Herramienta informática para vigilancia tecnológica -VIGTECH-

Technology Monitoring Software tool -VIGTECH-

Víctor A. Bucheli G., MSc, Fabio A. González O. PhD.
Universidad Nacional de Colombia, sede Bogotá Departamento de Ingeniería de Sistemas e Industrial
vabuchelig@unal.edu.co , fagonzalezo@unal.edu.co

Recibido para revisión 26 de Marzo de 2007, aceptado 15 de Junio de 2007, versión final 22 de junio de 2007

Resumen— **El artículo presenta una herramienta de software que apoya la vigilancia teconológica. La herramienta permite encontrar relaciones cognitivas y sociales en un conjunto de documentos extraidos de una base referencial tal como SCOPUS. Específicamente, la herramienta soporta las actividades de obtención de información de documentos científicos, extracción de metadatos, cálculo de estadísticas descriptivas, análisis de redes sociales, análisis de redes de palabras claves y visualización. El artículo presenta una descripción de las bases conceptuales que fundamentaron el desarrollo de la herramienta, así como una descripción de su arquitectura y funcionalidad.**

Palabras Clave— **Sociedad del conocimiento, vigilancia tecnológica, mapas científicos, aprendizaje de máquina aplicado a documentos científicos, análisis de redes sociales.**

Abstract— **The paper presents a software tool for supporting techniology monitoring tasks. The tools allows to find cognitive and social relationships in a set of documents, which have been extracted from a reference database such as SCOPUS. Specifically, the tool the following activities: gathering of information from scientific documents, metadata extraction, descriptive statistics calculation, social network analysis, keyword network análisis and visualization. the paper describes the conceptual bases that supported the development of the the tool, and describes its architecture and functionality.**

Key words— **Society of knowledge, technological monitoring, machine learning applied to scientific documents, social networks analysis.**

I. INTRODUCCIÓN

EL uso intensivo del conocimiento, el entorno globalizado y el avance en la informatización de la sociedad implica una nueva estructuración de las organizaciones, de las sociedades y del mercado[1]. Dicha estructuración, ha configurado un entorno con alto grado de incertidumbre, dinámico, de profundos cambios, donde la toma de decisiones y la gestión del conocimiento, son elementos clave para consolidar organizaciones eficientes, sostenibles, productivas e innovadoras. De está forma, se hacen necesarias prácticas sistemáticas que garanticen la estabilidad y el crecimiento económico de las organizaciones en dicho entorno [2] y que permitan basar sus procesos de producción en la incorporación intensiva del conocimiento [3], propiciando así el cambio tecnológico, la innovación y la competitividad de las organizaciones en una economía del conocimiento[1].

Una práctica que puede ser utilizada por las organizaciones para monitorear los cambios descritos, es la vigilancia tecnológica-VT-, la cual permite a una organización estar atenta al cambio de manera sistemática [4] a través del estudio permanente del mercado, del ámbito científico tecnológico, del ámbito político y del ámbito social [5]. Estas prácticas sistemáticas, principalmente las del estudio del ámbito científico tecnológico, permiten basar los procesos de producción en la incorporación intensiva del conocimiento involucrando principalmente información propia de dicho ámbito - artículos y patentes- en el desarrollo de nuevos productos o procesos. Así el ciclo de VT se compone de cuatro fases: planeación, búsqueda y captación, análisis y organización, inteligencia y comunicación [6].

La herramienta informática VIGTECH que se presenta en

1 Las sociedades a lo largo del tiempo han presentado elementos característicos que definen sus formas de producción y de construcción como una sociedad, dichos elementos que autores como [1]–[3] denominan principios de acción o principios organizadores del comportamiento humano, permiten reconocer cuáles son las formas que las organizaciones y las instituciones sociales utilizan para funcionar y responder a su entorno. Así, en este momento histórico, es la producción de conocimiento, la apropiación y el uso intensivo del mismo, el elemento clave para producir bienes, servicios y atender a las necesidades de la sociedad.

este artículo, es un instrumento para facilitar las prácticas de vigilancia tecnológica en una organización y está enfocada principalmente en el ámbito científico tecnológico, tomando como fuente de datos el servicio de información SCOPUS, apoyando así en las fases de captación y búsqueda; análisis y organización; e inteligencia. La herramienta VIGTECH automatiza los procesos de captación y búsqueda de datos mediante el modulo Crawler-VIGTECH- que permite descargar los documentos científicos de SCOPUS, y extraer características de dichos documentos, construyendo así, una base de datos relacional en la cual se almacenan estructuradamente los meta-datos del artículo y del autor. Esto permite realizar análisis descriptivos y análisis exploratorios de datos que apoyan la fase de análisis y organización. Por último, la herramienta informática VIGETCH utiliza técnicas de aprendizaje de máquina y de minería de datos apoyando así la fase de inteligencia [5], utilizando algoritmos para análisis de redes sociales [7]-[8], reducción de dimensionalidad, escalamiento multidimensional [9], agrupamiento [10]-[11] modelos gráficos probabilísticos [12], entre otros, que permiten vincular de una forma inteligible los resultados obtenidos presentando indicadores, mapas, socio gramas y en general representaciones relacionales de un tópico dado.

Así, esta herramienta permite encontrar las relaciones existentes en los documentos científicos, estas relaciones son de una parte cognitivas referidas a vínculos entre palabras clave y de otra parte relaciones sociales representadas en los vínculos de co-autoría, referenciales y de cooperación interinstitucional. Éstas relaciones en un plano general permiten reconocer en el ámbito científico tecnológico qué autores trabajan en que áreas; cuáles son las comunidades estructuralmente fuertes; oportunidades y amenazas que pueden afectar a una organización en áreas relacionadas con su campo de trabajo, cuerpos útiles de conocimiento; actores centrales y periféricos dentro de las redes sociales que se construyen en una comunidad científica [7]; posibles redes que permitan llevar a cabo proyectos conjuntos; rentabilidades; oportunidades de cambio científico tecnológico e innovaciones.

De esta forma el articulo presenta la herramienta informática VIGTECH, su funcionalidad, su arquitectura, los resultados obtenidos, y un estudio comparativo de las herramientas para el apoyo de practicas de VT existentes en el mercado (ver Tabla 2). VIGTECH se ha desarrollado buscando de una parte darle un componente fuerte para las fases de búsqueda, captación e inteligencia tal como se ha explicado en párrafos anteriores, y de otra parte se ha desarrollado la herramienta en un entono Web que permita interactuar de manera amigable e intuitiva a un empresario, investigador o tomador de decisiones.

