Organización y representación del conocimiento: fundamentos teóricos y metodológicos para la recuperación de la información en entornos virtuales

Marcos Luiz Cavalcanti de Miranda

Director y Profesor de la Escuela de Bibliotecología de la Universidad Federal del Estado de Río de Janeiro, Brasil. mlmiranda@unirio.br

Resumen

El presente trabajo tiene por objetivo la recuperación de la información en entornos virtuales, analizando el rendimiento de mecanismos de búsqueda que utilizan el lenguaje natural en el contexto de la organización del conocimiento. Nuestro punto de partida fueron las teorías y metodologías de la organización del conocimiento para la recuperación de la información, encontradas en la Bibliotecología y en áreas de Ciencias de la Información y Documentación. Diferentes formas y niveles de recuperación de la información en la Web, fueron caracterizados por medio del análisis de los mecanismos de búsqueda disponibles. Estas herramientas de búsqueda también fueron analizadas con el objetivo de identificar su potencial rendimiento en la recuperación de información y de verificar el índice de recuperación en la Web mediante la utilización del lenguaje natural. Los análisis realizados en los mecanismos de búsqueda (Altavista, Google y Yahoo) consideraron como variables documentos, términos, ocurrencias, subject scope y ranking. Los resultados obtenidos, interpretados con base en métodos teóricos y metodológicos sobre la organización del conocimiento para la recuperación de la Información en entornos actuales, revelaron que éste es un soporte valioso para ayudar a la organización del conocimiento en el entorno virtual. Los resultados también nos demuestran que los índices de recuperación de la información pueden ser mejorados si la organización del conocimiento para la recuperación de la Información en la web considera el procesamiento de informaciones, principalmente basado en relaciones conceptuales, alcanzando las estructuras cognitivas humanas.

Keywords: Motores de Búsqueda, Organización del Conocimiento, Recuperación de la Información, Web.

Abstract

The present work approaches information retrieval in virtual environments analysing search engines performance using natural language in the context of knowledge organization. The research has as starting point theories and methodologies of the knowledge organization for

information retrieval found in Library and Information Sciences and Documentation areas. Different forms and levels of information retrieval in the Web were characterized by analyzing search engines available. The search engines were also analyzed in order to identify their potential performance in information retrieval, and in order to verify the information retrieval rate in the Web using natural language. The analyses performed on the given search engines (Altavista, Google and Yahoo) considered as variables documents, terms, occurrences, subject scope and ranking. The results obtained, interpreted on the basis of theoretical and methodological approaches to knowledge organization for information retrieval in current environments, revealed that this is a valuable support to subsidize knowledge organization in virtual environment. The results also indicate that information retrieval rate may be improved if knowledge organization for Information retrieval in the web considers information processing, mainly based on conceptual relations, approaching human cognitive structures.

Keywords: Information Retrieval, Knowledge organization, Search Engines, Web.

1. Introducción

El fenómeno de la información tiene lugar en diversas áreas del conocimiento y en los ámbitos más variados. En esta sección tratamos de este fenómeno bajo la perspectiva de la Ciencia de la Información. De esta manera, el ámbito con el cual se está trabajando es el SRI - Sistema de Recuperación de la Información-, independientemente de los soportes de la información y de las instituciones de información y memoria. Nos basamos en Levy (1997) para presentar el entorno actual y el entorno virtual.

La información puede estar registrada en diferentes soportes al mismo tiempo, ya sea en el documento impreso, documento de imágenes, documento electrónico y/o documento digital. Estos documentos pueden estar disponibles en diversos entornos de la información: en el entorno actual, en el entorno electrónico, en el entorno digital y/o en el entorno virtual.
La World Wide Web, o simplemente la Web es un camino de acceso a la información en Internet. Es un modelo de dominio público de información, que fue construido para la búsqueda de informaciones, lo que puede llevarse a cabo mediante directorios, motores de búsqueda y metabuscadores.

El procesamiento de la información es una actividad compleja dado que la información, dependiendo del contexto y del dominio del conocimiento puede tener diversos significados – un mismo término en determinada situación puede representar un concepto y en otra otro, constituyendo relaciones conceptuales distintas.

En la literatura científica del ámbito de la Bibliotecología y la Ciencia de la Información encontramos, con frecuencia, los términos procesamiento de la información y tratamiento de la información como sinónimos. Sin embargo, para esta investigación, consideramos como procesamiento de la información el proceso que ocurre en la estructura cognitiva de un individuo, cuando tiene lugar la absorción y el almacenamiento de la información en el plano mental. Como tratamiento de la información consideramos todos aquellos procesos realizados por profesionales de la información en el plano físico, desde el momento en que la información que está registrada en determinados soportes se selecciona para incorporarla a la

base de datos de cualquier unidad de información, para que dichos soportes estén disponibles para utilización futura.

