos, sin embargo se usara la capacidad de categorizacion de los nodos en base a colores para mostrar las distintas redes de conocimiento que se vallan graficanto.

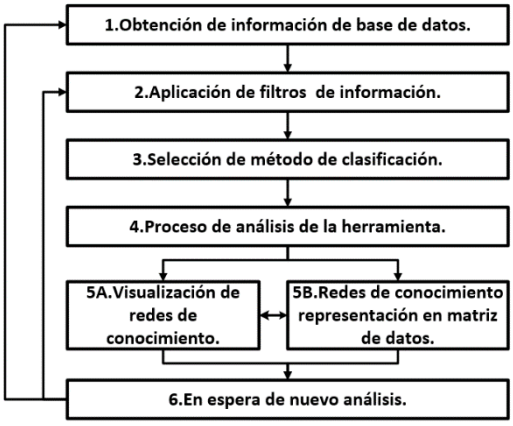
5.1 Modelo.

El proceso de desarrollo esta definido en 6 pasos, iniando con la obtencion de datos ya sea desde la base de datos local o alguna fuente de datos externa de algun proveedor de datos usando el protocolo OAI-MPH (*DBLP-Citation-network V4: 1,511,035* para esta investigacion). Se continúa con la aplicación de filtros de informacion que son especificados por el usuario al momento de requerir modelar algun tema o campo de investigacion.

DERECHOS DE AUTOR Y DERECHOS CONEXOS, año 1, No. 1, Agosto - diciembre 2017, es una Publicación semestral editada por el Clúster de Inteligencia Artificial (Ai Cluster), Bulevar Tomás Fernández 8945, Parques Industriales, 32470 Cd Juárez, Chih. Tel. (656) 2027902, https://aicluster.org/ Editor responsable: Jorge Rodas Osollo. Reserva de Derechos al Uso Exclusivo No. 00-2018-00000000-000, ISSN: 0000-0000, ambos otorgados por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este Número, Ai Cluster, Victor Hernández, Bulevar Tomás Fernández 8945, Parques Industriales, 32470 Cd Juárez, Chih., 1 de enero de 2018.

estamos buscando y se limitan los datos que se mostraran en el resultado, por ejemplo:

- Metodo de generacion de nodos o matriz de resultados
  - Excluir nodos individuales
  - Mostrar red de conocimiento de mayor rango.
  - Excluir nodos con relaciones menores a x.
- Metodo de relacion entre nodos.
  - Citas directas.
  - Uniones Bibliograficas.
  - Co-referencias.

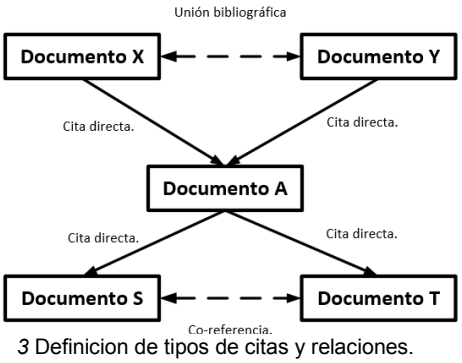


2 Modelo de ejecucion de herramienta.

Se procede a procesar los datos requeridos y metodos selecciónado en donde se evaluaran los nodos y sus relaciones limitando la informacion en base a lo establecido por el usuario y entregar el modelo visual o matriz de resultados.
Para finalizar el sistema esperara si continúa con un nuevo analisis o realiza un segúndo analisis en los resultados obtenidos.

6 METODOLOGIA.

Para identificar los nodos y relacionones existentes utilizaremos la definicion de tipos de citas tal y como las nombra (Kose T. & Sakata I. 2018) [4]. En donde las "citas directas" son las que se encuentran dentro de las referencias de cada artículo, las "uniones bibliograficas" son aquellos documentos X, Y que tienen una cita directa a un documento A, entre ellos existe una union bibliografica, por ultimo tenemos las "Co-Referencias", asi llamadas las relacion que existe entre los listados de documentos referenciados S y T en un documento inicial A tal y como se muestra en la figura 4.