El desarrollo de la herramienta informática VIGTECH está basado en una licencia GPL, buscando así que las organizaciones tengan acceso libre y puedan de esta forma apoyar sus prácticas de VT. Creemos que disminuyendo los altos costos, integrando una interfaz amigable y configurando una herramienta tan completa como sea posible es viable potenciar el uso de este tipo de herramientas[2].

Así el artículo está organizado de la siguiente forma: la Sección 2 plantea el acercamiento teórico y las técnicas computacionales utilizadas en la concepción de la herramienta informática; en la Sección 3 se hace un análisis comparativo de las herramientas existentes para el apoyo de practicas de VT; en la Sección 4 se describe la arquitectura, la funcionalidad y la implementación de la herramienta VIGTECH; en la Sección 5 se muestran y discuten los resultados obtenidos; finalmente, en Sección 6 se presenta las conclusiones y el trabajo futuro.

II. CONCEPCIÓN Y ESTABLECIMIENTO DE REQUISITOS DE LA HERRAMIENTA INFORMÁTICA VIGTECH

Las prácticas de vigilancia tecnológica buscan el estudio de comunidades científicas, dado que son éstas las que permiten el desarrollo del conocimiento y por lo tanto es el estudio de dichas comunidades el punto de partida para el desarrollo de la herramienta informática VIGTECH. El enfoque propuesto para el diseño de dicha herramienta, está basado en el modelo de unidades de análisis de una comunidad científica [13] donde se señalan tres unidades de análisis -cognitiva, científica y textual- las cuales permiten el mapeo del quehacer de los científicos. De esta forma, es posible reconocer las relaciones socio-cognitivas que se presentan en la producción del conocimiento que no son sólo las relaciones cognitivas (relaciones entre palabras) encontradas en un texto las que representan el quehacer de los científicos, sino también las relaciones que se dan en la construcción social del conocimiento, tomando así relevancia el estudio las estructuras y los vínculos sociales que aparecen en los documentos científicos, tales como co-autoría o cooperación interinstitucional.

En este sentido la teoría del actor red [14] aporta a esta discusión y propone un enfoque complementario reconociendo que para observar los procesos de producción de conocimiento es necesario encontrar los elementos vinculantes existentes al interior de un sistema de relaciones en el que participan entidades sociales y documentales. Entonces, para avanzar en el diseño de la herramienta planteamos los siguientes requerimientos: i) la herramienta debe estar enfocada a realizar análisis reticulares (relacionales) y por lo tanto análisis socio-cognitivos, ii) debe modelar, representar y obtener medidas de las estructuras sociales de la comunidad científica, iii) debe permitir el análisis de los documentos científicos en el contexto de la producción de conocimiento y iv) debe permitir la identificación de información útil [15] para llevara a cabo practicas de VT.

2 Para el caso colombiano se observa en la Encuesta TICs a las empresas manufactureras que encuestó a las empresas innovadoras en Colombia, que solo el 28% llevan a cabo prácticas de VT asistidas por herramientas informáticas.

Por último, identificamos las técnicas computacionales que nos permitan solventar los requerimientos, en la TABLA I se presentan los resultados.

Tabla 1.Técnicas Computacionales para Satisfacer Requerimientos de un Sistemas para el Análisis de Comunidades Científicas

Requerimiento	Métodos de análisis	Técnicas computacionales
Representar y obtener medidas de las estructuras sociales de la comunidad científica.	Análisis de redes sociales: el análisis de redes sociales, es una metodología de análisis cuantitativo y estructuralista [7] que busca reconocer las relaciones y sus estructuras para poder encontrar en dicho sistema de relaciones y de actores, comportamientos y en sí la estructura -o estructuras- sociales de la comunidad analizada.	El análisis de redes sociales, toma elementos del algebra de matrices al igual que de la teoría de grafos para construir desde un conjunto delimitado de actores vinculados entre sí, una representación de las relaciones existentes, dichos actores pueden vincularse de diferentes modos, adyacencia, afiliación o atributos. [16][17] qué significan los paréntesis. De esta forma es posible modelar, medir y visualizar las estructuras sociales, identificando patrones estructurales por ejemplo diadas, tríadas, hoyos estructurales, etc, y métricas de las estructuras relacionales tales como grado, centralidad y periferias, entre otras.
Analizar los documentos científicos en el contexto de la producción de conocimiento	Análisis estadísticos descriptivos: Permite construir indicadores basados en las propuestas hechas por la cienciometría o biliometría, donde el principal objetivo es medir el estado de la CyT+I [13]. Utilizando variables, tales como autores, palabras clave, revista donde fue publicado, fechas, palabras encontradas en el artículo o patente, entre otras.	Estadísticas descriptivas y multivariadas sobre el corpus construido, indicadores como número de investigadores, número de artículos por año o indicadores bibliográficos sobre el estado de la publicación científica, tales como co-ocurrencia de palabras, estudio de referencias, etc.
Identificar información útil para llevar a cabo practicas de VT.	Los métodos utilizados son propios de la minería de datos, de esta forma se utilizan técnicas de extracción de características de los documentos científicos tales como palabras clave, temáticas, etc. Así lo que se busca es identificar agrupamientos naturales de los documentos, identificar patrones, clasificaciones automáticas, entre otras [5][18]	Extracción de características: vectores de características documentales y funciones de distancia propias para este tipo de datos por ejemplo jaccard. Para identificar información útil: palabras asociadas, reducción de dimensionalidad, escalamiento multidimensional, agrupamiento de documentos y modelos gráficos probabilísticos.
La herramienta debe estar enfocada totalmente a realizar análisis reticulares	Vincular los resultados obtenidos por el análisis de redes sociales y la extracción de palabras clave, permite representar en un mapa tanto las relaciones sociales como las relaciones cognitivas.	Las técnicas utilizadas son principalmente para visualizar y obtener métricas de los resultados previamente obtenidos, así las técnicas utilizadas son reducción de dimensionalidad, escalamiento multidimensional, agrupamiento y visualización de grafos.

III. ANÁLISIS COMPARATIVO DE LAS HERRAMIENTAS EXISTENTES PARA EL APOYO DE PRÁCTICAS DE VT

Para llevar a cabo el análisis comparativo entre las herramientas existentes, se han tomado principalmente los trabajos de [2],[6], estos plantean que para hacer un análisis comparativo de las herramientas informáticas para el apoyo de prácticas de VT se debe atender al ciclo o proceso de VT, el cual se compone de cuatro fases: planeación; búsqueda y captación; análisis y organización; inteligencia y comunicación. Se construyó así, un cuadro comparativo, el cual se presenta en la TABLA II, este cuadro toma como base los cuadros comparativos propuestos por los autores mencionados pero se hace una nueva propuesta en tanto se han actualizando los costos, e incluído nuevos criterios de comparación principalmente en las fases de búsqueda y captación, e inteligencia.

