Analítica computacional y desarrollo de colecciones digitales

Dr. Valentino Morales López
Mtra. Zaira Lagunas

Introducción

El presente documento tiene por objeto plantear algunas reflexiones sobre el apoyo que puede brindar la analítica computacional para el desarrollo de colecciones. Si se revisan las tendencias es abrumadora la cantidad de información que se está generando y el bibliotecario es consciente de que está en una carrera que de antemano ya perdió, porque nunca podrá adquirir todo lo que se publica. Es absurdo llegar a aseverar que en un solo espacio se puede almacenar y poner a disposición de los usuarios todas las publicaciones generadas incluso en una sola área temática. Lo máximo a lo que se puede aspirar es a contar con redes que comparten recursos de información.

Hay que recalcar que la metodología que se presenta no busca sustituir a los métodos existentes en bibliotecología para el desarrollo de colecciones, sino la intención es complementarlos, para que el profesional de la información tenga otras opciones para efectuar su trabajo de forma eficiente. La metodología que se usa es el de análisis de términos mediante el análisis de redes sociales, a través de su recurrencia y su ubicación en artículos con determinada cantidad de citas. La metodología se aplica a dos ejemplos, en el primero es si se hiciera para el desarrollo de colecciones de INFOTEC, en el segundo es para conocer las tendencias en el tema de repositorios bibliotecarios.

El presente trabajo tiene dos partes: 1. Analítica computacional y desarrollo de colecciones, en la que se presenta el marco de referencia metodológico de la analítica computacional para el desarrollo de colecciones; 2. Ejemplo 1 de uso de análisis de redes sociales en TIC, en el que se muestra un pequeño ejemplo de la forma en la que se puede aplicar la analítica computacional para analizar términos derivados de la referencia y resumen

de los artículos con mayor cantidad de citas en la Tecnología de la Información y Comunicación (TIC) localizados en Web of Science; 3. Ejemplo 2 de uso de análisis Análisis del tema consorcio de bibliotecas en base Web of Science.

1.- Analítica computacional y desarrollo de colecciones

La información está en un proceso de crecimiento inusitado, gran parte de ello se debe a la influencia de las TIC en el crecimiento de las publicaciones científicas. Sumado a lo anterior, también se tiene la masificación de las bases de datos con diversas temáticas y que se constituyen en un elemento de intermediación entre los científicos, autores y públicos concretos.

Al mismo tiempo, la ecuación, desarrollo de colecciones y TIC dio por resultado la aparición de herramientas que comúnmente eran útiles para los profesionales de la información y documentalistas que deseaban medir el crecimiento de recursos informativos existentes en el contexto científico. No obstante, en la actualidad esas herramientas, basadas sustancialmente en la estadística son limitadas para identificar el considerable crecimiento de documentos, información y datos para dar oportunidad a que los expertos de dichas áreas conozcan las mejores publicaciones que se albergan en el ámbito digital.

En la bibliotecología ha existido una vertiente que requiere la evaluación de la calidad de las publicaciones periódicas para una adecuada adquisición de recursos de información, denominada desarrollo de colecciones. Existen diversos esfuerzos para lograr que la selección y adquisición de los recursos de información se realice con base en procesos bibliométricos, estadísticos y sociológicos. Con la finalidad de desarrollar una colección que verdaderamente contribuya a los requerimientos de la comunidad a la que la biblioteca está destinada. En ese sentidos, en actuales tiempo de los buscadores de la Web, el valor agregado que debe ofrecer la biblioteca es una colección compuestos por libros, revistas y bases de datos convenientemente seleccionada y organizada, que aporte a las necesidades de información que tiene sus usuarios.

Esa actividad podría desarrollarse adecuadamente al emplear

los métodos propuestos por la bibliometría/cienciometría. Sin embargo, esa área hace análisis de la producción científica con dos objetivos, que no necesariamente se vinculan al desarrollo de colecciones. El primero, de acuerdo con Pritchard (1969) se enfoca en el ámbito cuantitativo en función de la producción, diseminación y uso de la información (libros, documentos, revistas, artículos y autores). En una aproximación reciente, McGrath (1989) considera que el estudio se enfoca en la producción del conocimiento mediante las redes de colaboración y comunicación científica. A pesar de que los objetivos de la bibliometría no son cercanos al desarrollo de colecciones, las actividades anteriormente enunciadas tienden a entrelazarse para establecer un mejor panorama de la producción científica y de esa forma establecer cuáles son los ámbitos en los que debería hacerse un desarrollo de colecciones racional.