Como tratamiento de la información se entiende todas las actividades realizadas en el subsistema de entrada de la información en un SRI. Todavía, para fines de esta investigación, consideraremos el tratamiento temático de la información/documento – aquí no se separa el documento de la información y viceversa – , pues estamos de acuerdo con Briet cuando afirma que "no existe información sin documento, ni documento sin información"(Briet, 1951).

El tratamiento temático de la información o la indexación es una operación que consiste en identificar sobre lo que trata el contenido de un documento para poder obtener una síntesis -mediante el análisis y la representación de conceptos-, palabras-claves, términos, descriptores relevantes por medio de un lenguaje de indexación –aquí denominado sistema de organización del conocimiento–, objetivando la localización y la recuperación de documentos/información en cualquier ámbito.

Según Lancaster (2003) la indexación posee básicamente las siguientes etapas: análisis de los conceptos y traducción. El análisis conceptual consiste en identificar los conceptos tratados en el documento y seleccionar aquéllos que serán traducidos como términos de indexación en el contexto del SRI a que pertenece. La traducción se ocupa de la conversión de los conceptos identificados y seleccionados en términos de indexación en consonancia con un lenguaje normalizado adoptado por el SRI.

La eficiencia de un SRI depende mucho de la calidad del análisis conceptual tanto de los documentos como de las cuestiones de los usuarios. Muchas de las fallas en la recuperación de la información en los entornos actuales se deben a errores u omisiones en la interpretación del contenido de los documentos y en la percepción de la demanda de las personas a las cuales se destina el sistema, lo que también podrá ocurrir en el entorno virtual.

Actualmente, se han venido realizando diversos estudios sobre las formas de tratamiento, procesamiento, organización, almacenamiento y recuperación electrónica de la información. En este sentido, la atención de varios investigadores de las más diversas áreas del conocimiento se ha dirigido a Internet –una red de comunicación electrónica, que permite que sus usuarios tengan acceso a los asuntos más variados en diversas áreas del conocimiento humano, en cualquier parte del mundo. De hecho, la organización de toda esa información teniendo como objetivo una recuperación más precisa, es el gran desafío del profesional del área de información en el tercer milenio.

Si por un lado, las tecnologías y el desarrollo de las redes de información posibilitaron la difusión del conocimiento, por otro lado, estimularon la publicación directa de la "fuente al consumidor". Esto genera una falta de estándares para poder poner a disposición del usuario documentos/información en Internet y, como consecuencia, la búsqueda y la recuperación de la información en estos entornos virtuales es mucho más difícil.

La necesidad de información de los usuarios, naturalmente, varía de un dominio a otro y de un grupo a otro, de acuerdo con la fase de desarrollo del área de conocimiento, la naturaleza de los usuarios, sus objetivos. A pesar de esas variaciones es necesario que las informaciones

sean fiables, actuales, indexadas de manera adecuada para la recuperación y el acceso inmediato.

La recuperación de la información está relacionada con las formas de almacenamiento, y éstas a su vez con el tratamiento y la organización de la información. La información organizada se trataba inicialmente de forma manual, después se le realizó un tratamiento mecánico, a continuación electrónico y, en la actualidad, se somete a un tratamiento digital. Sabemos que todas esas formas de tratamiento y organización de la información coexisten.

En la recuperación de la información en entornos virtuales, en algunos momentos nos sentimos perdidos, sin la noción del área de conocimiento en la cual podríamos encontrar la información que deseamos. Todo esto como consecuencia natural de la característica intrínseca de la red mundial de informaciones, Internet que ha sido construida, desconstruida y reconstruida según una lógica bastante distante de los principios de la organización del conocimiento en el contexto de la Ciencia de la Información. Consideramos que este problema puede ser investigado a través de la utilización de los recursos del conocimiento teórico y metodológico que permitan la construcción de instrumentos para la organización del conocimiento con el objetivo de recuperar la información en entornos virtuales.

Como punto de partida, tomamos como base teorías y metodologías de la Organización del Conocimiento para recuperación de la Información, encontradas en el área de la Ciencia de la Información, aliadas con los elementos del área de Bibliotecología, Documentación, Comunicación y Ciencia de la Computación. Esto es debido a que la Organización del Conocimiento se constituye en disciplina científica, ínter y transdisciplinar, cuyo objetivo es generar y difundir, en un nivel de excelencia, la información en el ámbito de los archivos, bibliotecas, centros de información/documentación y museos, sea en entornos actuales o virtuales.