De esta forma, es posible ver cómo la herramienta informática VIGTECH obtiene un índice de 0.78, lo cual significa que cumple con las características de un software

Tabla 2. Cuadro comparativo de herramientas que apoyan las practicas de vigilancia tecnológica

Producto	Costo	Información que utiliza	Fases a las que apoya la herramienta							
			Planeación	Búsqueda y adquisición		Análisis	Inteligencia			
				Automatizada		No automatizada	Descriptivos	Multivariados	Cienciometría	Bibliometria	AnaliSís de redes sociales	Minería de datos	Comunicación	Índice de la herramienta
Copernic	119,95 US/Euro	Búsquedas sobre Internet.	No	Metabuscador	No	No	No	No	No	No	No	No	No	0,11
Strategic Finder	457 Euro	Búsquedas sobre Internet.	Si	Metabuscador	No	Si	Si	No	No	No	No	No	Si	0,44
TextAnalyst	No aplica	Búsquedas en texto	No	No	No	Si	Si	Si	No	No	No	No	No	0,33
T-LAB	618 US*	Búsquedas en texto	No	No	No	Si	Si	Si	No	No	Si	No	No	0,33
SPSS	1499 US	Información semi-estructurada	No	No	Si	Si	Si	Si	No	No	Si	No	No	0,44
Tetralogie	12000 Euro	Información semi-estructurada	No	No	Si	Si	Si	Si	Si	Si	Si	No	No	0,67
Matheo analizer	3450 Euro	Información semi-estructurada	Si	No	Si	Si	Si	Si	Si	Si	Si	No	No	0,67
Matheo patent	600 Euro	Información semi-estructurada	Si	Si	No	Si	Si	Si	Si	Si	Si	No	No	0,67
Goldfire	16000 US	Información semi-estructurada Búsquedas en texto Búsquedas sobre Internet	Si	No	Si	Si	Si	No	Si	Si	Si	Si	Si	0,78
Vigtech	Libre	Información semi-estructurada	No	Crawler	No	Si	Si	Si	Si	Si	Si	Si	No	0,78

captación, procesamiento y análisis con un modulo completo para la fase de inteligencia y en un entono Web, característica que las otras herramientas no tienen. De otra parte, la inexistencia de herramientas libres que apoyen las prácticas de VT da una ventaja comparativa importante. Por último, existen otros dos parámetros necesarios al momento de escoger una herramienta informática que apoye las prácticas de VT, el primero relacionado con la necesidad de personal experto en la herramienta, y el segundo con los costos [2].

Dado que la herramienta VIGTECH se ha desarrollado en un entorno Web la necesidad de personal experto se disminuye dadas las condiciones expuestas anteriormente los costos también; así los costos de las otras herramientas no

solo incluyen la compra de la licencia, sino, costos de instalación y capacitación; de otra parte, los costos de la tecnología dispuesta por la organización para dicho propósito, tal es el caso de servidores, computadores, acceso a Internet, entre otros, suponen una alta inversión de la organización, sin embargo la herramienta VIGTECH está diseñada bajo una arquitectura cliente servidor. Dicha infraestructura ha sido dispuesta en este momento por el Observatorio Colombiano de Ciencia y Tecnología, pero puede ser instalada en cualquier organización que cuente con una infraestructura similar.

Así, podemos ver cómo de una parte se atiende a los requerimientos planteados en la sección anterior brindando ventajas competitivas importantes pues no hay herramientas que integren técnicas propias de la cienciometría, bibliometría, análisis de datos, de análisis de redes sociales y de minería; de otra parte, podemos ver cómo el desarrollo de la ciencia y la tecnología en el país se pueden ver potenciado por el uso de prácticas de VT al interior de las organizaciones.

IV. ARQUITECTURA, FUNCIONALIDAD E IMPLEMENTACIÓN DE LA HERRAMIENTA INFORMÁTICA VIGTECH

A. Arquitectura y funcionalidad de la herramienta

Como ya se ha dicho en las secciones anteriores, la herramienta VIGTECH es el primer elemento en el desarrollo de un sistema que permita apoyar completamente las prácticas de VT, dado que está basada solamente en el ámbito científico tecnológico, y por lo tanto su diseño está enfocado en el análisis reticular de comunidades científicas, apoyando principalmente tres fases del ciclo de VT, estas son las fases de búsqueda y captación; análisis y organización; e inteligencia.

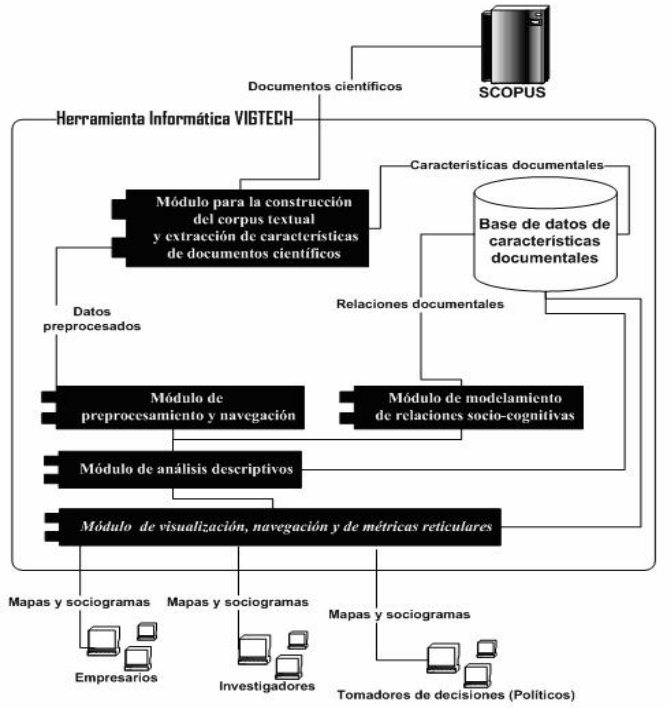

Figura 1. Arquitectura de la herramienta informática VIGTECH

La arquitectura de la herramienta informática VIGTECH se presenta a través de su visión estructural y su visión funcional. En la Figura 1 se presenta la arquitectura de la herramienta, la cual responde principalmente a los requisitos mencionados en la Sección 2, sus módulos y su funcionalidad.