La bibliotecología al ser una disciplina que de forma integral aglutina teorías y métodos provenientes de diversas disciplinas le permite asociarse de una manera más sencilla con otros campos temáticos para fortalecer teorías y sistemas que vinculados con la bibliometría/cienciometría sirven de apoyo a otras áreas del conocimiento, como la analítica computacional, que proviene de la computación y es usada para analizar grandes cúmulos de datos. Esta metodología emergente regularmente se asocia con el desarrollo de herramientas para la visualización de datos. Sin embargo, la forma en la que se entiende la analítica computacional en este trabajo es al proceso que va desde la recolección de datos, su compilación, organización, procesamiento y análisis mediante algoritmos de alto desempeño y una infraestructura de cómputo robusta.

La cuestión es que la bibliometría/cienciometría se basa en una metodología de corte estadístico, la que busca determinar el tamaño, crecimiento y distribución de la bibliografía científica. Esta área de la bibliotecología, también busca conocer cómo se produce, transmite y utiliza la ciencia en determinadas estructuras sociales. De tal forma que como parte de los objetos de estudio que se asocian con ambas áreas metodológicamente se busca demostrar una hipótesis “A”, en la que se indica que determinado autor, revista o área de dominio cuenta con una mayor cantidad de publicaciones o citas. Una posible hipótesis

"B" es demostrar las redes científicas a las que pertenece determinado autor. Indiscutiblemente es factible derivar otro tipo de análisis, pero dado que la metodología básica es estadística en un proceso inductivo-deductivo, restringido a un limitado número de variables predeterminadas, es complejo visualizar otro tipo de información o datos y las relaciones que darían otra perspectiva al análisis de la producción científica. En el gráfico que se muestra a continuación se puede observar la forma en la que se maneja el uso de variables que de manera selectiva se enfoca a una parte de todo el cúmulo de información a la que se podría tener acceso.

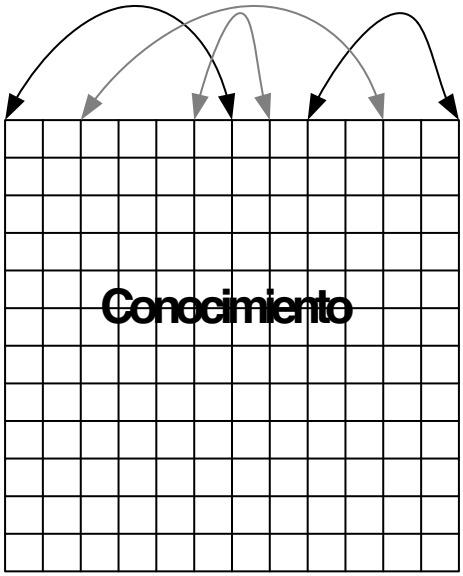


Fig. 1 Análisis estadístico de cúmulos de datos.

De este modo se entenderá que por eso la orientación de la bibliometría es descriptiva ya que regularmente los resultados de sus estudios concluyen con la afirmación de que cierto autor cuenta con un mayor número de publicaciones o citas, entonces la conclusión está preparada de antemano. Esto último es el cuestionamiento clásico a la premisa base de la denominada ciencia positivista: "si la realidad no se ajusta a la teoría, entonces pobre de la realidad".

Ahora bien, buscando hacer un replanteamiento de la bibliometría

integrando otras herramientas analíticas no limitadas por la estadística, porque de esa forma las muestras que arrojan las bases de datos podrían llegar a ser predictivas de tendencias en la investigación. Este replanteamiento permitiría modificar la limitación que se tiene en la actualidad y que solo es el reflejo en su totalidad de los índices de producción, distribución y uso de los recursos informativos.

Por lo tanto, lo que aquí se propone es complementar la aplicación de la bibliometría/cienciometría en apoyo al desarrollo de colecciones. Ya que la información disponible en el entorno digital resulta ser inmensa y muchos de los datos arrojados pueden ser erróneos, incluso hasta manipulados lo que provoca que la investigación discurra en los mismos derroteros y se genere una realidad equívoca que servirá de parámetro para la toma de decisiones.

En este sentido, será necesaria la implementación de la minería de datos, particularmente la minería de texto. Esta metodología permite ubicar comportamientos atípicos, correlaciones y patrones de los datos mediante técnicas predictivas que apoyan la toma de decisiones. De hecho, ante un cúmulo excesivo de datos difícilmente se podría efectuar una exploración basada en estadísticas de los mismos para emitir un pronóstico.