Algunas cuestiones se encuentran en esta línea de razonamiento: ¿cómo representar el conocimiento que no está organizado en estructuras organizadas, considerándose que el conocimiento, en el ambiente virtual, no es organizado, y se recupera solamente de manera jerárquica, pero también rizomática? ¿Podríamos, dentro de esas estructuras jerárquicas establecer *links* para la búsqueda de la información en áreas de conocimiento que son por excelencia ínter, trans o multidisciplinares? ¿Y los metadados? ¿Hasta qué punto serían útiles para conseguir resolver estas cuestiones? ¿Cómo lidiar con las barreras establecidas por el carácter virtual de la información ora disponible, ora “perdida”en los sitios? En la realidad, ¿Cuáles son las condiciones de las probabilidades de recuperación de información en la Web?

Como objetivo general de esta investigación, nos propusimos demostrar el potencial de los fundamentos teóricos y metodológicos para la organización del conocimiento con el objetivo de recuperar la información en entornos virtuales. Nuestros objetivos específicos son: analizar la recuperación de la información en los entornos virtuales a través del lenguaje natural y estudiar los mecanismos de búsqueda de información existente y utilizados para recuperar la información. En ello reside la esencia de nuestra investigación.

2. Topografia del Conocimiento: las relaciones conceptuales en la representación del conocimiento en la Web

Ya se sabe que la clasificación y la indexación automática, así como los metatados, han sido utilizados para la representación del conocimiento en entornos virtuales. A pesar de todo, debe concederse mayor atención al tratamiento de la información. Por esto, en esta sección destacamos la importancia de estas relaciones para la representación del conocimiento en la Web.

Diversos estudios se han realizado con el objetivo de discutir la teoría y ampliar los estándares que faciliten y le aseguren al usuario la recuperación y la visualización de la información y, que permita que información indebida no sea recuperada o no sea solicitada (Miranda, 1999, p. 68).

Las dificultades que están asociadas a las formas de representación en entornos actuales y virtuales constituyen desafíos para la racionalidad humana. El Hombre piensa en representación si a ésta le puede dar una expresión de racionalidad, a medida en que asocia principios, categorías, procedimientos y normas, en la medida exacta en la cual construye modelos y algoritmos con la finalidad de que todos esos aspectos, puedan tornar estable la tarea de representar la realidad de forma comprensible (Miranda, 1999, p. 68).

Vickery opina que muchas técnicas diferentes de la representación del conocimiento, fueron desarrolladas en los distintos campos, y que con frecuencia se producen variaciones debido a los diferentes tipos de manipulación. Pero existen algunas convergencias interesantes y parece que cada campo tiene algo que aprender de los otros. Vickery revisó diferentes formas de representación del conocimiento, buscando algunas que fuesen aplicables al área de la Ciencia de la Información (Vickery, 1986, p. 75) y que deben ser consideradas en los entornos virtuales.

La necesidad de organizar el conocimiento siempre fue reconocida por el hombre. Mientras que en la antigüedad era una tarea considerada necesaria exclusivamente por los Filósofos y Bibliotecarios, en el tercer cuarto del siglo XX, pasó a ser preocupación también de los Científicos de la Información. Actualmente, “las personas de las áreas de Inteligencia artificial, Sistemas expertos, Hipertextos, Educación, Psicología, entre otras, no dejan de demostrar interés en la aplicación de metodologías para la representación del conocimiento, buscando elementos que sustenten su organización” (Dahlberg, 1993, p. 211).

Los sistemas de organización del conocimiento bibliográfico y los alfabéticos, utilizados por bibliotecarios como instrumentos de representación y organización del conocimiento, con el paso de los años, fueron ignorados por los especialistas en las nuevas tecnologías, pues éstos creen que el conocimiento, por ser tan complejo, no podría ser capturado por ninguna teoría o sistema ordenado con principios tan generales (Dahlberg, 1993, p. 211).

Aun así, Dahlberg (1993, p. 211) afirma que “fundamentos teóricos desarrollados en el área de la Clasificación y Tesauros durante décadas, pueden ser utilizadas en todos los tipos [de sistemas] de organización del conocimiento, generales o específicos.”