Entonces la herramienta se describe como un sistema capaz de buscar, identificar, extraer y representar las estructuras relacionales presentes en los documentos científicos. Así el punto de llegada no es simplemente almacenar la información producida tal como sucede actualmente con SCOPUS, y otras bases referenciales, sino es la fuente de información que permite llevar a cabo prácticas de exploración, transformación de datos y en ultima instancia, extracción de conocimiento [19], pues es de esta forma, que podemos encontrar las estructuras que permiten contribuir a entender los procesos de producción de conocimiento y a la toma de decisiones estratégicas. Así, técnicas de análisis de datos, de minería de datos y de aprendizaje de máquina permiten el descubrimiento de patrones y la extracción de conocimiento en una compilación de textos, haciendo explícitas las relaciones existentes entre temáticas y autores. Por último, la herramienta cuenta con un administrador de proyectos, que permite organizar y mantener un historial del trabajo llevado a cabo en la herramienta.

B. Módulos de la herramienta informática VIGTECH

1) Módulo para la construcción del corpus textual y extracción de características de documentos científicos: este módulo permite obtener los documentos científicos de forma automática mediante el crawler-VIGTECH-. Ingresando una cadena de búsqueda, el sistema es capaz de ingresar a SCOPUS y descargar los reportes por documento y a partir de estos resultados generar una base de datos que extrae los meta-datos del documento, tales como año, revista, título, resumen, entre otros, las palabras clave y los meta datos del autor, tales como institución de afiliación, temas de interés del autor, entre otros. De otra parte, el sistema permite cargar datos a partir de un archivo plano.

2) Módulo de preprocesamiento y navegación: este módulo permite depurar los registros de la base de datos, depuración de forma y de repeticiones y navegar por el corpus construido. De esta forma teniendo el proyecto inicializado y la base construida se normalizan automáticamente las palabras clave a través de Stemming[3] y se normalizan los

3 Técnica de preprosesamiento que permite encontrar la raíz de una palabra, en este caso se utilizó el algoritmo de Porter.

autores a través del código de identificación único proporcionado por SCOPUS, en la Figura 2 se presentan las pantallas de estos módulos.

Figura 2. Módulo para la construcción del corpus textual y extracción de características de documentos científicos y Módulo de preprocesamiento y navegación:

3) Módulo de análisis descriptivos: con el corpus almacenado y preprocesado se construyen las frecuencias de las palabras clave y de los autores. Estos análisis estadísticos descriptivos de la producción científica se basan en la construcción de indicadores de conteos bibliográficos los cuales pueden ser datos de la producción clasificada por años, nombres de los autores, palabras contenidas en los títulos o resúmenes, descriptores e identificadores, citas que hace cada artículo, etc. De esta forma se construyen tablas de frecuencias que permiten construir indicadores básicos propios de la bibliometría y cienciometría y sus respectivas relaciones son útiles al momento de modelar los vínculos socio-cognitivos.

4) Módulo de modelamiento de relaciones socio-cognitivas: con el fin de construir métricas y mapas de las redes temáticas y sociales que se encuentran en los documentos científicos, se modelan los documentos de acuerdo con sus características documentales tales como autores, palabras clave, entre otras. Este módulo construye matrices que modelan el conjunto de documentos, de esta forma la matríz se compone de vectores de características por documento, el cual tiene 1 o 0 si la palabra clave, autor, etc. se encuentra o no en el documento.

El módulo permite calcular distancias de similitud propias para este tipo de representaciones tal es el caso de la medida de similitud Jaccard y coseno entre otras [10].

Para modelar las relaciones sociales se toman los documentos según sus relaciones de autoría, construyendo así una matríz de relaciones de co-autoría, donde un documento es descrito por un vector del total de autores el cual tiene 1 o 0 si el autor se encuentra o no en el documento. Así, lo que se busca en este módulo es de una parte construir matrices binarias, de palabras clave, autores, etc. y de otra parte modelar las relaciones, a través de la frecuencia de aparición de la palabra clave en el título o resumen o de número de coautores en un documento lo cual permite modelar el documento con funciones de frecuencia, frecuencia inversa, etc. [10]

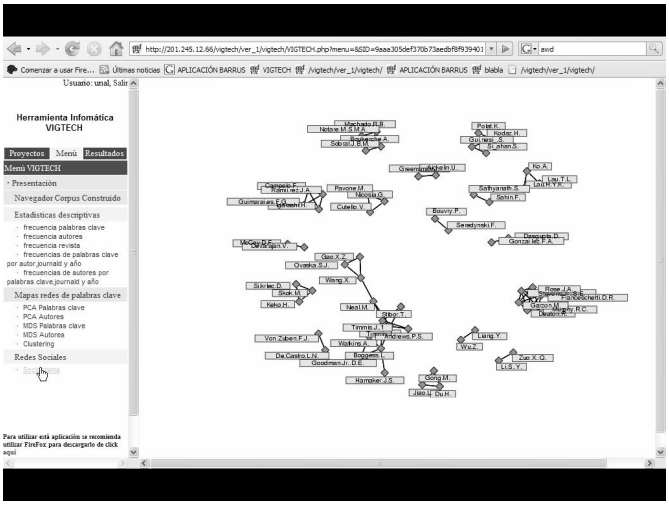

Figura 3. Módulo de visualización, navegación y de métricas

5) Módulo de visualización, navegación y de métricas reticulares: este modulo se desarrolló para realizar mapas tecnológicos, socio-gramas y construir métricas reticulares. El objetivo principal es utilizar técnicas de reducción de dimensionalidad, de escalamiento multidimensional, de agrupamiento [20] y de análisis de redes sociales para construir una representación en dos o tres dimensiones de un área científica de interés y métricas que den cuenta del estado y las dinámicas de dicha área. Buscando la representación visual de los datos obtenidos se presentan gráficos donde se describe el comportamiento de un tópico o grafos que de igual forma permiten construir medidas y obtener representaciones donde se vinculan la dimensión social y cognitiva. En la Figura 3 se presenta la pantalla de este módulo. Los nodos del grafo visualizado corresponden a los autores y los arcos las relaciones de co-autoría.

C. Implementación

Para el desarrollo de la herramienta informática VIGTECH se utilizó la arquitectura cliente servidor. Como servidor Web se utilizó Apache 2.0 Handler, y como motor de base de datos

MySQL 4.0.18. Los desarrollos se realizaron en PHP Versión 4.3.4 y JavaScript.

Para lograr dicha implementación se tuvo en cuenta que el desarrollo de ésta supone un gran esfuerzo y por lo tanto se planteó la necesidad de integrar en una sola herramienta informática, software disponible para propósitos específicos tal es el caso de R (*Foundation for Statistical Computing) versión 2.0* y *Weka* (Waikato Environment for Knowledge Análisis), software que dispone de bibliotecas especializadas para hacer aprendizaje de maquina y minería de datos, de esta forma utilizando las implementaciones de algoritmos de visualización, clustering y análisis de redes sociales se integró a través de PHP a la herramienta VIGTECH.