Asimismo, hay que tener en cuenta que para el análisis de grandes cantidades de información es necesario contar con un procesamiento de cómputo de alto desempeño. Esto se debe a que un equipo de cómputo convencional no tendrá la velocidad para un análisis adecuado. En ese sentido, una buena alternativa es tener métodos de almacenamiento de cómputo distribuido que optimiza el espacio de reserva, pero permite tener a disponibilidad del análisis los datos y la información y no se quedan oculto, por la comprensión.

Dicho lo anterior se enlistan los siguientes elementos para efectuar el proceso:

1) Creación de modelos: el propósito es la ubicación de la información histórica sobre la cantidad de artículos en un área del conocimiento, autores, revistas, países, etcétera. Entonces

se crean dos grupos:

i) Los artículos con una mayor cantidad de citas
ii)Los autores con mayor cantidad de citas.

Mediante ciertos algoritmos, como las redes neuronales, árboles de decisiones, y análisis de regresión binaria, se establecen los perfiles de las diferencias entre ambos grupos. Este proceso permitirá tener ciertas características de los artículos en esas áreas del conocimiento, lo que ayudará al establecimiento de patrones que facilitarán el análisis.

El resultado de esta primera fase serán varios modelos, de los que no se puede afirmar que uno sea mejor que otro, ya que han sido desarrollados de diferentes maneras.

2) Selección de los mejores modelos: una vez que se identificaron diversos modelos, hay que proceder a la identificación del mejor modelo, en términos de los requerimientos que tiene el analista y las bases de datos a analizar. La identificación del mejor modelo, es con el propósito de que se establezca con cuál se podría dar los mejores resultados al usarlo en el análisis de una base de datos. Hay que tener en cuenta que esa nueva base de datos debe tener las mismas características que la base de datos en la que se desarrolló el modelo seleccionado, porque de otra forma el análisis no será apropiado.

El proceso de selección garantiza que el modelo seleccionado permita una mejor descripción de la base de datos en la que fue desarrollado, sino que también puede generalizarse y aplicarse a otras bases de datos.

3) Uso del modelo seleccionado: La forma en la que se pone a prueba el modelo es haciendo que establezca si los perfiles de cada una de las bases de datos nueva a las que analice tenga la información y los datos sobre las publicaciones. Posteriormente se relaciona esa información con los resultados de las bases de datos históricas, los resultados permitirán ver cuál de los modelos plantean las tendencias generales de acuerdo al área de conocimiento analizada.

Cuando se llega a esta etapa del proceso de minería de datos, ya se comienza a interpretar los modelos. De acuerdo a los diversos tipos de algoritmos será la forma en la que se puedan interpretar los modelos y los datos. De cualquier manera esta fase permite generar conocimiento sobre los problemas o alternativas que tiene el objeto estudiado, como lo ilustra la siguiente imagen.