3 Definicion de tipos de citas y relaciones.

6.1 Detección de comunidades.

Antes de que se use un método de análisis de las relaciones existentes en la red, es necesario eliminar los artículos que no contienen ninguna cita directa de o hacia otro artículo y solo la información con relaciones a otros nodos es tomada para el análisis.
El método que usaremos para la identificación de los nodos y relaciones dentro de la base de datos que se analiza para la obtención de las redes de conocimiento será por medio de la detección de comunidades en estas redes sociales.
Para la analizar la estructura de comunidades se pueden utilizar varios métodos que nos pueden brindar exactitud, tiempo o calidad, pero para cada caso nos puede convenir utilizar uno u otro ya que no existe uno que brinde el mejor costo de cómputo y que nos dé el mejor resultado [8].
Uno de los métodos más aceptados y del cual aremos uso es el método de Modularidad Q presentado por Newman y Girvan [9]:

$$Q = \sum_{r} (e_{rr} - a_r^2)$$

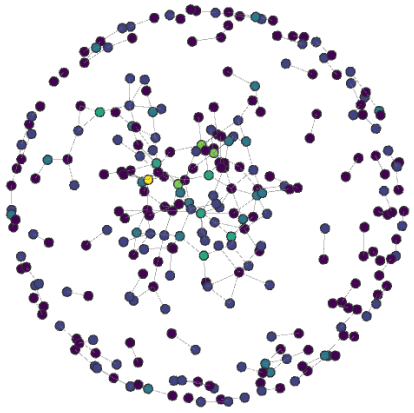
4 Formula de Modularidad Q Newman y Girvan.

Donde err son las fracciones de enlaces que conectan dos nodos dentro de una comunidad r , ar es la fracción de enlaces que tienen uno o ambos vértices dentro de la comunidad r, y la suma se extiende a todas las comunidades en r en dada red. La modularidad es un criterio para evaluar la calidad de dividir una red en clusters [8].
Con el uso de la modularidad Q de Newman y Girvan se identificaran las comunidades en donde sus enlaces son más densos y en donde son muy débiles, para así generar clusters que delimitaran ciertos grupos de relaciones dándonos como resultado las redes de conocimiento. El tipo de relación usado, ya sea referencias directas, uniones bibliográficas o co-referencias influyen directamente en el resultado pero estas opciones son a elección del usuario.

DERECHOS DE AUTOR Y DERECHOS CONEXOS, año 1, No. 1, Agosto - diciembre 2017, es una Publicación semestral editada por el Clúster de Inteligencia Artificial (Ai Cluster), Bulevar Tomás Fernández 8945, Parques Industriales, 32470 Cd Juárez, Chih. Tel. (656) 2027902, https://aicluster.org/ Editor responsable: Jorge Rodas Osollo. Reserva de Derechos al Uso Exclusivo No. 00-2018-00000000-000, ISSN: 0000-0000, ambos otorgados por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este Número, Ai Cluster, Victor Hernández, Bulevar Tomás Fernández 8945, Parques Industriales, 32470 Cd Juárez, Chih., 1 de enero de 2018.

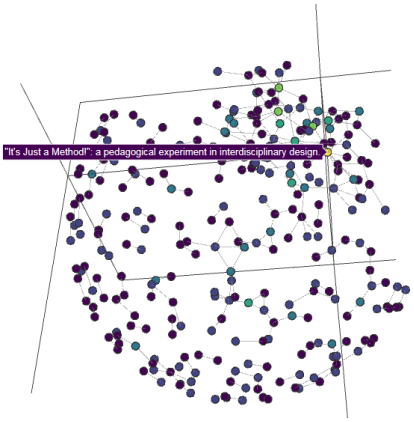
7 Pruebas

Usando la base de datos de aminer.org (*DBLP-Citation-network V4)* se realizaron una serie de pruebas en donde se generaron grafos con las relaciones registradas en las referencias directas. La cantidad de datos graficada fue limitada a 800 nodos con referencias, esto debido a que la cantidad de nodos a graficar por la librería de python plotly se limita a cierta cantidad saturando el grafo de información.