Para identificar e implementar las técnicas de visualización, minería de datos y análisis de redes sociales se hizo un trabajo exploratorio el cual permitió reconocer las técnicas computacionales que permiten desarrollar mapas científicos y definir las funciones y librerías útiles para la herramienta, de esta forma las funciones y librerías de R utilizadas se presentan en la siguiente tabla:

Tabla 3 Librerías y Funciones de R utilizadas

Técnica computacional	Funciones y librerias utilizadas
Análisis de componentes principales	prcomp de la librería STATS
Escalamiento multidimensional	isomds, sammon, cmdscale prcomp de la librería STATS
Agrupamiento	SOM, Hculst, pam incluidas en la librería cluster,

V. MÉTODOS PARA EL PROCESAMIENTO Y EXTRACCIÓN DE CONOCIMIENTO

En esta sección se describe las fases de extracción de características de los documentos, el modelamiento de dichos documentos, las técnicas utilizadas en la extracción de conocimiento y en la visualización a través de mapas de los resultados obtenidos.

A. Extracción de características y representación de documentos.

Un documento puede considerarse como un vector D de características hasta un total de *j*, donde un valor que pertenece a los números naturales expresa en que grado el documento posee la característica en la posición *i*.

$$D = \left(c_{1,}, c_2, c_{3,\ldots,} c_j\right) \tag{1}$$

$$c_i \in N \tag{2}$$

La característica en este caso es la ocurrencia o no de determinas palabras, autores o de la frecuencia de aparición de uno de estos en el documento.

Para calcular la similitud de un documento a otro se ha utilizado el calculo de distancias propias para datos binarios [(presencia (1) y ausencia (0)] así tenemos principalmente Jaccard, y coseno.

El coeficiente de Jaccard (porcentaje de presencia-ausencia) puede variar entre 0 y 1, donde 0 indica ausencia de características en común y 1 en el caso que los documentos sean idénticos.

$$J = \frac{2a}{2a + b + c} \tag{3}$$

Donde (a) representa dos presencias (1:1), (b) representa presencia-ausencia (1:0) y (c) ausencia-presencia (0:1).

La distancia coseno permite medir la similitud entre un documento y otro, está se puede representar como el ángulo entre sus representaciones en el espacio vectorial, noventa grados sin similitud (perpendicular), cero grados máxima similitud (idénticos).

$$D(x, y) = 1 - x \cdot y \Big/ \left(|x||y|\right) \tag{4}$$

Donde x.y es el producto punto entre x e y, y |x| es la norma del vector x.

B. Construcción y visualización de redes cognitivas.

Para la construcción y visualización de redes cognitivas o de palabras clave se utilizar técnicas que obtienen mapas que representan las características más relevantes según un criterio y las relaciones de estas características, dichas técnicas son análisis de componentes principales, escalamiento multidimensional y agrupamiento.

$D(x, y) = 1 - x \cdot y \Big/ \left(|x||y|\right)$

1) Análisis de Componentes Principales (PCA): el objetivo del análisis de componentes principales es reducir la dimensión de un conjunto de variables a un conjunto *m* de menor número de variables que permita manejar el problema de la multidimensionalidad, obteniendo así una representación que ofrece la mayor información disponible en el conjunto de datos, de esta forma se busca la proyección de los datos dentro de nuevo conjunto de ejes. Así el análisis de componentes principales busca centrar los datos en la media, escalar la varianza y rotar los ejes principales producidos por una transformación lineal ortogonal.

2) Escalamiento multidimensional (MDS): es un método basado en la información de las distancias de un conjunto de datos multivariados, busca reducir la dimensión *L* encontrando un conjunto de vectores que pertenezcan a los

reales y que reproduzca las distancias del conjunto inicial. En términos generales podemos decir que un MDS es un PCA que previamente ha utilizado una función de distancias para hacer la reducción de dimensionalidad de ésta y no de los datos originales. Un MDS comienza con un conjunto de ejes tomados de PCA y busca minimizar el "stress" o media del error cuadrático entre el conjunto inicial de ejes y la matriz de distancia original, la técnica comúnmente utilizada es sammon, esta técnica define la medida del stress como la relación existente entre la matriz de distancias y una matriz randomica creada con igual distribución en un espacio de dos dimensiones.

3) Agrupamiento (CLUSTERING): esta técnica permite la organización de una colección de patrones usualmente representados por un vector de características o un punto en un espacio multidimensional. Encontrando agrupaciones naturales de los datos basadas en una función de similitud o disimilitud entre los vectores o puntos. Así, los puntos que comparten características se agrupan dado que son homogéneos y por lo tanto estarán dentro de los mismos grupos (mínima varianza) y los datos disímiles quedaran en grupos diferentes y separados entre ellos (máxima varianza). Las técnicas de agrupamiento utilizadas en la herramienta son aglomerativas, y de grafos de partición. Para el caso de las técnicas aglomerativas, se encuentran agrupaciones asignado inicialmente cada objeto a un grupo y repetidamente se unen pares de grupos hasta cumplir con un criterio de parada, las formas de unión utilizadas son enlace simple y enlace completo, para el caso de las técnicas de agrupamiento basadas en grafos de partición, se construye un grafo como modelo de afinidad de las relaciones entre las palabras y después se particiona o se exige el grafo.

Los métodos particionales consideran solamente las relaciones de afinidad entre un objeto y un pequeño número de los ejemplos similares, lo cual permite identificar los grupos de temáticas en un tópico dado.

C. Construcción y visualización de redes sociales.

Para la construcción de redes sociales se utilizan elementos del Análisis de Redes Sociales-ARS-, esta es una metodología cuantitativa y estructuralista que permite reconocer los sistemas de relaciones presentes en una comunidad identificando de una parte actores centrales, periféricos, de paso obligado, etc. y de otra, los comportamientos, patrones de enlaces, uniones estables que dan cuenta de vínculos irreversibles, que expresan la existencia de uniones internas fuertemente ligadas en las cuales se presentan normas, valores y orientaciones propias de la comunidad. Así el ARS, utiliza elementos tomados del álgebra matricial al igual que de la teoría de grafos para construir desde un conjunto delimitado de actores vinculados entre sí, una representación de las relaciones existentes. De esta forma las relaciones de coautoría se modelan como un grafo en donde en los nodos se encuentran los autores y los enlaces representan las relaciones. De esta forma se obtiene un grafo de coautorías al cual se pueden aplicar diferentes medidas tales como grado, centralidad por intermediación, etc.