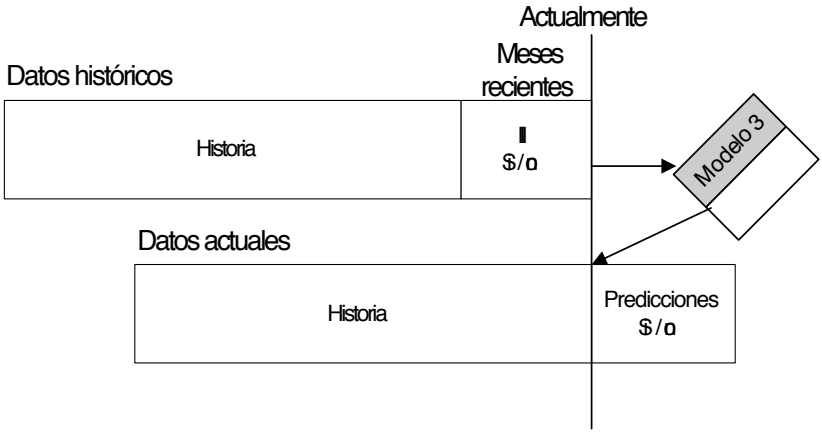


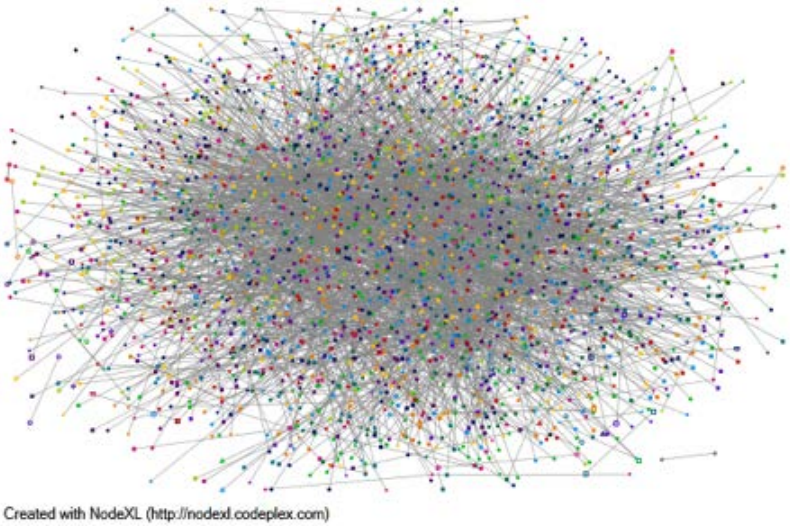
Fig. 2 Análisis histórico de datos

Una vez que se ha hecho una explicación sucinta de la minería de datos, en el siguiente apartado se desarrollará un ejemplo para el uso de analítica computacional en el análisis de información bibliométrica para el establecimiento de tendencias temáticas.

2.- Ejemplo de uso de análisis de redes sociales en TIC

Se parte del supuesto de que el encargado del desarrollo de colecciones en INFOTEC necesita determinar cuáles son los temas claves y tendencias en TIC. Por lo tanto se recurre a Web of Science para efectuar una búsqueda de la cadena de términos Information Communication and Technology. Una primera búsqueda arroja resultados con diversas temáticas y tipos de publicaciones, se hacen las respectivas limitaciones a la búsqueda y se obtienen cinco mil referencias, de las que se seleccionan las de mayor cantidad de citas que es un total de 500 artículos. A Web of Science se le pide un archivo en txt con la referencia y el resumen de esos 500 artículos. A esas referencias y su resumen se les hace un análisis de texto, basado en redes sociales, que toma en cuenta dos cuestiones:

recurrencia de los términos y su ubicación en los artículos con mayor cantidad de citas. Una vez hecho ese análisis el sistema NodeXL arroja la siguiente red:



Il 2 Sociograma de los artículos mas citados sobre TIC en Web of Science

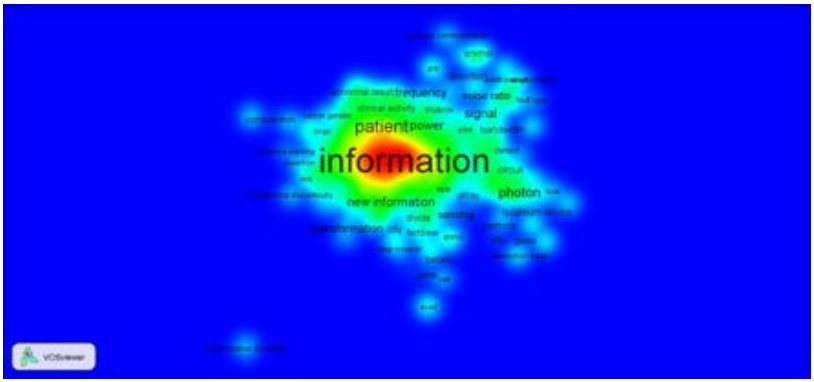
Obviamente el análisis de la anterior red es complicado, aunque gráficamente es sumamente interesante. Sin embargo, sus métricas se consignan en la siguiente tabla:

Métrica del Gráfico	Valor
Tipo de Gráfico	Indirecto
Vértices	1611
Relaciones Únicas	27209
Relaciones con Duplicados	0
Total de Relaciones	27209
Autorelaciones	0
Componentes Conectados	1
Componentes conectados con una sola relación	0
Máximo de Vértices en un Componente Conectado	1611
Máximo de Relaciones en un Componente Conectado	27209
Maximum Geodesic Distance (Diameter)	8
Average Geodesic Distance	3.997956
Graph Density	0.020980757
Modularity	0.894764

Tabla 1 Métricas de la red de artículos sobre TIC con mayores cantidades de citas en Web of Science

Entre los datos a poner en relevancia es que son 1611 los términos identificados y 27209 las relaciones entre esos términos. A esto se hace referencia cuando se plantea que son varios datos que un sistema no podría procesar de forma rápida y acertada.

