

**Borrado CERO - HERRAMIENTA INFORMÁTICA DE VIGILANCIA  
TECNOLÓGICA PARA ANÁLISIS SOCIO-COGNITIVOS DE  
COMUNIDADES CIENTÍFICAS**

Director:  
ING. FABIO GONÁLEZ OSORIO, PHD.

Autor:  
VÍCTOR ANDRÉS BUCHELI GUERRERO

MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN  
UNIVERSIDAD NACIONAL DE COLOMBIA

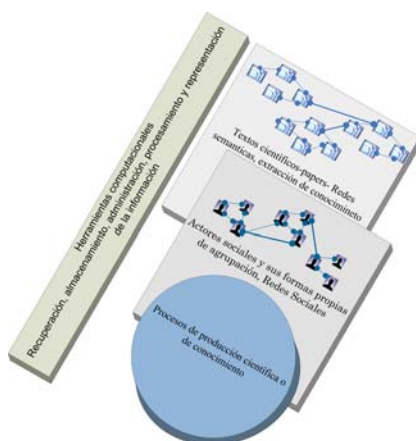
## Capítulo 1 Análisis de comunidades científicas.

el análisis de las comunidades científicas a través de herramientas computacionales, se focaliza en el análisis de redes semánticas y redes sociales como dos marcos de trabajo tanto conceptuales como metodológicos, así se aborda el tópico desde diferentes metodologías y técnicas propias de la bibliometría, la lingüística computacional, la cienciometría y la recuperación de información-IR-. Dichas herramientas permiten procesar mediante técnicas de extracción de conocimiento las publicaciones científicas, para así construir métricas sobre la producción científica y representar a través de mapas y socio gramas las estructuras relacionales.

El análisis de comunidades científicas, presentado en este artículo tiene un enfoque reticular, así es importante mencionar el cambio de paradigma en la construcción de métricas apropiadas de la producción científica, debido al desarrollo de dichas comunidades en una sociedad del conocimiento. El estudio de las comunidades científicas es de facto un problema de investigación en el que se involucran diferentes áreas del conocimiento en un proyecto transdisciplinar, integrando herramientas computacionales, métodos de manejo, recuperación, almacenamiento, administración, procesamiento y representación de la información.

Se toma el modelo propuesto por Leydersdorf, acerca de las unidades de análisis de una comunidad científica y se propone como una alternativa a este modelo, tres dimensiones para exponer el análisis de comunidades científicas desde un enfoque reticular. En la primera dimensión se encuentran los textos científicos-papers-, en la segunda los actores sociales y sus formas propias de agrupación y en la tercera los procesos de producción científica o de conocimiento, de esta forma podemos ubicar áreas del conocimiento delimitadas, que permiten hacer un acercamiento al problema desde su enfoque particular, así la sociología de la ciencia o del conocimiento, la cienciometría, y las teorías de la información y la comunicación, son integradas en un marco lógico, permitiendo así, avanzar en un análisis socio-cognitivo, vinculando el análisis cuantitativo de textos, la sociometría, la bibliometría, la recuperación de información -IR-, y en última instancia la cienciometría. Se construye así, un aparato metodológico que permite cuantificar la producción científica, dar cuenta de los procesos de producción de conocimiento, las estructuras sociales y de poder de las comunidades científicas.

El análisis de la información científica y tecnológica, permite dar cuenta del fenómeno científico-tecnológico como un proceso de producción de conocimiento donde juegan un papel importante los vínculos sociales. Así identificar las estructuras relacionales de la comunidad científica, el análisis de los documentos científicos en el contexto de la producción



de conocimiento, constituyen una unidad donde la recuperación, procesamiento y representación de la información entran a ser parte del núcleo de la discusión. Podemos seguir entonces la siguiente clasificación para presentar los trabajos realizados en el tópico acá trabajado: representación de las estructuras sociales (Ej. análisis de redes sociales), análisis estadísticos descriptivos (Ej. estadísticas sobre los datos bibliográficos, co-ocurrencia de palabras, estudio de citas), procesamiento de lenguaje natural (Ej. extracción de términos, palabras clave, semántica latente construcción y utilización de tesauros y ontologías), clasificación automática de documentos (Ej. método de las palabras asociadas, clustering de documentos). El objetivo es presentar los instrumentos (es decir, indicadores, métodos y herramientas computacionales) de análisis de la información científica y tecnológica.

## 1.1 Cienciometría, bibliometría, indicadores de ciencia y tecnología.

Los estudios sociales de la ciencia, al igual que la cienciometría, tienen como objeto de estudio las comunidades científicas, la cienciometría se preocupa por los aspectos cuantitativos de la ciencia como disciplina o actividad, y encuentra aplicación en el establecimiento y evaluación de políticas de ciencia y tecnología, vigilancia tecnológica, comunicación científica, así la cienciometría emplea, técnicas y métricas para la evaluación del fenómeno científico.

La cienciometría tiene un enfoque cuantitativo de la ciencia (número de investigadores, número de publicaciones), en el desarrollo de las disciplinas (estudio de palabras asociadas), en la relación entre ciencia y tecnología (estudios sobre innovación), en la estructura de comunicación entre los científicos (estudio de citas), en las relaciones entre el desarrollo científico y el crecimiento económico (porcentaje del PIB destinado a la investigación).

Así la cienciometría es justamente el proyecto de cuantificar la producción científica, dar cuenta del estado de la ciencia y la tecnología, sin embargo esta no se preocupa por la unidad de análisis cognitiva, tiene un enfoque netamente estadístico teniendo como unidad la contabilidad de las publicaciones y citas. Sin embargo el problema puede ser un tanto más complejo si entra en la discusión la adquisición, representación y gestión del conocimiento.

En este sentido la teoría del actor red, es de gran utilidad si planteamos la producción de conocimiento como una construcción social (Leydesdorf,2001), así observamos los procesos cognitivos dentro de un sistema de relaciones en el que participan entidades sociales y documentales que permiten encontrar elementos para la identificación del tejido –vínculos- en una comunidad científica, se preocupa por la observación, descripción y medición de las estructuras de relaciones que tiendan a construir vínculos coordinados y consolidados para la producción de conocimiento.

### 1.1.1 *Estadísticas de datos bibliográficos*

La producción científica genera grandes volúmenes de información Ej. Bases de texto completo de artículos científicos o bases de datos de patentes, dicha estructuración de la información ha permitido realizar análisis de la producción científica mediante la construcción de indicadores bibliográficos los cuales pueden ser: años, nombres de los autores, palabras contenidas en los títulos o resúmenes, descriptores e identificadores, citas que hace cada artículo, códigos de clasificación de patentes, etc.

Dichos indicadores se pueden clasificar en: *a)* el tamaño y las características de la producción científica y tecnológica, *b)* el impacto de las publicaciones (medido a través de las citas que reciben) y *c)* los aspectos estructurales de la ciencia o indicadores. Mientras que este último grupo sirve para la elaboración de los *mapas conceptuales o socio gramas*, los dos primeros, denominados indicadores de actividad, constituyen el núcleo alrededor del cual se evalúa la investigación. Algunos indicadores de actividad que se utilizan son: El crecimiento de cualquier campo de la ciencia según la variación cronológica del número de trabajos que se publican en él, el envejecimiento de los campos científicos según la *vida media* de las referencias de sus publicaciones, la evaluación cronológica de la producción científica según el año de la publicación de los documentos, la productividad de los autores o instituciones, medida por el número de sus trabajos.

### 1.1.2 *Estudio de citas*

El estudio de citas es uno de los análisis más recurrentes en esta área busca medir las citas efectuadas por los autores de artículos, éstas suponen un reconocimiento formal respecto a la investigación previa publicada ya que indican aquellos trabajos anteriores que se han considerado relevantes en el tema y que han influido en el artículo. Conocer los artículos más citados tendrá gran interés, ya que se trata, probablemente, de los documentos más influyentes en el área considerada.

El ISI (*Institute for Scientific Information*) construye métricas basadas en las citas de revistas, lo cual proporciona un indicador del desarrollo de la investigación en un determinado campo. El JCR (*Journal Citation Reports*) produce dichos indicadores de estudio de citas, incluye más de seis mil publicaciones de sesenta países sobre todas las especialidades en ciencia, tecnología y ciencias sociales. El JCR ofrece una perspectiva para evaluaciones de revistas, autores y temas, mediante la tabulación y agregación de citas algunos indicadores son: El número de citas recibidas por una revista (citas de artículos publicados en la revista) durante un año determinado. Este indicador muestra el uso que los investigadores hacen de cada revista. Como sabemos, un pequeño porcentaje de las publicaciones recibirá la mayoría de las citas, el *Factor de Impacto de las Revistas (Impact Factor)*, ideado por (Garfield, 1999), mide la frecuencia en que un artículo promedio de una revista ha sido citado en un año determinado. Es evidente que el Factor de Impacto indica la categoría científica de una revista. De ahí la importancia que tiene para un autor publicar sus artículos en revistas con un elevado Factor de Impacto, ya que conseguirán una gran *visibilidad*.

Así la siguiente lista permite ver qué se puede analizar mediante un estudio de citas.