D. Construcción y visualización de redes socio-cognitivas.

La técnica de modelos gráficos probabilísticos, permite encontrar la dependencia existente entre las relaciones de autores y palabras clave frente a la categoría. Entoces se busca determinar qué variables influencian dicha categoría, esto es posible dado que existen relaciones documentales que clasifican dichos documentos por sus palabras clave y por sus relaciones de co-autoría, que evidencian las relaciones sociales existentes en una comunidad académica. para ello, se modelan los documentos y sus relaciones como una red de probabilidades donde se pueden hacer inferencias, encontrar variables relacionadas no conocidas o correlaciones. En los casos donde todas las variables son no conocidas estas técnicas nos permiten modelar las posibles correlaciones entre las variables de las que sí tenemos información y de esta forma construir una red Bayesiana que permita hacer inferencias y por lo tanto caracterizar las relaciones probabilísticas existentes entre las variables y la clase.

VI. RESULTADOS OBTENIDOS

Los resultados que se presentan a continuación están organizados de acuerdo a las fases del ciclo de VT; se presentan los tiempos de descarga, los indicadores, las matrices y visualización de mapas y socio-gramas obtenidos e implementados en la herramienta informatica VIGTECH, para este fin se utilizaron diferentes conjuntos de datos por lo cual los resultados que se presentan a continuación dan cuenta solamente, de los resultados que se han obtenido como desarrollo de la herramienta y no deberían ser tomados como un ejercicio de minería de datos o de análisis de un tópico.

A. Fase de búsqueda y captación:

El crawler-VIGTECH-, permite conectarse a SCOPUS y descargar automáticamente los artículos científicos, el tiempo

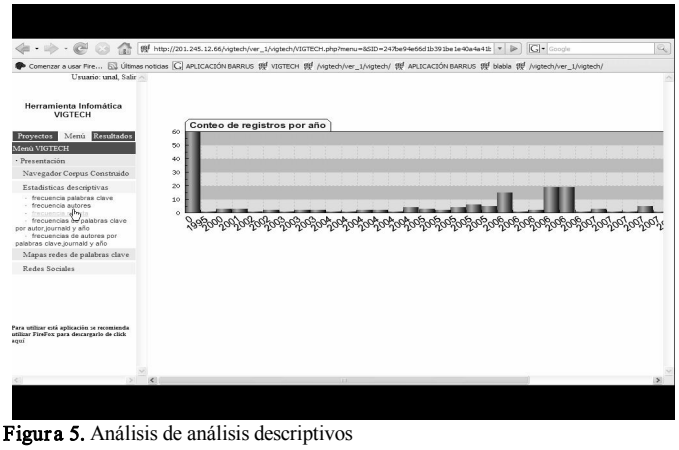

Figura 5. Análisis de análisis descriptivos

promedio de descarga es de 0,3 segundos/artículo, una vez descargados los documentos se extraen las siguientes características del documento: nombre de la revista y meta-

datos, título del artículo, tipo de documento, número de citas en SCOPUS, resumen, número de referencias utilizadas en el artículo, correo del autor, palabras clave, autor y meta-datos del autor tales como id del autor en SCOPUS, afiliación institucional y áreas de interés del autor, para la extracción de características e inserción en la base de datos el sistema se demora 1 segundo/artículo. Fase de análisis de análisis descriptivos

El sistema permite llevar a cabo diferentes análisis desde la construcción de indicadores tales como número de artículos por palabra clave, número de artículos por año, autores por revista, número de publicaciones de un autor, entre otros. De igual manera, el sistema es capaz de obtener estadísticas básicas, en la siguiente figura se presentan algunos resultados al respecto.

De otra parte, el sistema permite llevar a cabo análisis de

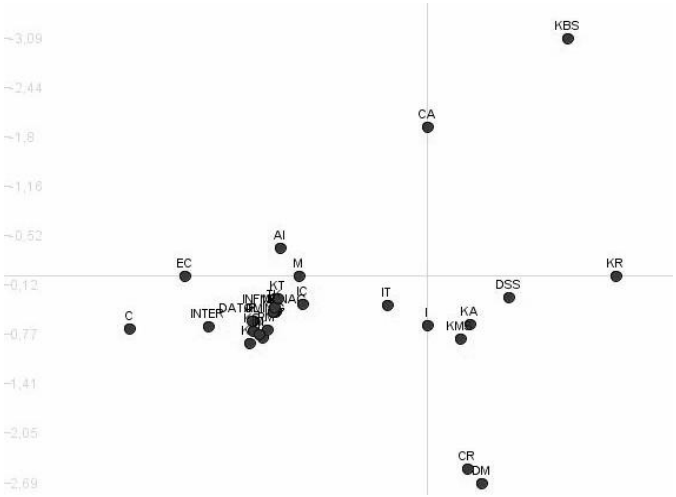

Figura 6. Análisis de componentes principales

componentes principales PCA para el caso de palabras clave en la figura 6 se presenta un PCA obtenido para un conjunto de datos de gestión del conocimiento. Esta técnica da una representación de los documentos pero no permite una fácil interpretación de los resultados, en el siguiente gráfico se ve cómo la categoría knowledge management(KM) está cerca del centro y cerca de la media y hay un alto grado de cercanía de los datos de las categorías Internet(I), management(M), information retrieval(INFR), sin embargo, no se puede decir que son grupos de temáticas simplemente se puede decir que hay una alta correlación entre estas categorías. Por último, se puede decir que hay unas categorías que se encuentran alejadas tales como Data mining (DM) o knowledge representation (KR), lo cual nos da un acercamiento al problema y un primer acercamiento a la construcción de mapas científicos.

Otras tres técnicas de exploración de datos utilizadas son escalamiento multidimensional, agrupamiento y análisis de redes sociales. Para el caso de escalamiento multidimensional los resultados obtenidos permiten tener una representación más acertada dado que como dijimos anteriormente se