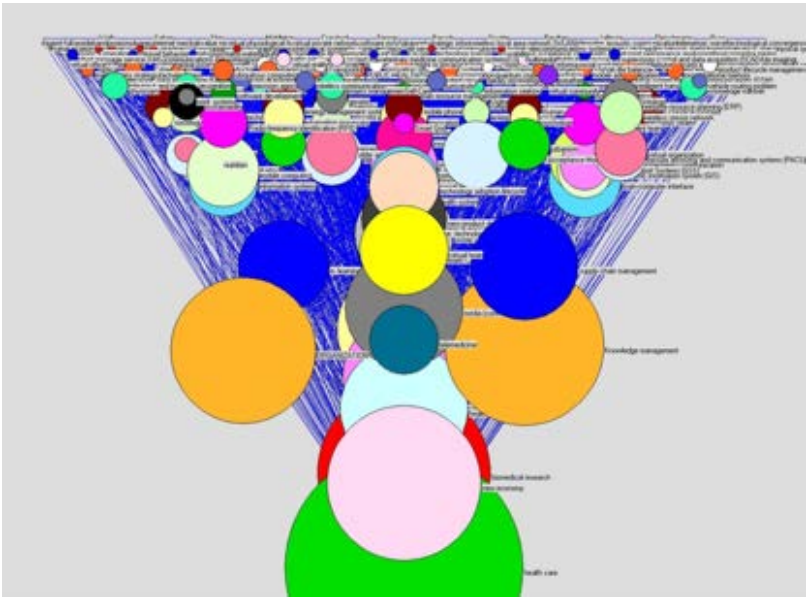
El siguiente proceso que se llevó a cabo fue el de la visualización calórica, por esos los colores de rojo a frío de la relevancia de los términos. Esto permitió obtener una imagen en la que mediante closteo se visualizan los temas vinculados a la cadena de términos bajo la que hizo la búsqueda inicial.



II. 3 Imagen calórica de los clusters temáticos de los artículos sobre TIC mas citados en Web of Science

En la imagen resalta el término información que se podría decir es obvio, pero no es así, porque no aparece de forma relevante technology o communication. Otro término al que hay que prestar atención es el de patient, cuya relevancia va en función de que por un lado varios de los sistemas de información están enfocados al área de la salud, sobre todo de tipo preventiva. De hecho en INFOTEC se están llevando a cabo proyectos que están orientados al soportar con TIC a área médica. El otro término que salta a la vista es el de photon que está vinculado al desarrollo de hardware de alto desempeño, al respecto habría que prestar atención a término quantum device, porque una de las tendencias en desarrollo de equipo de cómputo son las computadoras cuánticas que tienen mayores capacidades y ofrecen mejor seguridad. El ejemplo es el desarrollo de la supercomputadora cuántica en China.

De acuerdo a lo que se ha comentado ya, esta imagen derivada de un procesamiento analítico ofrece varias posibles interpretaciones sobre los temas de frontera en la TIC. Sin embargo, continua siendo limitado, por esa razón en Pajek se agrupan los temas por jerarquía y se obtuvo como resultado la siguiente imagen.



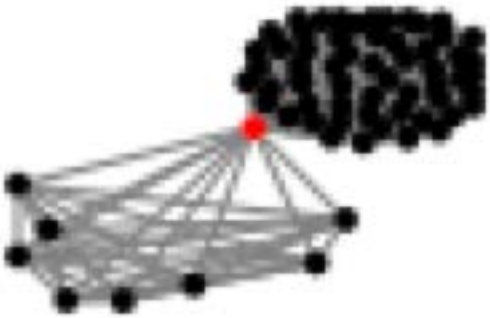
II 4 Red jerarquizada de los artículos sobre TIC más citados en Web of Science

En la imagen se resaltan los términos de mayor recurrencia y difieren de los apuntados previamente. Por ejemplo information, no aparece, pero si health care que iría conforme a lo comentado previamente, sobre la relevancia que tiene el tema de la salud preventiva para la investigación en TIC. Sin embargo, aparece un asunto que no era relevante en la imagen calórica, new economy, éste en el caso de INFOTEC es relevante, porque el centro no solo está interesado en el desarrollo de software o infraestructura, también en las implicaciones sociales y económicas que tiene la TIC. En un tercer nivel aparecen dos términos que son interesantes, organization y knowledge management, la razón es que los actores sociales que están interesados en incorporar la TIC a sus procesos son las organizaciones y por otra parte en la administración del conocimiento se reconoce que la TIC es un elemento sustancial para la administración de conocimiento

para las organizaciones.