- Identificar las tendencias y el crecimiento del conocimiento en las distintas disciplinas.
- Estimar la cobertura de las revistas secundarias.
- Identificar los usuarios de las distintas disciplinas.
- Identificar autores y tendencias en distintas disciplinas.
- Medir la utilidad de los servicios de disseminación selectiva de información.
- Predecir las tendencias de publicación.
- Identificar las revistas núcleo de cada disciplina.
- Formular políticas de adquisición ajustadas al presupuesto.
- Adaptar políticas de descarte de publicaciones.
- **Estudiar la dispersión y la obsolescencia de la literatura científica.**
- **Diseñar normas para estandarización.**
- **Diseñar procesos automáticos de indización, clasificación y confección de resúmenes.**
- **Predecir la productividad de editores, autores individuales, organizaciones, países.**

### 1.1.3 *Citaciones de patentes y familias de patentes*

Dentro de la solicitud de una patente ésta también lleva información citacional logrando así obtener familias de patentes. Dichas referencias son un conjunto de patentes o artículos científicos referidos que el solicitante vincula reconociendo la utilidad de estas en su patente. Una patente muy citada es, probablemente, más útil que una patente aislada.

### 1.1.4 *Los indicadores relacionales e índices*

Mientras los indicadores anteriormente mencionados proporcionan datos sobre el volumen y el impacto de las actividades de investigación, los indicadores relacionales se proponen conocer las relaciones y las interacciones entre los diferentes elementos bibliográficos: investigadores, campos, sectores..., intentando describir el contenido de las actividades y su evolución (Callon, 1993).

A continuación los indicadores relacionales más utilizados: *a) las citas o cocitaciones y b) la coocurrencia de palabras (co-word analysis).*

### 1.1.5 *cocitaciones*

El análisis de las cocitaciones detecta la aparición simultánea de dos citas que se repiten en gran número de artículos. De la cogitación de artículos se puede pasar a la cocitación de autores. El análisis contabiliza el número de coapariciones de parejas de referencias citadas. Así, si un artículo de un autor A y otro de un autor B son citados conjuntamente por un autor C, estamos ante un ejemplo de cocitación.

La frecuencia de la cocitación mide el grado de asociación entre dos documentos o dos autores y permite identificar grupos (*clusters*) de artículos próximos que son cocitados frecuentemente y revelar las líneas de investigación de una determinada área.

## 1.2. Estudios sociales de la ciencia y análisis de redes.

## 1.3 Teoría del actor red y sociología de la traducción.

La cienciometría estudia los aspectos cuantitativos de la ciencia como disciplina o actividad, forma parte de la sociología de la ciencia y encuentra aplicación en el establecimiento y evaluación de políticas de ciencia y tecnología, vigilancia tecnológica, esta tiene en cuenta los aspectos cuantitativos de la ciencia, la comunicación científica y la política, así la cienciometría emplea, técnicas y métricas para la evaluación de la ciencia.

Esta enfoca sus estudios en el crecimiento cuantitativo de la ciencia (Número de investigadores, número de publicaciones), en el desarrollo de las disciplinas (estudio de palabras asociadas), en la relación entre ciencia y tecnología (estudios sobre innovación), en la estructura de comunicación entre los científicos (estudio de citas), en las relaciones entre el desarrollo científico y el crecimiento económico (porcentaje del PIB destinado a la investigación).

Así la cienciometría es justamente el proyecto de cuantificar la producción científica, dar el estado de la ciencia y la tecnología, sin embargo esta no se preocupa por la unidad de análisis cognitiva, esta tiene un enfoque netamente estadístico teniendo como unidad de contabilidad las publicaciones y citas, sin embargo el documento se plantea el problema de la adquisición, representación y gestión del conocimiento difundidos bajo una forma escrita a través de las publicaciones científicas.

En este sentido la teoría del actor red planteada por Michelle Callon y Latour investigadores de la escuela de minas de París (Ecole des Mines de Paris), es de gran ayuda si planteamos la construcción de conocimiento como una construcción social (Leydesdorf, 2001), así observamos los procesos cognitivos dentro de un sistema de relaciones en el que participan entidades sociales y documentales que permiten buscar los elementos para la identificación del tejido –vínculos- en una comunidad científica, esta se preocupa por la observación, descripción y medición de las estructuras de relaciones que tiendan a construir vínculos coordinados y consolidados para la producción de nuevo conocimiento.

### *Teoría del actor red.*

La noción metodológica de red (Latour, 1996) permite tener en cuenta no solo las relaciones sino también el tipo de conexión logrando identificar la estabilidad, intensidad y conectividad para luego explicar el tipo de conexión que se establece entre los actores participantes de la comunidad. Las relaciones identificadas no están dadas e inscritas definitivamente no son entendidas en el sentido orgánico de las relaciones sistémicas sino son basadas en la noción de la sociología de la traducción

Los actores (individuales y colectivos, humanos y no humanos) trabajan constantemente para traducir sus lenguajes, sus problemas, sus identidades o sus intereses. Es a través de este proceso que el mundo se construye y se deconstruye, se estabiliza y se desestabiliza. Por esto “la identidad de los actores y sus tallas respectivas son situaciones, apuestas permanentes en las controversias que se desarrollan” (Callon, 1986:174).

Los actos de traducción ejercidos por la comunidad, nos dan la posibilidad de identificar vínculos. Una comunidad existe si entre sus actores encontramos actos de negociación -problematización, de interesamiento, enrolamiento y de movilización-. Entonces la comunidad entendida como actores alrededor de un tema, quienes reconocen un problema generan procesos de interesamiento que ajustan a los actores en unos roles a través de mecanismos de negociación, enrolamientos que se producen por un conjunto de estrategias definidas por propósitos e interrelaciones entre los actores. Así se vinculan y se ajustan los actores en la red. La traducción es un proceso constante, que permite que la comunidad se vincule en preocupaciones comunes, estas relaciones de poder (poder hacer) son las que deseo encontrar en la comunidad aunque estos actos de traducción hayan terminado en casos fallidos.

“La noción de red intenta la interpretación del establecimiento, nunca definitivo y en continua construcción, de las relaciones entre humanos y objetos. Pero la estabilización de las formas de la vida social debe ser considerada más como un punto de llegada que como un punto de partida del análisis. Se trata de reabrir las cajas negras (lo que va de sí ya no es interrogado como un hecho científico, una técnica, un procedimiento o una institución) cerradas por los actores. La red es el resultado más o menos solidificado de procesos de traducción y de su inscripción en “cajas negras”; “la palabra indica que los recursos están todos concentrados en algunos lugares – los nudos – pero que estos nudos están ligados unos con otros por mallas; gracias a estas conexiones”.(Arellano, 1989).

Así podríamos decir que la comunidad se construye en diferentes espacios, la comunidad en constante construcción es definida por la interacción constante entre actores heterogéneos, y son sus relaciones de poder las que como resultado permite el identificarse e identificar la comunidad como una red donde el reconocimiento de sus actores y la identificación de los conceptos, actos vinculantes y los diferentes espacios comunes permiten representar la red propia mente dicha. La red es la identificación de una comunidad-red entendida en términos de la construcción de vínculos coordinados y consolidados, esta no es construida a priori, es a través de la observación de la comunidad que observamos la red y no observamos la red a través de la comunidad.

La coordinación de las redes es mediada a través de mecanismos de inscripción de las mismas en objetos tecnológicos o en medios donde circula el conocimiento codificado como artículos o papers, esta intermediación provee a la red de formas de coordinación logrando la convergencia o grado de acuerdo o desacuerdo. Grado utilizado por una serie de traducciones que permiten la alineación y la coordinación de la red.

La convergencia de la red y el número de relaciones y actores definen la frontera y por ende la comunidad y son sus actos de traducción los que definen la complejidad de la estructura de relaciones, en este sentido la comunidad-red es delimitada por los actos de

traducción existentes y es la convergencia la que permite la identificación de la estructura relacional.

Las estructuras relacionales son representadas a través de grafos así la unidad de análisis cognitiva es representada mediante mapas de conocimiento, estos se construyen por nodos (conceptos) y vínculos (relación entre los conceptos), o las relaciones sociales representadas por socio gramas así apoyado de las teorías de la sociología de la ciencia su estudio se basa en el concepto estructural de la conformación de comunidades y como dichas comunidades se relacionan y se enrolan en relaciones de poder. De esta forma podemos ahondar en la construcción de análisis de comunidades científicas, desarrollando el análisis mediante el estudio de los procesos de producción de conocimiento y articuladamente el análisis de estructuras sociales las comunidades científicas.

#### 1.4 Análisis socio-cognitivo.

La observación de las métricas de producción científica permiten estudiar los aspectos generales del fenómeno científico-tecnológico para el caso de los indicadores de producción, por ejemplo índices de citas científicas<sup>1</sup> los cuales son creados para evaluar el impacto social de la producción de un investigador, permiten dar cuenta de factores generales y no pueden responder a preguntas tales como el proceso de producción de conocimiento en el sentido de las teorías, métodos, técnicas, etc , que se relacionan o en el sentido de las relaciones sociales que interviene en la construcción de dicho conocimiento, así podemos hacer la pregunta de si el indicador es una construcción que permite medir los procesos de producción de conocimiento, sin embargo podemos hacer una representación mas compleja del problema haciendo una análisis reticular, así las redes de conceptos o mapas de conocimiento o socio gramas.

Este documento desarrolla la teoría del actor red planteada por Michelle Callon y Latour investigadores de la escuela de minas de Paris (Ecole des Mines de Paris), dicha teoría plantea buscar los elementos para la identificación del tejido –vínculos- en una comunidad científica, se preocupa por la observación, descripción y medición de las estructuras de relaciones que tiendan a construir vínculos coordinados y consolidados para la producción de nuevo conocimiento. Las relaciones son evidenciadas a través de los actores y su reconocimiento, el cual, permite identificar a aquellos que han construido y construyen la comunidad.