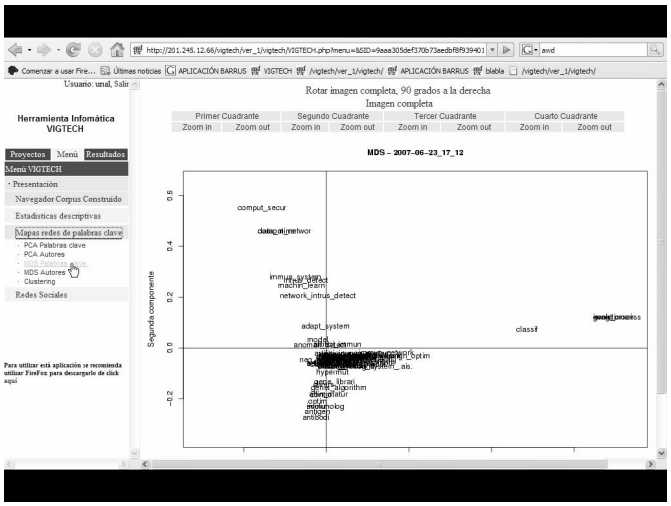

Figura 7. MDS aplicado a palabras clave

mapean las distancias y no las relaciones entre los datos permitiendo así tener una representación con mayor información. Para la evaluación de las técnicas de reducción de dimensionalidad y de análisis de componentes principales se utilizaron medidas de explicación y de estréss que expresen la representatividad que tienen esos resultados del conjunto de datos particular. La técnica de PCA da un valor de explicación del 54.12% para el conjunto de gestión del conocimiento lo cual no es un valor bueno dado que no explica bien el modelo, esto se puede evitar utilizando las técnicas de escalamiento multidimensional, buscando mapear las distancias y no los datos, obteniendo así un valor de estréss del 70 % para el mismo conjunto. En la siguiente Figura se presenta la implementación del MDS y aplicado a las palabras clave del conjunto de datos *(artificial immune systems)*.

B. Fase de análisis socio-cognitivos

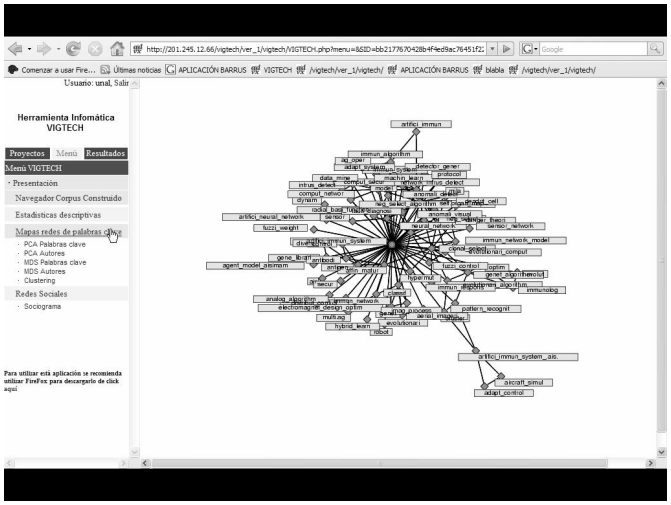

Figura 8. Red de palabras clave.

Para la fase de inteligencia se han utilizado principalmente técnicas de redes sociales y agrupamiento para obtener las relaciones existentes entre las palabras clave y los autores. Un

primer acercamiento al análisis de redes por palabras clave (ver Figura 8) muestra cuál es el núcleo del cuerpo de conocimientos de un tópico dado, de esta forma es posible encontrar las palabras clave que tienen mayor transitividad en el grafo y están más correlacionadas formando así componentes y representando el núcleo del cuerpo de conocimiento. De otra parte, existen otras palabras clave que no tienen vínculos, lo cual sugiere preguntas tales como, si estas temáticas están en emergencia o son temáticas que perdieron vigencia, de esta forma encontramos en el centro de la estructura a knowledge management (KM) y una componente principal de 28 nodos lo cual representa el núcleo del área del conocimiento. En la siguiente Figura se presenta la implementación de la red de palabras clave para el conjunto de datos *(artificial immune systems)*.

La técnica de agrupamiento es la que mejores resultados ofrece, así para la exploración de temáticas del área Gestión del Conocimiento se utilizaron algoritmos de agrupamiento tales como algoritmos aglomerativos, de grafos de partición y se variaron las funciones de distancia, obteniendo varios resultados que permiten tener un mapa de las áreas en las que se está trabajando en el tópico de gestión del conocimiento KM., Obtuvimos los siguientes grupos descritos por las palabras clave más representativas del cluster:

- Knowledge Management, e-learning, information technology
- Knowledge Management, data mining, ontology, knowledge modeling,
- Knowledge Management, organizational development, organizational learning
- Knowledge Management, document management, Knowledge Management system
- Knowledge Management, semantics, communities of practices
- Knowledge Management, Knowledge networks, communities of practices.

Por último, es a través del análisis de redes sociales que se busca reconocer y medir las estructuras sociales al interior de una comunidad científica, tal como se presentó en la Figura 3. el grafo de co-autorias del tópico (*artificial immune systems*), permite reconocer estructuras o componentes en el grafo que dan cuenta de la existencia de relaciones fuertes en ese campo de estudio, Este grafo dirigido está compuesto por 70 autores que tienen una frecuencia de aparicion mayor a 4, dicho resultado permite identificar grupos que son fácilmente reconocibles por su centralidad dentro de la estructura de relaciones, y una componente del grafo de compuesta por 11 autors y las otras componentes de menos numero que representando las tematicas y que las relaciones de coautoria estan fuertemente ligadas a las relaciones tematicas. De esta forma podemos obtener un socio-grama en el que se representan los vínculos de co-autoría del área y la estructura de relaciones existente en esta comunidad científica. Es

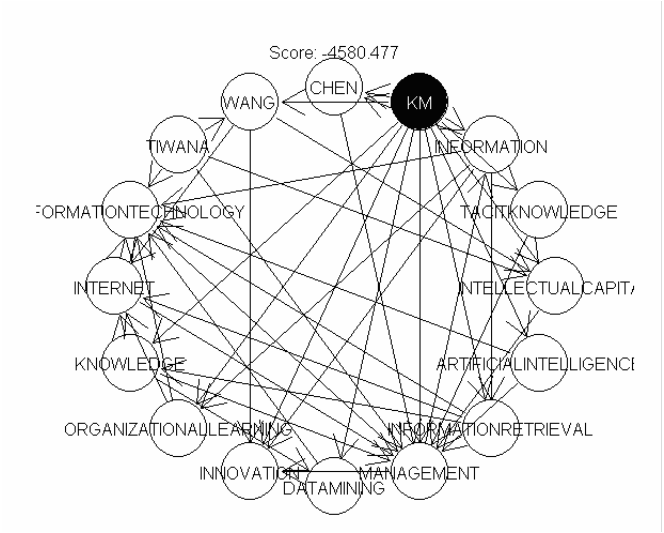

Figura 9. *Directed acyclic grapas (DAG),* resultado modelos gráficos probabilísticas

claramente reconocible un subconjunto de autores que tienen mayor centralidad en el grafo, el resto se encuentra en la periferia, la medida utilizada para este trabajo es centralidad por intermediación-centrality betweness- la cual es de 0,13 en promedio indicando que es un tópico en el cual no hay redes de co-autoría fuertes, consolidadas sino aún en emergencia.