Ahora bien, ya se mencionó el análisis macro del grupo de datos que se recuperó, a continuación se planteará el análisis a detalles de dos términos. En este caso se seleccionaron esos dos términos porque uno tiene el mayor EigenVector y el otro tiene mayor centralidad.

El EigenVector en un análisis de redes permite identificar los términos que son puente entre grupos o vértices de una red El término seleccionado es equipment con un grado de 60 de EigenVector. En la siguiente ilustración se presenta con que agrupaciones tiene relación ese término.

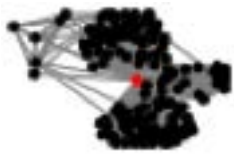


Il 5 Término Equipment, presentación de su EigenVector

La razón de la relevancia del equipamiento para la TIC es porque éste permite que los procesos de cómputo se lleven a cabo. Además, a mayor cantidad de información es necesario contar con equipo más robusto en términos de velocidad de procesamiento, almacenamiento y ahorro de energía. Además, se debe tomar en cuenta que los programas pueden ayudar en la optimización del aprovechamiento del equipo de cómputo. En ese sentido, para INFOTEC cobra relevancia prestar atención a los nuevos desarrollos de equipos, ya sea para una selección adecuada de los mismos o si es posible para el desarrollo de software que permita aprovecharlos al máximo.

Ahora bien, si el interés recae en conocer cuál es el término de mayor centralidad, que quiere decir el que tiene la mayor cantidad de relaciones con otros vértices de la red. El término

que se identifica es information processing del cual se obtuvo un nivel de 145 y su red se constituye de la siguiente forma:



Il 6 la red semántica del término Information processing

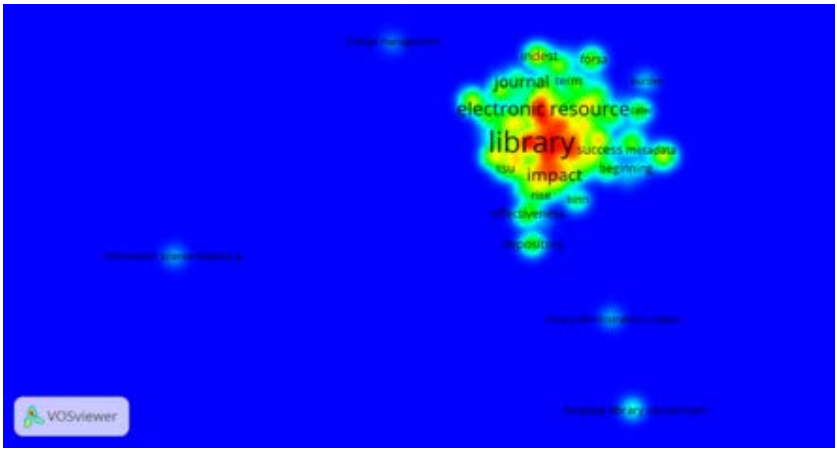
En resumen se identifica que el tema con mayor cantidad de relaciones con los otros términos es information processing, cuestión no menor, por la importancia que tiene el análisis de mayores cantidades de información y datos. Este tópico no sólo debe referirse a software, también a hardware, por eso su centralidad.

A fin de concluir el ejemplo que se mostró, se puede decir que para orientar el desarrollo de la colección en INFOTEC éste es un insumo valioso, que junto con otros criterios ayudarán a orientar la selección de los recursos de información para este centro de investigación.