#### 1.5 Análisis estadísticos descriptivos y métodos de análisis textuales

##### 1.5.1 *Método de las palabras asociada*

El método o análisis de las palabras asociadas es una herramienta cuantitativa desarrollada inicialmente en el Centre de Sociologie de l'Innovation (CSI) de l'Ecole Nationale Supérieure de Mines de Paris y en el Institut de l'Information Scientifique et Technique (antiguo CDST) del CNRS. Este método visualiza la estructura de las redes de palabras, de acuerdo con la teoría actor-red, y calcula una serie de parámetros que nos permiten estudiar el comportamiento de cada uno de los actores, tanto en su aspecto puramente estructural como en su aspecto evolutivo o dinámico. Para la puesta en

---

<sup>1</sup> ISI – índice de citaciones

marcha de este método se desarrolló un conjunto de programas informáticos denominado LEXIMAPPE (1988).

Leximappe se aplica a todo tipo de documentos indizados mediante palabras clave y en especial a los artículos científicos y técnicos, patentes, etc. Por tanto, la gran ventaja que aporta el método de las palabras asociadas frente al análisis de co-citas radica en que el primero puede tomar la información de diferentes base de datos SCI, SSCI, AHCI, MEDLINE, BIOSIS mientras que el segundo está limitado prácticamente a la utilización de las bases SCI, SSCI, AHCI .

*i) Matriz de ocurrencias y Matriz de asociaciones*

El método de las palabras asociadas tiene en cuenta la ocurrencia de palabras así el contenido de un documento es definido dichas palabras o palabras clave. Se parte, por tanto, de una matriz de datos "documentos x palabras clave", denominada matriz de ocurrencias, que representaría el contenido conceptual del campo científico en estudio (COURTIAL, J. P. y MICHELET, B., 1990). En las celdas la matriz tiene un uno en el caso de que la palabra "i" este en el documento "j". El número de veces que una palabra clave "i" aparece u ocurre se denota por  $C_i$ .

	Pal.1	pal.2	pal.i	pal.j	Pal.1000
Doc.1	1	0	0	1	0
Doc.2	1	1	0	0	0
Doc.i	0	1	1	0	0
Doc.j	1	0	0	0	0
Doc.3000	1	0	0	0	1
	120	98	25	20	3
	$c_1$	$c_2$	$c_i$	$c_j$	$c_{1000}$

Así dos palabras co-ocurren cuando aparecen simultáneamente en el mismo documento. Dos palabras estarán más ligadas o asociadas entre sí cuanto mayor sea la co-ocurrencia entre ellas. Por tanto, la medida del enlace entre dos palabras de una red será proporcional a la co-ocurrencia de esas dos palabras en el conjunto de documentos que se tome como muestra.

	Pal.1	pal.2	Pal.i	Pal.j	Pal.1000
Pal.1	-	20	20	0	2
Pal.2	-	-	0	5	0
Pal.i	-	-	-	20	0
Pal.j	-	-	-	-	0
Pal.1000	-	-	-	-	-

La matriz de asociaciones, de co-ocurrencias o de "palabras clave x palabras clave" es una matriz de adyacencia cuadrada simétrica. Cada elemento representa la asociación entre los descriptores. En la celda  $C_{ij}$  colocamos el número de documentos en los que la palabra "i" y la palabra "j" aparecen simultáneamente(Moreno B,1998).

*iv) Redes de palabras y temas, agrupaciones y subredes*

La matriz de asociaciones normalizada es la matriz de adyacencia del grafo que representa la red. Cada nodo de este grafo es un descriptor o palabra y cada índice de equivalencia entre cada dos descriptores es la ponderación de los arcos que une estas parejas de vértices.

Así lo que se busca con esta representación es extraer de la red de palabras, agrupaciones o subredes significativas. Estas subredes representarían los temas de investigación y definirían los temas que forman la red global, así como también la capacidad de calcular parámetros que cuantifiquen la red y los caractericen según suposición estratégica y poder seguir su evolución temporal o dinámica de la temática, por último la red debe tener estabilidad frente a la posibilidad de errores de indexación.

*ii) Algoritmo de clasificación por enlace simple.*

Los elementos de la matriz de asociaciones son ordenados en una lista decreciente según su índice de equivalencia. Esta lista está formada tan solo por aquellas palabras que tengan una ocurrencia mínima y pares de asociaciones también con una co-ocurrencia mínima preestablecidas. El programa recorre la lista desde el principio y va construyendo duplas, tripletes, etc. de palabras asociadas de forma que suministra un grafo conexo que no exceda de un valor máximo de palabras preestablecido (por ejemplo 10 ó 15) Cada vez que se obtiene un grafo, elimina las palabras de éste de la lista y comienza el proceso de construcción de nuevos grafos hasta agotar el total de palabras disponibles.

*iii) Algoritmo de agrupación sobre centros simples.*

Este algoritmo también ordena los pares de asociaciones por orden decreciente de índice de equivalencia y sólo pueden formar parte de esta lista las palabras con una ocurrencia mínima y los pares con una co-ocurrencia mínima establecidas previamente. Se inicializa un contador para cada descriptor y comienza a recorrer la lista desde el principio incrementando el contador de las palabras que van apareciendo. Cuando el contador de una palabra alcanza un valor igual al número de palabras máximo estipulado para los temas menos uno, el algoritmo toma esta palabra como centro de una agrupación. El conjunto resultante estará formado por las uniones de esta palabra central y todas aquellas que se han asociado con ella. El resultado es una estructura en forma de estrella. Las palabras que han aparecido se eliminan de la lista y se comienza de nuevo el proceso para generar más agrupaciones. Si después de recorrer toda la lista ningún contador llega al valor máximo preestablecido, éste se disminuye en tantas unidades como sea necesario para formar una nueva agrupación. El proceso finaliza cuando el valor máximo del contador disminuya hasta un valor mínimo preestablecido o se terminen todas las palabras de la lista ordenada de pares.

La esencia de las redes cuantitativas y sociocognitivas es la de la presencia de fronteras difusas, por lo que no es de extrañar que no sea posible definir las exactamente. Según el algoritmo utilizado, trazaremos más hacia un lado o hacia otro de la frontera difusa, la línea divisoria que nos servirá de referencia, pero debe entenderse que esta línea es sencillamente una guía para adentrarnos de forma simplificada en el estudio de las redes que de por sí son muy complejas, cuando se realizan mapas de la ciencia, se pretende representar sobre un plano, las relaciones complejas y multidimensionales de una red de difícil representación.

### *1.5.2 Co-ocurrencia de palabras*

La coocurrencia de palabras estudia la aparición conjunta de dos o más palabras representativas en campos tales como títulos de artículos o de patentes, resúmenes o *abstracts*, palabras clave (*key words*) de artículos (descriptores e identificadores), códigos de clasificación, reivindicaciones (*claims*) de patentes o bien directamente el texto libre. La repetición de dos palabras juntas -por ejemplo, como descriptores o bien en las palabras de los títulos- en muchos artículos, indica también una relación o *proximidad* entre ellas.

La frecuencia de aparición conjunta de dos indicadores que pueden ser o no de la misma naturaleza (palabras clave, nombres de organizaciones, autores, años, citas, etc.) mediante análisis que se denominan de coocurrencia.

La coocurrencia de palabras consiste, pues, en la detección de las palabras que caracterizan el contenido de los trabajos sobre un tema y en contar la coaparición de éstas. Los conceptos de *proximidad* o *lejanía* se pueden representar gráficamente, lo que constituyen la base para la elaboración de los *mapas* conceptuales.

Un trabajo realizado en esta área es el Ketan Mane titulado “Mapping topics and topic bursts in PNAS”, que trata el tema del análisis de textos científicos mediante el método de palabras asociadas, toma como fuente de datos los proceedings de la national academy of science (PNAS), obtiene grafos de los temas haciendo un corte en el 50% de co-ocurrencia y los representa por años, el artículo también se centra en el problema de la visualización, utilizan el fruchterman-reingold 2D un algoritmo de visualización ya implementado en algunos software como Pajek.

#### 1.6. Análisis de Redes sociales-ARS-

El análisis de redes sociales, es una metodología de análisis cuantitativo y estructuralista que busca reconocer las relaciones y sus estructuras para poder encontrar en este sistema de relaciones y de actores, comportamientos y en si la estructura -o estructuras- social de dicha comunidad analizada; 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.

##### 1.6.1 Elementos conceptuales

- i) Actores. Son las unidades sociales o entidades. En los socio gramas o grafos estos se identifican por los nodos.
- ii) Relaciones. Estas son las representaciones de los actos vinculantes, establecen un vínculo entre un par de actores, ejemplo asociación, citas, trabajos en común, afiliación. Son representadas como las aristas en los grafos.
- iii) Un grafo  $G$  consiste en dos conjuntos de información: un conjunto de nodos,  $N = \{n_1, n_2, \dots, n_g\}$  y un conjunto de lazos,  $L = \{l_1, l_2, \dots, l_h\}$  entre pares de nodos. En un grafo hay  $g$  nodos y  $h$  lazos. Un grafo se representa como  $G(N, L)$ . Se dice que dos nodos son adyacentes si el lazo  $l_k = (n_i, n_j)$  está

incluido en el conjunto de lazos  $L$ . Esta representación de la red permite reconocer propiedades estructurales de las relaciones y sus actores

- iv) Red social, conjunto de actores vinculados entre si por una estructura de relaciones, dichos actores colectivos pueden vincularse de diferentes modos, adyacencia, afiliación o atributos.

### 1.6.2 Medidas de centralidad

El análisis de redes sociales tiene la potencialidad de permitir hacer una representación formal de las estructuras sociales, para así construir medidas que den cuenta del estado de la estructura representada, las medidas son tomadas de (Wasserman,2004).

- i) Densidad: Mide la proporción de lazos existentes en relación con los posibles.

$$D=2L/[n(n-1)]$$

Donde L es el numero total de lazos y n el numero total de nodos.