Por último, se presenta la red obtenida a través de la técnica de modelos gráficos probabilísticos; en la siguiente figura se presentan los resultados obtenidos, estos describen claramente las relaciones entre las palabras clave, co-autoría, y la clase KM, esta representación se acerca más a una red socio-cognitiva, dado que encontramos relaciones tales como, las del autor Chen con innovación, o de Tiwana con Chen, sobre gestión del conocimiento con capital intelectual o la inexistencia de relaciones de capital intelectual con minería de datos o de Wang con Chen, esto nos permite ver cómo esta técnica ofrece resultados más inteligibles para construir un mapa científico tecnológico de un tópico dado.

VII. CONCLUSIONES Y TRABAJO FUTURO

Los avances conseguidos con el desarrollo de la herramienta informática VIGTECH han permitido el modelamiento de los documentos por sus palabras clave y autores, el análisis descriptivo del corpus documental y el análisis de los datos , de esta forma se ha automatizado los procesos de búsqueda y captación; análisis e inteligencia del proceso de VT. La versión beta de la herramienta se encuentra en el link http://201.245.12.66/vigtech/ver_1/vigtech/index.php

Este acercamiento permite la recuperación de información y la extracción de conocimiento de la documentación científica, disminuyendo el desgaste, el tiempo y la curva de aprendizaje, se espera que el uso de esta herramienta permita el avance en la incorporación de conocimiento científico el

desarrollo de nuevos productos o procesos; el desarrollo del estado del arte de una investigación, el reconocimiento del estado y la dinámica de la ciencia y la tecnología en un departamento, país u organización, y que a través de esta obtengan información útil para la toma de decisiones y planeación de políticas en materia científica.

Se ha logrado hacer una primera validación de la herramienta informática VIGTECH construyendo un conjunto de artículos científicos y caracterizar los documentos a través de palabras clave y autores, representar sus estructuras relacionales y encontrar métricas de dichas representaciones. Se encontró que las técnicas utilizadas permiten dar una representación visual de los documentos y que las técnicas más acertadas para dar dichas representaciones, son los modelos gráficos probabilísticos, agrupamiento y análisis de redes sociales dado que permiten una interpretación más intuitiva de las representaciones mientras que las técnicas de reducción de dimensionalidad PCA o MDS dan una representación de los documentos pero no permiten una fácil interpretación de los resultados.

Como trabajo futuro se integrará completamente la herramienta como un sólo paquete de software, y se validará en diferentes centros de desarrollo tecnológico, grupos de investigación y empresas. De otra parte, es determinante seguir desarrollando y mejorando esta primera versión buscando aumentar la eficiencia de la herramienta pues por ser una herramienta web, en los procesos de captura, pre-procesamiento, construcción de matrices, análisis de redes sociales, los tiempos de espera son amplios, por tal razón, se propone incluir técnicas eficientes de computación de matrices con el fin de minimizar la complejidad que suponen las representaciones matriciales.

AGRADECIMIENTOS

Manifestamos nuestro agradecimiento al Observatorio Colombiano de Ciencia y Tecnología por sus valiosos aportes y participación en el proyecto, y por permitir utilizar sus recursos e infraestructura informática con lo cual se ha podido avanzar en el desarrollo de está herramienta.

REFERENCIAS

[1] M. Castells, La Sociedad Red. La Era de la Información, Madrid: Ed. Alianza, 1996

[2] A. Leon , O. Castellanos y F. W. Vargas., "Valoración, selección y pertinencia de herramientas de software utilizadas en vigilancia tecnológica", Revista de Ingeniería e investigación, Vol. 26(01), p. 92-102, 2006.

[3] F. Chaparro, "Apropiación Social del Conocimiento, Aprendizaje y capital social," Medellín, Universidad de Antioquia, Simposio Internacional sobre Ciencia y Sociedad, 2003.

[4] M. Ramon, De la vigilancia a la inteligencia competitiva. Madrid: Prentice Hall, 2001.

[5] A. Porter,S. Cunningham, Tech Mining, Exploiting new technologies for competitive advantage. New Jersey: John Wiley & Sons, 2005.

[6] M.Sanches, F. Palop.(2007, Enero 20). Herramientas de Software especializadas para Vigilancia Tecnológica e Inteligencia Competitiva [Online].disponible en www.intempres.pco.cu

[7] S.Wasserman, K. Faust, D. Iacobucci, M. Granovetter, Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences).Cambridge: Cambridge University Press, 2004.

[8] CH. Chen, The centrality of pivotal points in the evolution of scientific networks. En: Proceedings of the 10th international conference on Intelligent user interfaces, p. 98-105, 2005.

[9] R. Jhonson y D. Wichern, Applied multivariate statistical analysis. New Jersey: Prentice Hall, 2002.

[10] S. Sebastiani, Machine learning in automated text categorization. En: ACM Comput Suv,34, p. 1-47, 2002.

[11] J. Zhu ,J. Hong y J. S. Hughes, PageCluster: Mining conceptual link hierarchies from Web log files for adaptive Web site navigation. En: ACM Trans. Inter. Tech, Vol 4 (2), p. 185-208, 2004.

[12] G. Bottcher, Learning Bayesian Networks with Mixed Variables. En Artificial Intelligence and Statistics, p. 149-156, 2001.

[13] L. Leydesdorff, The Self-Organization of the Knowledge-Based Society. Budapest: Typotext.2005.

[14] M. Callon, Réseau et cordination. Paris: Ed Economica, 1999.

[15] M. Lewkowicz, Summary of COOP'04 workshop on interaction and knowledge management. En: SIGGROUP Bull, 24, p. 2-5,2004.

[16] B. Yu, P. Munindar, Searching social networks. En: Proceedings of the second international joint conference on Autonomous agents and multiagent systems, July 14-18, 2003. Melbourne: ACM Press, p. 65-72, 2003.

[17] Bekkerman y A. Mccallum, Disambiguating Web appearances of people in a social network. En: Proceedings of WWW 2005 bibtex ppt. Disponible http://citeseer.ist.psu.edu/bekkerman05disambiguating.html el 13 de noviembre de 2006, 2005.

[18] E. Weippl, Visualizing content based relations in texts. En: Proceedings of the 2nd Australasian conference on User interface. Queesland: ACM International Conference Proceeding Series, Vol. 14, p. 34-41, 2001.

[19] M. Kobayashi, K.R Takeda, Information retrieval on the web. En: ACM Comput. Surv, 32, p.144-173, 2000.

[20] E. Weippl, Visualizing content based relations in texts. En: Proceedings of the 2nd Australasian conference on User interface. Queesland: ACM International Conference Proceeding Series, Vol. 14, p. 34-41, 2001.