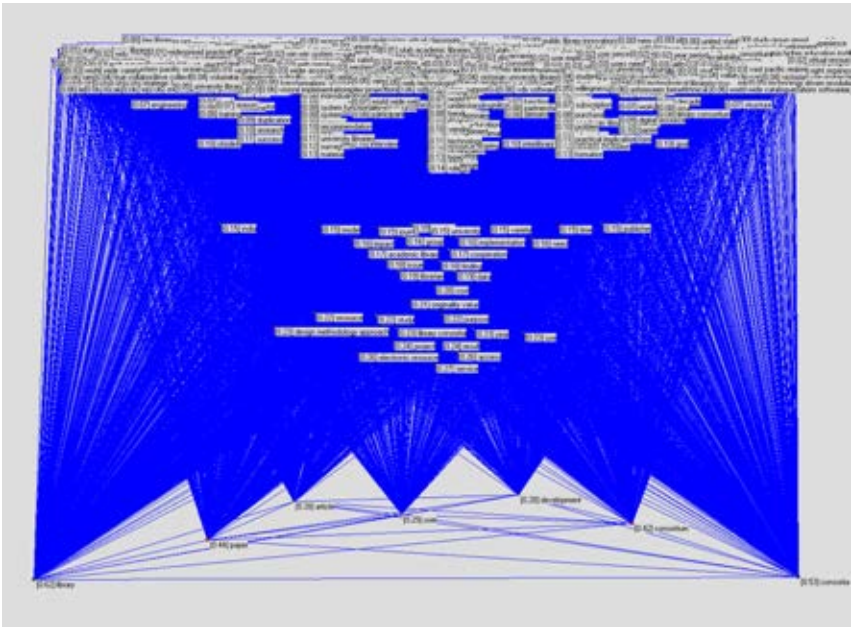
3.- Ejemplo 2 de uso de análisis Análisis del tema consorcio de bibliotecas en base Web of Science

A pesar de que el artículo de Cuadra es desde 1972, plantea varias actividades que son vigentes en la actualidad y a las que los consorcios tendrían que prestar atención para establecer una agenda de trabajo. No obstante, hay que señalar un asunto importante, desde hace 40 años los consorcios bibliotecarios han evolucionado. En consecuencia se hizo una búsqueda en Web of Science de todo el siglo bajo la frase "library consortia", el primer resultado fue de 1074 documento, posteriormente se restringió a la temática Information Science Library Science y se tuvieron 473 registros. Esos registros se depuraron de acuerdo a dos tipos de criterios: los registros con mayor número de citas y los registros recientes para poder contrastarlos. A los registros seleccionados, mediante análisis de texto y redes se les analizó el título y el resumen.

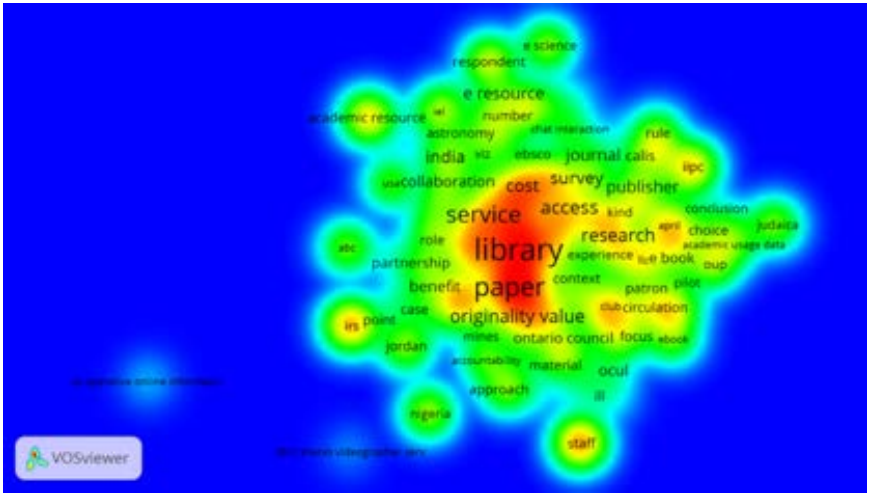
En el primer grupo de registros, los que tenían de una a más citas, se obtuvieron 87 registros y el análisis de closteo calórico arrojó el siguiente resultado:



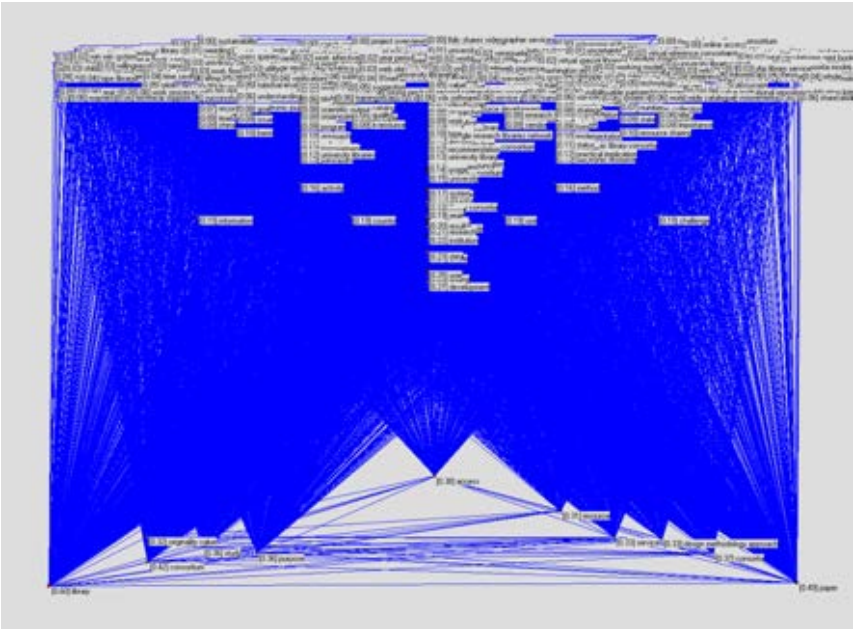
En un análisis de redes sociales se obtuvo el siguiente resultado jerárquico:



El otro análisis fue de los documentos publicados los últimos 10 años, se obtuvieron 63 registros y de su análisis se elaboró el siguiente gráfico:



En el análisis jerárquico se tuvo el siguiente gráfico:



La comparativa de ambos resultados arroja lo siguiente:

a) La Investigación está alejada de los temas que afectan la cotidianidad del trabajo de los profesionales de la información o bibliotecarios.
b) Hay mayor agrupación en temas, en el caso de los artículos con

mayor número de citas, mientras que en los que se publicaron en los últimos 10 años tiene mayor dispersión.
c) En lo referente a los temas de mayor cantidad de citas son relevantes los recursos electrónicos, revistas, el impacto. Mientras que en las que se han publicado en los últimos 10 años son el servicio, el costo, el acceso.
d) En una comparativa que se muestra en la siguiente tabla, de los 12 términos de mayor relevancia en el análisis previamente hecho. Resulta que los términos que aparecen en los artículos con mayor cantidad de citas y en el otro no son user, article, electronic resource, Project y result. Mientras que en los publicados en los último 10 años son purpose, study, desisgn methodology approach, originality value y resource. De donde los términos que sí se encuentra en ambos casos son library, consortia, paper, consortium, development, service y access. Es posible incluir electronic resource y resource.

Con mayor cantidad de citas	Publicados los últimos 10 años
library	Library
consortia	Paper
paper	consortium
consortium	consortia
user	purpose
article	study
development	service
service	design methodology approach
electronic resource	originality value
access	resource
project	access
result	development

Por lo tanto, de forma aventurada se podría aseverar que los temas predominantes en los consorcios bibliotecarios (descartando los términos library, consortia y consortium) son artículo (paper), desarrollo (development), servicio (service) y acceso (acccess). Esto lleva a cuestionar que la adquisición de recursos de información sea la acción predominante de un consorcio, cuando existen temas que parecen ser de mayor relevancia, como el servicio y el acceso.

Conclusiones

La cantidad de información que se genera es excesiva y no es posible una adquisición de todo lo que se publique, además de resultar elevada la compra de membresías y recursos para los diferentes centros de información. Además, hay que prestar atención que la biblioteca no solo se trata de un centro de servicio, también es una colección que es la construida de acuerdo a las necesidades de la comunidad a la que sirve.

Por otra parte, resulta indispensable generar políticas de adquisición precisas que se apeguen a las normas y disposiciones presupuestales determinadas, lo que necesitaría de criterios aún más definidos en función de racionalizar la producción, transferencia y consumo de la comunidad de investigadores que demande de los recursos informativos, por lo tanto, deberá forjarse un panorama realista y concreto de las futuras transacciones.

Por ende hoy en día deben utilizarse las diversas herramientas de cómputo que permitan generar un análisis previo al momento de selección y compra de recursos informativos. Además, hay que tener información histórica que ayude a hacer comparativos, porque de otra forma se corre el riesgo de cada vez inventar criterios.

Bibliografía

Cuadra, Carlos y Ruth J. Patrick (1972) Survey of academic library consortia in the U. S. College and Research Libraries (july), 271-283.

MCGrath, W. (1989) "What bibliometricians, scientometricians and informetricians study; a typology for definition and classification; topics for discussion". En: International Conference on Bibliometrics, Scientometrics and Informetrics. Ontario: The University of Western Ontario, 1989.

Pritchard, A. (1969) "Statistical bibliographic or bibliometrics". En: Journal of Documentation. 25 (4), 348-349.