- ii) n – clique: esta medida se refiere al agrupamiento de los actores, el cual puede ser definido como: un actor es miembro de un grupo si está conectado con todos los miembros del grupo a una distancia mayor que uno, usualmente se utiliza la distancia de trayecto dos.

- iii) Cercanía: el índice de la cercanía de un nodo es una medida del nodo con el resto de la red. Para ello se calcula la suma de los geodésicos (o caminos más cortos) que unen a cada vértice o nodo con el resto de la red.

El índice relativo de la centralidad proximidad de un punto  $RC(i)$ , para el punto i es  $RC(i) = (n-1)/D_{i+}$ , donde  $D_{i+}$  es la suma de las distancias desde i a todos los demás puntos, que puede ser representado como la suma de las filas i de la matriz de distancias  $D_{i+}$ .

$$D_{i+} = \sum_{j=1}^n D_{ij}$$

Intermediación: Para todos los puntos no ordenados, i,j,k, donde  $i < j$ , n es el número de nodos de la red y  $g_{ij}(k)$  es el número de geodésicas (caminos más cortos) entre i y j, que pasan por k. Por tanto si k está en el camino más corto del par (i,j), K tiene alta centralidad-mediación.

$$C_B(K) = \frac{2 \sum_{i=1}^n \sum_{j=1}^n (g_{ij}(k) / g_{ij})}{n^2 - 3n + 2}$$

Las medidas aquí presentadas representan las propiedades estructurales de las relaciones a través de las medidas de centralidad, la proximidad o cercanía, es interpretada como la capacidad de comunicación con el conjunto de la red, y la mediación como aquel actor intermediador a través del cual se comunican otros actores. Es claro identificar quiénes

tienen más oportunidades y mejores posiciones dentro de la estructura relacional. En este sentido, la medida de centralidad es una métrica de la capacidad de vincularse de un actor con el total de la red, y la intermediación es la medida de cómo un actor es intermediador dentro de la estructura de relaciones ya que este es un paso obligado para la vinculación de otros actores. En este aparte hay que hacer un trabajo encaminado a identificar cuales son los algoritmos y métodos útiles a la investigación, como por ejemplo neighbor, eigenvector, CONCOR, técnicas de cluster.

#### *Aplicaciones y herramientas.*

Existen varias herramientas disponibles para hacer ARS, estas se caracterizan por servir para graficar o para construir medidas o hacer las dos funciones, las mas comerciales y utilizadas son: Usinet, Netminer, Pajec, Socnet, Jung, Social Network Visualiser For Linux (Socnetv), NV2D.

#### 1.6.3 Medidas de cohesión

#### 1.6.4 Equivalencia estructural

### 1.7. Redes semánticas y semántica latente.

La representación del conocimiento se hace posible a través de redes de conceptos o mapas cognitivos, dichos sistemas de organización del conocimiento estructuran conceptos no jerárquicamente sino como red. Los conceptos o nodos, se relacionan y se vinculan a través de arcos, estas relaciones pueden ser del tipo todo-parte, causa-efecto, padre-niño, es\_un o es\_parte. Así las redes semánticas son grafos orientados que proporcionan una representación declarativa de objetos, propiedades y relaciones.

Una red semántica se puede definir como un grafo dirigido, con etiquetas en los arcos y en los nodos, el nodo representa el concepto y la etiqueta en el arco representa la clase de relación que existe entre los nodos.

$$G = (N, R, T_N, T_R, \varphi_1, \varphi_2)$$

Así los nodos  $N$ , son un conjunto no vacío, con un conjunto  $R \subset M \times N$ , de pares entre los nodos, estos son llamados arcos,  $T_N$  y  $T_R$  conjuntos de elementos llamados etiquetas de los nodos y de relaciones respectivamente y las funciones  $\varphi_1, \varphi_2$ .

$$\varphi_1 : N \rightarrow T_N \text{ y } \varphi_2 : R \rightarrow T_R$$

Si  $\alpha \in T_R$ , entonces  $R_\alpha = \varphi_2^{-1}(\alpha)$

Las redes semánticas tiene la potencialidad de permitir construir reglas de inferencia, logrando así agregar o eliminar etiquetas de los arcos, obteniendo grafos extendidos o grafos reducidos. Un grafo que no puede ser extendido se llama completo y un grafo que no puede ser reducido se llama kernel.

Las redes semánticas son estructuras que permiten representar el conocimiento incluido en un documento, de esta forma los mapas conceptuales (Novak, 1988), son una forma de representar la estructura cognitiva, permiten organizar jerárquicamente a través de

redes de proposiciones conceptos y relaciones entre estos. Novak y Gowin (1988) definen los mapas conceptuales como “recursos esquemáticos para representar un conjunto de significados conceptuales incluidos en una estructura de proposiciones” dichas proposiciones se estructuran formando una jerarquía de inclusión.

Un mapa conceptual, consta de diferentes tipos de relaciones, derivativa, correlativa, supraordinaria o combinatoria. El tipo de enlace explica el tipo de significancia de la relación, además los enlaces pueden ser, directos, recíprocos, generar enlaces cruzados, contribuyendo con el “grado” de significancia de la relación.

Los esquemas de representación son los modelos formales de representación en los cuales encontramos los elementos básicos de las redes de representación de conocimiento o que comúnmente denominados redes semánticas, estos son:

- i) Estructuras de datos en nodos, que representan conceptos, unidas por arcos que representan las relaciones entre los conceptos.
- ii) Un conjunto de procedimientos de inferencia que operan sobre las estructuras de datos.

#### 1.7.1. Modelos y Conceptos.

Como veremos más adelante las redes semánticas se clasifican según su esquema de representación, sin embargo lo importante es que estos esquemas comparten características fundamentales, entre las que se destacan la herencia por defecto. En una red semántica, los conceptos (estructuras, clases, marcos, dependiendo del esquema concreto) están organizados en una red en la que existe un nodo superior (T) al que se le asigna uno o varios nodos hijos, que a su vez tienen otros conceptos hijos y así sucesivamente hasta que se alcanza el final, cuyos nodos pueden ser o bien conceptos, o bien instancias. El concepto de herencia es fundamental para entender el funcionamiento de las redes semánticas. Siguiendo a Shastri (1988), definimos la herencia como el sistema de razonamiento que lleva a un agente a deducir propiedades de un concepto basándose en las propiedades de conceptos más altos en la jerarquía.

- i) Redes IS-A, en las que los enlaces entre nodos están etiquetados.

Las redes semánticas son entendidas normalmente como redes IS-A, y muchas veces se menciona como un sinónimo de red semántica. Una red IS-A es una jerarquía taxonómica cuya columna está constituida por un sistema de enlaces de herencia entre los objetos o conceptos de representación, conocidos como nodos. Las redes IS-A son el resultado de la observación de que gran parte del conocimiento humano se basa en la adscripción de un subconjunto de elementos como parte de otro más general. Las taxonomías clásicas naturales son un buen ejemplo: un perro es un cánido, un cánido es un mamífero, un mamífero es un animal.

Algunos problemas y desventajas importantes: 1) la elección de los nodos y arcos es crucial en la fase de análisis, 2) una vez que se ha decidido una estructura determinada, es muy complicado cambiarla, 3) dificultad para expresar cuantificación. Por ejemplo en expresiones tales como "algunos pájaros vuelan" o "todos los pájaros pían".

- ii) Grafos conceptuales: en los que existen dos tipos de nodos, de conceptos y de relaciones.

Los problemas anteriormente mencionados, llevaron a buscar una idea más compleja de representación con una estructura más compleja que simples nodos y arcos, que fuesen capaces de dar cabida a éstas y otras situaciones. Concretamente, John Sowa (1984) propuso los grafos conceptuales. En estos se representan las relaciones y los conceptos como nodos, esto permite el agrupamiento conceptual de una colección de textos representados por un conjunto de grafos conceptuales, que son una representación simple pero con mayor información del contenido de los textos. Este método emplea una estrategia de aprendizaje no supervisado, que construye incrementalmente una jerarquía de los grafos conceptuales. Además este método incorpora algunas características que lo hacen atractivo para la minería de texto. Por ejemplo: considera toda la información estructural de los grafos conceptuales, emplea conocimiento del dominio, y considera los intereses del usuario.

- iii) Redes de marcos (frames): en los que los puntos de unión de los enlaces son parte de la etiqueta del nodo.

De los tres tipos de redes semánticas, el esquema basado en marcos es el que permite una mayor flexibilidad.

Así, un marco es una estructura de datos compleja que representa una situación estereotipo. Cada marco posee un número de casillas donde se almacena la información respecto a su uso y a lo que se espera que ocurra a continuación. Al igual que las redes semánticas, podemos concebir un marco como una red de nodos y relaciones entre nodos (arcos). Una base de conocimiento basada en marcos es una colección de marcos organizados jerárquicamente, según un número de criterios estrictos y otros principios más o menos imprecisos tales como el de similitud entre marcos. A nivel práctico, podemos considerar los marcos como una red semántica con un número de posibilidades mucho mayor, entre las que destacan especialmente, la capacidad de activación de procesos (triggering) y de herencia no-monotónica mediante sobrecontrol (overriding), en la que un nodo hijo hereda todas las casillas de su padre a menos que se especifique lo contrario.

Principales características de los marcos:

**Precisión:** los objetos, las relaciones entre objetos y sus propiedades se describen de forma precisa; en ausencia de evidencia contraria se usan valores por omisión.

**Activación dinámica de procesos (Triggering):** es posible adjuntar procedimientos a un marco o alguno de sus componentes de forma que se llamen y ejecuten automáticamente tras la comprobación de cambio de alguna propiedad o valor (p. ej. IF-NEEDED, IF-ADDED).