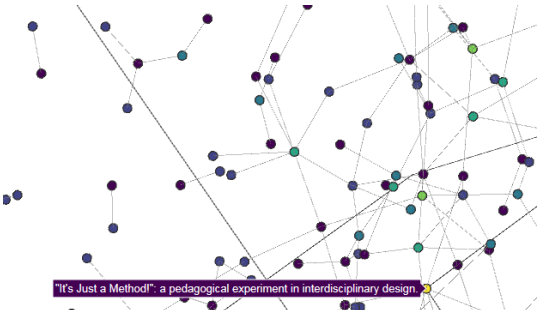
5 Grafo de relaciones DBLP-Citation-network V4 limitada a 800 nodos relacionados (vista 3D).

Gracias a las propiedades de la librería utilizada se pueden realizar análisis interactivo de la red girando y magnificando el grafo según se requiera y utilizando el puntero de la computadora ver el nombre de los artículos referenciados en cada nodo.



6.Grafo de relaciones DBLP-Citation-network V4 limitado a 800 nodos relacionados (vista 2).

Durante el proceso de los 800 nodos relacionados se identificó en base a clases mediante colores la cantidad de referencias directas encontradas en cada nodo, de este modo, los nodos de color más oscuro contienen menos referencias que los colores más brillantes, en este ejemplo podemos encontrar que el artículo de nombre:"It's Just a Method!": a pedagogical experiment in interdisciplinary design", es el nodo con mayor número de relaciones.



7.Grafo de relaciones DBLP-Citation-network V4 limitado a 800 nodos relacionados (Vista de nodo con mayor número de referencias).

8 Conclusiones

La limpieza de nodos, identificación de relaciones y clasificación por número de referencias fue posible gracias al uso de la herramienta descrita y nos muestra las capacidades para la implementación de los métodos de identificación de comunidades en redes sociales que se quieren utilizar, como lo es el método de modularidad Q descrito en esta investigación.

Referencias

[1] SciELO, hemeroteca virtual conformada por una red de colecciones de revistas científicas en texto completo y de acceso abierto y gratuito: https://analytics.scielo.org/w/publication/article 2018.

[2] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). Pp.990-998.

[3] Tu, Y. N., & Seng, J. L. (2012). Indices of novelty for emerging topic detection. Information processing & management, 48(2), 303-325.

[4] Kose, T., & Sakata, I. (2018). Identifying technology convergence in the field of robotics research. Technological Forecasting and Social Change.

[5] Hashimoto, M., Kajikawa, Y., Sakata, I., Takeda, Y., & Matsushima, K. (2012). Academic landscape of innovation research and National Innovation System policy reformation in Japan and the United States. International Journal of Innovation and Technology Management, 9(06), 1250044.

[6] Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. Research Policy, 43(8), 1450-1467.

[7] Small, H. G. (1977). A co-citation model of a scientific specialty: A longitudinal study of collagen research. Social studies of science, 7(2), 139-166.

[8] Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005(09), P09008.

[9] Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), 026113.

DERECHOS DE AUTOR Y DERECHOS CONEXOS, año 1, No. 1, Agosto - diciembre 2017, es una Publicación semestral editada por el Clúster de Inteligencia Artificial (Ai Cluster), Bulevar Tomás Fernández 8945, Parques Industriales, 32470 Cd Juárez, Chih. Tel. (656) 2027902, https://aicluster.org/ Editor responsable: Jorge Rodas Osollo. Reserva de Derechos al Uso Exclusivo No. 00-2018-00000000-000, ISSN: 0000-0000, ambos otorgados por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este Número, Ai Cluster, Victor Hernández, Bulevar Tomás Fernández 8945, Parques Industriales, 32470 Cd Juárez, Chih., 1 de enero de 2018.