**Herencia por defecto no-monotónica:** los marcos están conceptualmente relacionados, permitiendo que los atributos de los objetos sean heredados de otros objetos predecesores en la jerarquía.

**Modularidad:** la base de conocimiento está organizada en componentes claramente diferenciados.

Observando los modelos mencionados anteriormente, podemos preguntarnos si en realidad es necesario emplear un complejo sistema de representación del conocimiento, en la práctica, sólo es indispensable un sistema que permita la creación de jerarquías conceptuales, la asignación de atributos, características a los conceptos y la explicitación de relaciones entre conceptos, garantizando la coherencia interna del conjunto de conceptos: por ejemplo, que no existan conceptos repetidos, ni estructuras circulares, además, por supuesto, de permitir la asignación de términos a estos conceptos.

### 1.7.2. Semántica Latente.

El Análisis Semántico Latente (LSA), considerado en sus inicios como una teoría y un método de representación del conocimiento humano, es en sí un método de análisis semántico basado en un modelo estadístico del uso de palabras que permite comparar las similitudes semánticas entre piezas de información textual. Esta comparación se realiza en un espacio semántico multidimensional, generado a partir de un valor singular de descomposición (SVD). Con este análisis es posible determinar las distancias y relaciones entre palabras, palabras y párrafos, y entre párrafos. Esta forma de análisis permite abordar el problema de la representación de los procesos de producción de conocimiento por medio de inferencias de relaciones a partir de grandes cantidades de textos.

Esta teoría descansa en la noción de que algunos dominios de conocimiento contienen diversos números de interrelaciones débiles o latentes, que si son aprovechadas se pueden amplificar produciendo aprendizaje a través de procesos de inferencia. El método de inducción propuesto depende de la reconstrucción de un sistema de relaciones de similitud múltiples en un espacio multidimensional. En la co-ocurrencia de eventos, en particular de palabras, en contextos locales se generan y se reflejan por su similitud en algún lugar de este espacio multidimensional. Utilizando los métodos estadísticos referidos, se concluye que el LSA puede usarse para predecir fenómenos tales como la coherencia textual, comprensión, desambiguación contextual de homógrafos y generación del significado central inferido de un párrafo. Se define al Análisis Semántico Latente como una teoría y un método para extraer y representar el significado contextual en uso de palabras a través de computación estadística aplicada a un gran corpus textual (Landauer y Dumais, 1997).

También, "es una técnica matemático-estadística totalmente automática para extraer e inferir relaciones de uso contextual esperado de palabras en pasajes de discurso. No es un procesamiento de idioma natural tradicional o programa de inteligencia artificial; no usa ningún diccionario construido humanamente, bases de conocimiento, redes semánticas, gramáticas, segmentadores sintácticos, o morfologías y toma como input sólo la segmentación del texto en palabras, pasajes, frases o párrafos" (Landauer, Foltz y Laham, 1998).

En trabajos más actuales, Kintsch (2000) define al LSA como un procedimiento totalmente automático de técnicas matemáticas estándar que sirve para analizar un gran corpus de texto digitalizado. Este intenta integrar esta técnica a su teoría de Construcción e Integración como una herramienta que le permita representar la

macroestructura como vectores en el espacio. Finalmente, Landauer (2002) establece que el LSA es un modelo de semántica natural del idioma, sin embargo, plantea:

"If any of my presentations of LSA have given cause to believe that LSA is a to be considered a complete theory of language and knowledge, or even lexical semantics, I regret it profoundly. LSA is a theory of (about) those things, but not of everything about them" (Landauer, 2002: 32).

Para cerrar esta discusión se puede decir que el método es netamente una técnica matemático-estadística que permite la creación de vectores multidimensionales para el análisis semántico de las relaciones existentes entre palabras, palabras y párrafos y entre párrafos, por cuanto tanto es válido para representar el conocimiento expuesto en documentos científicos.

### *Clasificación automática de textos*

Este es uno de los problemas de la IA, del machine learning y del IR, la categorización automática puede entenderse como un proceso de aprendizaje, durante el cual un programa capta las características que distinguen cada categoría o clase de las demás, es decir, aquellas que deben poseer los documentos para pertenecer a esa categoría. Así se busca la construcción de vectores patrón que contengan las características de distintas clases o categorías de documentos, utilizando técnicas basadas en aquellas aplicadas en la expansión de consultas por relevancia.

Estas características no tienen por qué indicar de forma absoluta la pertenencia a una clase o categoría, sino que más bien lo hacen en función de una escala o graduación. De esta forma, por ejemplo, documentos que posean una cierta característica tendrán un factor de posibilidades de pertenecer a determinada clase. De modo que la acumulación de dichas cantidades puede arrojar un resultado consistente en un coeficiente asociado a cada una de las clases existentes. Este coeficiente lo que expresa en realidad es el grado de confianza o certeza de que el documento en cuestión pertenezca a la clase asociada al coeficiente resultante.

Las técnicas de recuperación de información son usadas en tres fases:

BI-estilo: indexar los documentos a partir de un corpus inicial para su posterior clasificación.

BI-estilo técnicas para hacer búsquedas y refinar búsquedas se utiliza en la construcción inductiva de clasificadores.

BI-estilo evaluación, es la evaluación de clasificación la efectiva.

### *D El Modelo Vectorial y la Representación de Categorías*

Un documento puede considerarse como un vector  $D = (c_1, c_2, c_3 \dots c_j)$ , es decir, como un conjunto de características, hasta un total de  $j$ , y en el cual  $c_1$  es un valor numérico que expresa en qué grado el documento  $D$  posee la característica 1,  $c_2$  lo mismo para la característica 2, y así sucesivamente. El concepto 'característica' suele concretarse en la ocurrencia de determinadas palabras en el documento, aunque nada impide tomar en consideración otros factores.

Se han propuesto diversos sistemas para calcular dicho valor numérico, es decir, el peso de cada término contemplado, para cada documento. En general, se tiene en cuenta para esto la frecuencia inversa (IDF), combinándola de alguna forma con la frecuencia del término dentro del documento.(Salton y Buckley, 1987).

El objetivo del IR es construir vectores patrón representativos de cada categoría o clase. Muchos sistemas aplican un mecanismo de realimentación, a través del cual, después de una primera consulta y sus correspondientes resultados, utilizan aquellos documentos señalados por el usuario como más relevantes para reformular de forma automática la consulta, extrayendo términos de estos documentos más relevantes y añadiéndolos a la consulta original y recalculando los pesos de los términos.

Así pues, si disponemos de una colección de documentos categorizados manualmente, y adscritos a una clase determinada, es posible aplicar dichos mecanismos de realimentación para construir un vector patrón, representativo de esa clase. Los nuevos documentos a categorizar pueden ser confrontados con ese vector patrón, calculando la similitud entre ambos y, en función del grado de ésta, se puede determinar su asignación o no a esa clase. Diversos sistemas se utilizan en los procesos de realimentación, para construir un nuevo vector de consulta, y que pueden ser aplicados a la categorización, para construir los vectores patrón de cada clase o categoría.

#### 1.7.3. Las redes semánticas – conceptuales.

#### 1.7.4. Tesoros, Ontologías y Estándares.

##### *E. Tesoros.*

Schauble propuso una nueva estructura de la información: el espacio conceptual. Este construyó una teoría de tesoros conceptuales, expuesta como un sistema formal mediante la lógica matemática: el dominio algebraico. Este sistema revela una estrecha relación entre los tesoros y el modelo espacial, y en ella las relaciones entre términos son definidas con mayor precisión que en los tesoros jerárquicos. Lo que es más interesante para nuestros propósitos es la construcción automática de tesoros conceptuales, cuyos desarrollos descansan básicamente en el análisis estadístico de la frecuencia de las palabras que componen los textos.(lopez,2000).

La organización de la información mediante los tesoros conceptuales permite mejorar la precisión en las recuperaciones con lenguaje natural, y su integración con las reglas de la base de conocimientos de un agente experto.

Así, los tesoros son una macroserie del sublenguaje controlado en un dominio científico específico, que se usan durante: El proceso de indexación, como ayuda en la identificación de los conceptos y en el proceso de recuperación, como fuente de nuevos términos que identifiquen conceptos y aumenten la precisión de las búsquedas.

Las redes semánticas - conceptuales:

1) Facilitan el encaminamiento de las hojas hacia una raíz y viceversa, gracias a la función de orientación presentada por las facetas (clases de discriminantes).

2) Permite la transición de un árbol hacia otro a través de las relaciones asociativas (nodos, polijerárquicos).

3) Permite la extensión a un árbol completo, la restricción a un subárbol, o incluso la restricción a árboles parciales

4) Permite la exploración transversal de la red, buscando configuraciones correspondientes a una o varias facetas en una reunión o en una intersección de árboles.

#### *F. Ontologías*

El término ontología, es tomado de la filosofía y la epistemología. La Ontología es una rama de la metafísica que se ocupa del estudio de la naturaleza de la existencia, de los seres y de sus propiedades transcendentales. Siguiendo esta línea, el término ontología se usa en el ámbito de la ingeniería del conocimiento para referirse a un conjunto de conceptos organizados jerárquicamente, representados en algún sistema informático cuya utilidad es la de servir de soporte a diversas aplicaciones que requieren de conocimiento específico sobre la materia que la ontología representa. Una ontología ha de interpretarse como un entendimiento común y compartido de un dominio, que puede comunicarse entre científicos y sistemas computacionales. Podemos decir que el sinónimo comúnmente utilizado de ontología es conceptualización.

Así una ontología es una entidad computacional, y no ha de ser considerada como una entidad natural que se descubre, sino como recurso artificial que se crea (Mahesh 1996)

Entonces la ontología, es un cuerpo estructurado de conocimiento, el cual se basa en mayor o menor medida en un modelo de marcos, aunque, en general, no emplea sistemas de activación de procesos por tratarse de representaciones con un marcado carácter estático, las ontologías están enfocadas a surtir de información a otras aplicaciones que contienen los procedimientos.

“An ontology is a database describing the concepts in the world or some domain, some of their properties and how the concepts relate to each other”. Weigand (1997)

Así las ontologías como "sistemas de representación de conocimiento" se utilizan para especificar a qué tipo de sistemas nos referimos. En realidad, las ontologías se están empleando en todo tipo de aplicaciones informáticas en las que sea necesario definir concretamente el conjunto de entidades relevantes en el campo de aplicación determinado, así como las interacciones entre las mismas. Algunas ontologías se crean con el objetivo de alcanzar una comprensión de la Unidad del Discurso pertinente, ya que su creación impone una especificación muy detallada. Otras ontologías han sido creadas con un propósito general, como por ejemplo el proyecto Cyc (Guha & Lenat 1990), que está orientado a la construcción de una base de conocimiento que contenga el conocimiento humano necesario para hacer inferencias.

Tres tipos fundamentales de ontologías:

i) Ontologías de un dominio, en las que se representa el conocimiento especializado pertinente de un dominio

ii) Ontologías genéricas, en las que se representan conceptos generales y fundacionales del conocimiento como las estructuras parte/todo,

iii) Ontologías representacionales, en las que se especifican las conceptualizaciones que subyacen a los formalismos de representación del conocimiento, por lo que también se denominan meta-ontologías.

funcionalidades de las ontologías:

i) Almacenan las restricciones de selección y otros elementos de conocimiento del mundo.

ii) Ayudan en la resolución de ambigüedades semánticas y en la interpretación del lenguaje no literal, realizando inferencias basadas en la topología de la ontología para medir afinidad semántica entre significados.

iii) Suponen una herramienta para clasificar personas, lugares, roles sociales y organizaciones.

iv) Conforman el substrato sobre el que los significados de las palabras de cualquier lengua están fundados.

Las ontologías son precisamente el tipo de recurso independiente de la lengua que sirve de punto de encuentro entre dos o más lenguas, permitiendo una conceptualización muy concreta ya que debe ser hecha explícita y de forma detallada. asegurando, que todos y cada uno de los términos estén asignados a un concepto determinado. Cada uno de los conceptos, debe formar parte de una estructura bien definida y debe ser posible especificar diversos tipos de relaciones entre ellos.

El segundo aspecto, también fundamental, se refiere a la posibilidad de que un concepto pueda "ubicarse en más de una clase de conceptos", debido a la priorización de diferentes características o atributos. Esto es posible gracias a los mecanismos de herencia, ya que una de las ventajas de las ontologías de conceptos es que pueden diseñarse para dar soporte a mecanismos de herencia múltiple, por medio de la cual un concepto hijo puede asignarse a más de un padre y aparecer, de este modo, en lugares diferentes de la ontología.

i) Estándares

OWL es un lenguaje de marcado para la publicación de ontologías en la WWW y tiene como objetivo facilitar un modelo de marcado, construido sobre RDF y codificado en XML que permita representar ontologías a partir de un vocabulario más amplio y una sintaxis más fuerte que la que permite RDF<sup>7</sup>. Por este motivo OWL puede ser utilizado para representar de forma explícita el significado de términos pertenecientes a un vocabulario y definir las relaciones que existen entre ellos<sup>8</sup>.

RDF El fundamento o base de RDF es un modelo para representar propiedades designadas y valores de propiedades. El modelo RDF se basa en principios perfectamente establecidos de varias comunidades de representación de datos.

## **Capítulo 2, Técnicas para Construcción Automática de Mapas de Áreas Científicas**

La ciencia se desarrolla dentro de lo que se conoce como la sociedad del conocimiento, donde el conocimiento científico es traducido en productos científicos (artículos y patentes) y son dichos productos los que permiten a una sociedad seguir produciendo más conocimiento para de esta forma buscar el bienestar y desarrollo de la sociedad, así los investigadores, los empresarios y los tomadores de decisiones necesitan de herramientas informáticas que permitan hacer por una parte seguimiento y control y de otra parte, métricas y mapas que permitan dar cuenta de el estado y las dinámicas de la investigación científica. Una de las áreas que se encarga de esta tarea es la vigilancia tecnológica, esta requiere del esfuerzo sistemático y organizado de observación, captación, análisis, difusión precisa y recuperación de información sobre los hechos del entorno científico.

Así la sociedad ha construido servicios de artículos y patentes donde la producción de las comunidades científicas se ve reflejada, estos servicios se caracterizan por tener almacenado los meta datos de los artículos, estos meta datos son: palabras clave, autores, abstract, referencias y en algunos casos artículos completos. Dichos servicios también se caracterizan por tener la información almacenada en estructuras heterogéneas y por tener grandes volúmenes de información, lo cual supone un gran trabajo a la hora de buscar y navegar por la información científica.

La propuesta para este trabajo es tomar los documentos científicos en este caso artículos y construir un conjunto de datos que permita representar las temáticas mas relevantes dentro de un área del conocimiento, así a través de mapas y representaciones reticulares delimitar la búsqueda y la navegación por las categorías mas representativas.

El artículo muestra el proceso de automatización de la construcción de mapas de la ciencia y entregar los resultados previos obtenidos, en esta vía se preocupa por el modelamiento de documentos en tanto sus palabras clave y la exploración de técnicas que permitan construir automáticamente mapas de áreas del conocimiento, por lo cual se han desarrollado algunas herramientas informáticas en php y se han utilizado algunas implementaciones en R.

Dichas herramientas contemplan descargas de los documentos científicos del área definida, la construcción del CORPUS textual, el pre-procesamiento, la extracción de características de documentos científicos, el desarrollo de un modelo para la caracterización de redes de palabras clave y la construcción de representaciones que permitan la fácil navegación y búsqueda.

Los métodos de clasificación, agrupamiento y visualización de los documentos a través de las palabras clave son: reducción de dimensionalidad, de escalamiento multidimensional, SOOM, partitioning methods.

### 2.1 Proceso de automatización de la construcción de mapas de la ciencia

El presente trabajo el cual es exploratorio y busca desarrollar de una parte un modelos que permitan caracterizar los documentos científicos en tanto sus palabras clave y de otra parte identificar las técnicas computacionales que permitan desarrollar mapas científicos, presenta los siguientes pasos organizado para construir un mapa de la ciencia.

Como primer desarrollaron permiten documentos forma mediante una búsqueda a los publicaciones consulta de datos en extraiga las los autores y datos del como año, abstract.

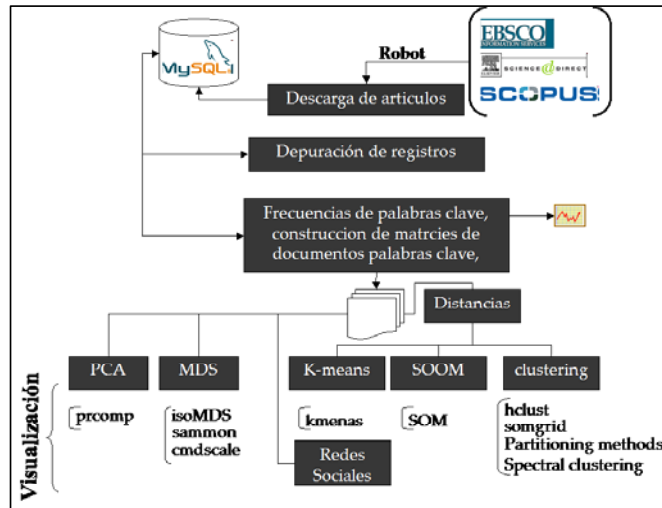


Fig. 1. Metodología de trabajo

ítem se aplicaciones que descargar los científicos de automática función de servicios de y a partir de esta generar una base MySQL que palabras clave, algunos meta documento, tales journal, título y

Como segundo ítem se desarrollaron aplicaciones que permiten depurar los registros de la base de datos, depuración de forma y de repeticiones.

Con este conjunto de datos se desarrollaron análisis de frecuencias y dinámicas del tópico relevantes a las palabras clave. Así como también se construye una matriz de documentos contra palabras clave y se calcularon distancias de manhhathan y jaccard para aplicar clustering y MDS.

Para realizar los mapas se utilizaron técnicas de reducción de dimensionalidad, de escalamiento multidimensional, SOOM, partitioning methods y de análisis de redes.

### 2.3 Construcción de matrices.

La matriz de relación de documentos palabras clave se construye teniendo en cuenta las frecuencias de las categorías mas representativas, de tal manera que tenemos un

vector de características por documento el cual tiene 1 o 0 de si la palabra clave se encuentra o no en el documento.

$X=\{x_i\}$  Conjunto de  $n$  palabras clave  
 $Y=\{y_j\}$  Conjunto de  $m$  documentos  
 $E=\{w_{ij}\}$  Conjunto de ejes que indican 1 si la palabra  
esta en el documenton y 0 en el otro caso

## 2.4. Cálculo de distancias.

La idea principal del calculo de distancias es encontrar que tan cercanos o distantes están dos puntos, para este estudio se busca obtener la similitud de cada una de las palabras clave de los documentos representados en la matriz.

Existen varias funciones de similitud para calcular las distancias en datos binarios [(presencia (1) y ausencia (0)] coseno, de Dice, de Jaccard, etc. Los coeficientes de distancias que utilizaremos en este trabajo son el de manhattan y el de jaccard.

El coeficiente de Jaccard (porcentaje de presencia-ausencia) puede variar entre 0 y 1, donde 0 indica ausencia de palabras clave en común y 1 que ambos documentos son idénticos.

$$\text{Distancia Jaccard} = \frac{2a}{2a+b+c}$$

donde (a) representa dos presencias (1:1), (b) representa presencia-ausencia (1:0) y (c) ausencia-presencia (0:1). (Digby y Kempton, 1987).

La distancia manhattan o de calles urbanas entre dos casos es la suma de los valores absolutos de la diferencia entre observaciones para cada variable.

## *Reducción de dimensionalidad*

### 2.5. Reducción de dimensionalidad

Se propone para este articulo utilizar técnicas que producen mapas en dos y tres dimensiones de datos multivariados. este articulo da un primer acercamiento a las técnicas de reducción de dimensión dado el conjunto de datos.

#### 2.5.1 Análisis de componentes principales

El análisis de componentes principales-PCA- se enmarca en el conjunto de técnicas multivariadas conocidas como métodos correlacionales, se busca la proyección de los datos dentro de nuevo conjunto de ejes, que contengan la máxima varianza en el primer eje, y la máxima varianza no correlacionada en el segundo, 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.

El objetivo del análisis de componentes principales es reducir la dimensión de un conjunto de  $p$  variables a un conjunto  $m$  de menor número de variables para mejorar la interpretación de los datos.

#### 2.5.1 Escalamiento multidimensional

Multi-Dimensional Scaling-MDS- es un método basado en la información de las distancias de un conjunto de datos multivariados y busca reducir la dimensión  $L$  encontrando un conjunto de vectores que pertenezcan a los reales y que reproduzca las distancias del conjunto inicial. El acercamiento clásico al MDS es el PCO el cual utiliza las dos primeras componentes de la descomposición de la matriz de distancias, si la medida de distancias utilizada es euclidiana la técnica se convierte en PCA.

MDS comienza con un conjunto de ejes tomados de PCO y busca minimizar el “stress” esta medida es la 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 randominca creada con igual distribución en un espacio de dos dimensiones.

Las técnicas utilizadas para este trabajo son: isomds, sammon, cmdscale.

#### 2.6. Clustering

Estas técnicas permiten agrupar las observaciones de forma que los datos sean muy homogéneos dentro de los grupos (mínima varianza) y que estos grupos sean lo más heterogéneos posible entre ellos (máxima varianza). Las técnicas utilizadas son SOM, Hculst y parm.

##### 2.6.1 Técnicas jerárquicas

##### 2.6.2 Técnicas particiónales

##### 2.6.3 Espectral clustering

#### 2.7. Clasificación.

#### 2.8 Mapas auto-organizativos.

### **Capítulo 3, Herramienta informática de vigilancia tecnológica para análisis socio-cognitivos de comunidades científicas.**

#### 3.1 Antecedentes y justificación

La investigación científica se desarrolla dentro de lo que se ha denominado como la sociedad del conocimiento[1], en ella la producción científica (artículos en revistas indexadas y títulos de propiedad industrial) tiene un lugar destacado en el aparato productivo de un país, así el conocimiento generado y los productos que del mismo se desprenden constituyen herramientas básicas para generar mayor bienestar y desarrollos sociales[2].

La sociedad del conocimiento y la producción científica necesitan de herramientas informáticas que permitan hacer por una parte seguimiento y control y de otra parte, métricas y mapas que permitan dar cuenta de el estado y las dinámicas del conocimiento producido y encontrar la traducción del conocimiento en objetos puestos en el mercado.

En el marco de esta sociedad del conocimiento, aparecen las prácticas de vigilancia, inteligencia competitiva y de gestión del conocimiento para su explotación económica y social. La vigilancia tecnológica de la cual se tratará en este trabajo, se ocupa del monitoreo de las tecnologías disponibles o que acaban de aparecer capaces de intervenir en nuevos productos o procesos. Ésta consiste en la observación y el análisis del entorno científico, tecnológico y de los impactos económicos presentes y futuros, para identificar las amenazas y las oportunidades de desarrollo [3].

La vigilancia tecnológica requiere del esfuerzo sistemático y organizado de observación, captación, análisis, difusión precisa y recuperación de información sobre los hechos del entorno económico, tecnológico, social o comercial, relevantes para tomar decisiones con menor riesgo y poder anticiparse a los cambios [3].

En esta vía se han realizado los siguientes desarrollos entre los más importantes están la aplicación *Tetralogie* (la interfaz de usuario esta en francés), *matheo analyzer*, *Leximappe* y *Dataview*.

*Tetralogie* forma parte de la estación cuantitativa ATLAS, Tiene capacidad para realizar análisis independientemente del formato de la información original, sin necesidad de operaciones previas, el conjunto de los métodos estadísticos de los que dispone son componentes principales, análisis factorial de correspondencias, clasificación jerárquica ascendente, rotación procustiana. Permite obtener indicadores bibliográficos unidimensionales o cálculos de coocurrencia doble y triple con indicadores relacionales, También permite elaborar filtros y *thesaurus* multitérminos para depurar la información y delimitar los análisis.

En este trabajo se entiende la sociedad del conocimiento, como un sistema que está en constante construcción, evolución y transformación, recurrentemente esta obteniendo nuevos productos científicos, y sus actores modifican constantemente sus capacidades[4]. Éstas características se ven reflejadas en bases de datos de artículos científicos y de patentes las cuales están distribuidas en diferentes servicios donde la información es no homogénea, dado el alto volumen de información; es un problema interesante para la ingeniería, debido a que existe la necesidad de desarrollar herramientas informáticas que permitan a diferentes actores y usuarios contar con información estructurada y con herramientas computacionales que le permitan tener métricas y mapas de las dinámicas científico tecnológicas.

Ahora bien, teniendo en cuenta el modelo de la triple hélice (modelo utilizado por la política científica de los países latinoamericanos donde se considera que los sistemas nacionales de ciencia y tecnología están compuestos por tres tipos de actores: Gobierno, academia y empresa)[4] los usuarios y actores que se tienen en cuenta son: los tomadores de decisiones quienes necesitan métricas de la producción, los investigadores quienes necesitan mapas conceptuales y sociales de las temáticas que estudian y los empresarios quienes necesitan mapas tecnológicos del área.

De esta forma, existe la necesidad de construir una herramienta que permita hacer una representación del conocimiento científico a través de mapas que contengan la dimensión social y cognitiva y de mapas tecnológicos de una área de conocimiento, así como la necesidad de desarrollar análisis reticulares y métricas que den cuenta de cómo se construye dicho conocimiento. Se propone esta temática como un primer elemento en la construcción de un sistema que permita hacer vigilancia tecnológica.

Teniendo en cuenta que el desarrollo del conocimiento no dado solo en el discurso, sino también como una construcción social se propone desarrollar una herramienta informática o Frame Work-FW- de análisis de comunidades científicas desde un enfoque socio-cognitivo, este enfoque se basa en el modelo propuesto por Leydersdorf [5], acerca de las unidades de análisis de una comunidad científica.

De esta forma el FW debe permitir encontrar redes donde se vinculen la componente social y cognitiva, así como clasificar mediante las redes producidas documentos, por temáticas, palabras clave, autores, citas, referencias bibliográficas y diferentes tipos de relaciones inscritas en los documentos científicos artículos y patentes.

Así desde un acercamiento de Machine Learning-ML- se propone desarrollar una herramienta informática que permita, a partir de un conjunto delimitado de artículos científicos del área, construir representaciones reticulares. La herramienta permitirá, construir automáticamente diferentes tipos de redes según sus tipos de relaciones, así como visualizarlas, navegar y construir métricas de dichas redes.

Las redes que se proponen construir desde los documentos científicos artículos y patentes son: redes de artículos, temáticas, preposicionales, citacionales, autoría y co-autoría, de cooperación institucional y referencias bibliográficas para el caso de los artículos científicos y para el caso de patentes redes de autoría y co-autoría, cooperación internacional, clasificación internacional, y sector de aplicación.

Así los enfoques para el análisis de las comunidades científicas a través de herramientas computacionales, es el de redes semánticas[6][7], redes sociales[8][9] como dos marcos de trabajo tanto conceptuales como metodológicos y la vinculación de estos dos marcos permitirá observar los procesos de producción científica, la herramienta permitirá procesar mediante técnicas de extracción de conocimiento las publicaciones científicas, para así construir métricas sobre la producción científica y representar a través de mapas y sociogramas las estructuras relacionales.

Clasificación	DESCRIPCIÓN	Métodos de análisis
Representación de las estructuras sociales	Permite identificar y medir las estructuras sociales presentes en una comunidad. [8][9][10][11]	Análisis de redes sociales
Análisis estadísticos descriptivos	Permite construir indicadores sobre el estado del la CyT+i tales como Número de investigadores, Número de grupos de investigación o indicadores bibliográficos sobre el estado de la publicación científica.[12]	Estadísticas sobre los datos bibliográficos, co-ocurrencia de palabras, estudio de citas
Representación de las estructuras cognitivas	Permite la construcción de herramientas computacionales que buscan modelar los procesos de comunicación y los procesos cognitivos a través de lenguajes naturales. Los modelos aplicados para el caso de la evaluación de una comunidad científica se enfocan a la comprensión del lenguaje utilizado y a aspectos generales cognitivos.[13][14][15][16][17][18]	Extracción de términos, palabras clave, semántica latente construcción y utilización de tesauros y ontologías
Clasificación automática de documentos	Permite obtener agrupamientos de documentos, esta clasificación automática para documentos científicos se realiza a través de diferentes variables tales como autores, palabras clave, journal donde fue publicado, fechas, palabras encontradas en el paper, etc.[19][20][21][22]	Método de palabras asociadas, clustering de documentos.
Visualización	Permite visualizar las representaciones reticulares construidas[23]	Visualización de grafos

TABLA I, CLASIFICACIÓN DE MÉTODOS PARA ANÁLISIS DE COMUNIDADES CIENTÍFICAS

### Identificación del problema

La información de la producción científico tecnológica (artículos en revistas indexadas y patentes) se encuentra alojada en bases de datos no homogéneas las cuales contienen volúmenes altos de información, por lo que, la búsqueda y navegación en estos documentos supone una cierta especialización en el tópico tratado y varias horas de trabajo.

Pero es posible construir una representación del conocimiento científico a través de mapas tecnológicos de una área de conocimiento o mapas que contengan la dimensión social, cognitiva del tópico tratado. De esta forma desde un enfoque socio-cognitivo, se propone extraer las relaciones a partir de documentos científicos y encontrar diferentes representaciones reticulares tales como redes temáticas, redes de palabras clave, redes

de autores, redes de citas, redes de referencias bibliográficas y redes de documentos científicos, teniendo como resultado no solo la representación del conocimiento, también la posibilidad de desarrollar análisis reticulares y métricas que den cuenta de cómo se construye dicho conocimiento.

### 3.3 Objetivo general y objetivos específicos

Desarrollar un sistema de extracción de relaciones socio-cognitivas a partir de documentos científicos.

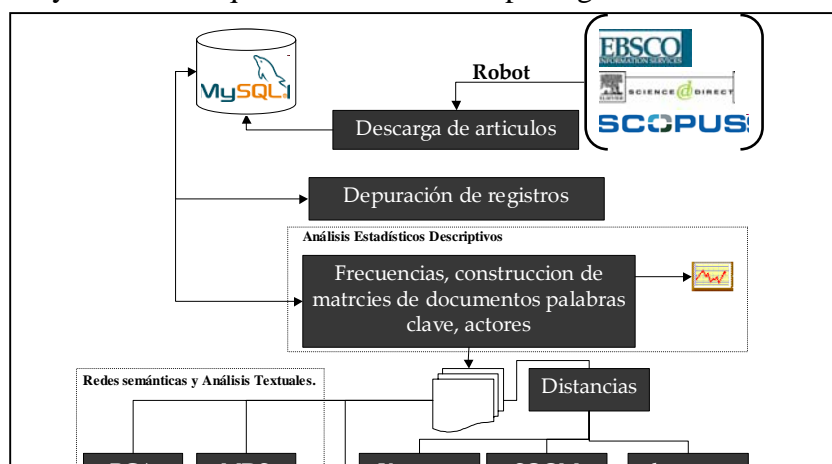
#### OBJETIVOS ESPECIFICOS

- Desarrollar un sistema para la obtención, preprocesamiento y extracción de características de documentos científicos y patentes.
- Construir un modelo para la construcción de redes sociales de las características obtenidas de los documentos
- Construir un modelo para la construcción de redes cognitivas de las características obtenidas de los documentos
- Desarrollo de un sistema para análisis descriptivos y dinámicos de las redes construidas
- Desarrollar una herramienta prototipo que integre los modelos generados con una interfaz de usuario, que permita la visualización, navegación y la construcción de métricas de las redes obtenidas.
- Evaluar la herramienta mediante dos casos de estudio.

### 3.4 Metodología

Dado el problema del manejo y representación de la información de la producción científico tecnológica y basado en los acercamientos de la cienciometría, bibliometría[12], el análisis de redes sociales[9] y redes semánticas [14] la metodología propuesta es utilizar técnicas del aprendizaje de maquina-ML- para la extracción de relaciones socio-cognitivas.

Como se muestra en el gráfico 2, se propone como parte de la metodología desarrollar algoritmos que permitan obtener documentos científicos y patentes, dichas búsquedas se realizaran en bases de datos de ISI, base de datos de patentes SIC. Con dichos CORPUS textuales preprocesar la información y extraer características, así como construir los datos y meta datos que serán útiles en etapas siguientes.



## Grafico 2, metodología de trabajo

### Análisis de Redes sociales-ARS-

El análisis de redes sociales, es una metodología de análisis cuantitativo y estructuralista que busca reconocer las relaciones y sus estructuras para poder encontrar en este sistema de relaciones y de actores, comportamientos y en sí la estructura -o estructuras- social de dicha comunidad analizada; 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.

### 3.5 Documentación de la herramienta informática

3.5.1 Documentación del desarrollo de algoritmos para la construcción del CORPUS textual.

3.5.2 Documentación de las descargas de los documentos científicos y patentes del área definida.

3.5.3 Documentación del desarrollo de algoritmos para el preprocesamiento y extracción de características de documentos científicos y patentes.

3.5.4. Documentación del desarrollo de un modelo para la construcción de redes sociales a partir de las características documentales.

3.5.4.1 Redes de autoría y co-autoría

3.5.4.2 Redes de cooperación institucional

3.5.4.3 Redes de referencias

3.5.4.4 Redes de referenciados

3.5.5. Documentación del desarrollo de un modelo para la construcción de redes cognitivas a partir de las características obtenidas.

3.5.5.1. Redes palabras clave

3.5.5.2 Redes de temáticas

3.5.5.3 Redes de clasificación internacional(patentes)

3.5.5.4 Redes de sector de aplicación(patentes)

3.5.6. Documentación del desarrollo de una herramienta de análisis descriptivos y dinámicos de las redes construidas.

3.5.7. Documentación del desarrollo de una herramienta de visualización y navegación de las redes construidas y de las métricas y de los meta datos.

3.6 Análisis de servicios de búsqueda de publicaciones científicas, disponibles en la Universidad Nacional de Colombia.

3.6.1 Criterios fuentes de datos

*Criterios fuentes de datos:*

*Cobertura ISI 16000 revistas(2001), Scopus 14.000 revistas(2004).Índice de Bradford  
Número principal de literatura en cualquier disciplina esta compuesta por menos de  
1000 revistas*

*Cobertura Temática. Ciencias Humanas, Biotecnología, etc.*

*Temas, países, idiomas y editores...Procesos de indización, meta datos, inclusión.*

*Calidad editorial*

*Proceso de evaluación(periodicidad), incluir – excluir.*

*Monitoreo*

*Radio de selección ISI 10 al 12 % de 2000 títulos evaluados por año*

*Patrones básicos de evaluación*

- *Periodicidad*
- *Meta datos completos*
- *Revisión por pares de las revistas*
- *Contenido*
- *Internacionalidad*
- *Análisis de Citas*

**Descriptor.** Recoge todos los conceptos del tesauro

**Identificadores.** Recoge los nombres propios de personas, instituciones con sede, títulos de obras etc.

**Otras fuentes...**

3.6.2 Criterios servicios de búsqueda y de recuperación de información

3.6.2.1 Análisis documental

El análisis documental es un trabajo mediante el cual por un proceso intelectual extraemos unas nociones del documento para representarlo y facilitar el acceso a los originales. Analizar, por tanto, es derivar de un documento el conjunto de palabras y símbolos que le sirvan de representación.

- Tipo de usuarios y necesidades de información más o menos especializadas
- Documentos a analizar: Libros, artículos de revistas, literatura gris, prensa, legislación.
- Finalidad técnica del análisis: catálogación o recuperación especializada.

3.6.2.2 Los lenguajes controlados y la indización

**LOS LENGUAJES CONTROLADOS Y LA INDIZACIÓN** El concepto de indización se identifica con el análisis del contenido en la medida que dichos lenguajes se utilizan para elaborar los índices temáticos por los que se recupera la información. Indizar es extraer una serie de conceptos que responden a los temas tratados en el documento, y que servirán como puntos de acceso para su recuperación.

**Las clasificaciones universales**, (CDU, LC, UNESCO) Su estructura jerárquica impide la combinación de los múltiples aspectos de una investigación, y no permite recoger temas muy específicos o novedosos.

**Las clasificaciones especializadas**, elaboradas para una disciplina o un sistema documental específico, complementarias a los descriptores. Permiten englobar en un marco amplio todos aquellos documentos de una categoría.

Tesauros: normas UNE 50-106-90 y 50-125-97.

**Normalización de singulares y plurales.** *Normas para el desarrollo de tesauros monolingües.*

Coherencia

Especificidad

Exhaustividad

**Pertinencia.** Metodologías utilizadas ejemplo norma UNE 50-121-91.

Recuperación de la información

Metodologías utilizadas ejemplo norma UNE 50-137-99.

Motor de búsqueda y recuperación de la información.

*Tabla con criterios presentados por servicio de indexación y resumen, basados en indexación, tesauros, resumen.*

*Relación palabras clave, título, abstract.*

*Repeticiones de búsquedas, precisión, error, curva ROC.*

*Búsquedas conocidas a priori, precisión, error, curva ROC.*

### 3.6.2.3 Modelos cognitivos de IR

*SMART (System for Manipulation and Retrieval of Text)* El modelo cognitivo, a diferencia del tradicional, no se pone en marcha a partir del momento en que el usuario realiza una consulta al SRI, sino que comienza incluso antes de que se produzca la necesidad informativa del usuario.

*Ingwersen label effect,*

Modelo probabilístico de RI de Robertson que sugiere que cuantas más pruebas o evidencias tengamos sobre la consulta, sobre los documentos y sobre las relaciones entre ellos, mayores serán las probabilidades de que los resultados se adecúen a la necesidad informativa del usuario

### 3.7 Evaluación por expertos.

## 4. Conclusiones.

Otro punto de partida que necesidades documentales y de recuperación de información exige la Vigilancia tecnológica