Para Dahlberg, el punto esencial en la teoría de la organización del conocimiento, se encuentra en el hecho de que cualquier organización del conocimiento debe tener bases en

unidades de conocimiento, que no son otra cosa sino conceptos. Y éstos se componen de elementos, también denominados características de conceptos. Dahlberg afirma que son exactamente estos factores los que permiten que un sistema de conceptos sea construido, de la misma forma como un sistema de organización del conocimiento. La autora considera que el conocimiento jamás podrá ser representado sin las unidades de conocimiento y sus posibles combinaciones en palabras, símbolos y términos (Dahlberg, 1993, p.211).
Este argumento es reforzado por Campos, que a partir de las bases de la Teoría de Clasificación Facetada, de la Teoría del Tesauro-con-Base-en-Conceptos y la Teoría General de la Terminología, encontró principios en común en el trabajo de los autores que dieron origen a las teorías. Estos principios establecen: el concepto y su denominación, las relaciones entre los conceptos y la representación de estos conceptos en un sistema, considerando que, estas teorías, tienen como finalidad la elaboración de estructuras conceptuales para la representación/recuperación o comunicación de la información/conocimiento (Campos, 2001).

En la concepción de Vickery, *apud* Binwal, la representación del conocimiento es la llave para la recuperación y diseminación efectiva de los datos, información y conocimiento. Las técnicas utilizadas, tales como: categorías del predicativo y argumento, la gramática de casos de Fillmore, los primitivos semánticos en análisis lingüístico; la lógica de los predicativos, estructuras y redes semánticas en la inteligencia artificial; facetas, categorías fundamentales, indicadores de función, operadores, reenvíos “ver” y “ver también”, términos genéricos, términos específicos, términos relacionados en el campo de la recuperación de la información, son todas tentativas de representación del conocimiento (Binwal, 1992, p. 197).

El conocimiento registrado es generalmente representado en una declaración de asuntos que consiste en términos. Los términos sostienen conceptos y los combinan en una declaración de asuntos, de acuerdo con un estándar de relaciones. Esto significa que cada asunto tiene su propia estructura. Las unidades, o sea, conceptos o ideas aisladas identificadas en la descripción de un asunto en particular, son puntos en la estructura o redes de relaciones. La estructuración de los conceptos que forman un asunto desempeñan un papel vital en la comunicación, el aprendizaje y la memoria (Vickery, 1986, p.72).

De esta manera, mediante los conceptos y sus relaciones podrán obtenerse resultados satisfactorios en la búsqueda y la recuperación de la información en la Web.

3. Material y metodología

Nuestro campo de observación fue el entorno virtual y el área de cobertura la propia área de Organización del Conocimiento, por ser nuestra área de actuación en el campo de la docencia y la investigación, colocándonos, de esta forma, como usuario real de los mecanismos de búsqueda y recuperación de la información. En un primer momento seleccionamos algunos términos del área de Organización del Conocimiento y algunos de los servicios de búsqueda fueron escogidos según nuestra experiencia con la representación y recuperación de la información y con el www.searchenginewatch.com.

Seleccionamos términos en el idioma inglés que representan conceptos del área de la Organización del Conocimiento. En total fueron quince los términos adoptados para las búsquedas en lenguaje natural: automatic classification, automatic classifying, automatic indexing, knowledege organization, organization of knowledge, knowledege organization

systems, classification systems, thesaurus, subject headings list, Classification Research Group, ontology, classification, classifying, indexing e information retrieval.

Cada concepto fue definido conforme el Diccionario de Bibliotecología, Documentación y Ciencia de la Información de UNESCO, reunido por Wersig y Neveling (1976), registrado en fichas terminológicas propias, teniendo en consideración puntos tales como: asunto, categoría, término, definición o explicación, nota explicativa, término(s) no preferente(s), término(s) genérico(s), término(s) específico(s), término(s) partitivo(s), término(s) relacionado(s), observaciones y fuentes, establecidos por Gomes (1996, p. 40-42) según la Teoría de la Terminología y posteriormente organizados de acuerdo con sus relaciones conceptuales.

Los términos seleccionados, fueron introducidos, uno de cada vez, en las cajas de búsqueda de Altavista[1], de Google[2] y de Yahoo[3]. Se escogieron estos motores, debido a que la relevancia de sus resultados en la búsqueda de informaciones académicas, científicas y técnicas, ha sido contrastada en estudios recientes[4].
De estos resultados se analizaron las primeras diez referencias que se obtuvieron a partir de cada término, en cada uno de los motores de búsqueda utilizados. Se realizó su identificación con una codificación que estaba constituida y a la cual se le atribuían los siguientes elementos:

a)los términos, numerados de 1 a 15

1 Automatic Classification
2 *Automatic Classifying*
3 Automatic Indexing
4 Classification
5 Classification Research Group
6 Classification Systems
7 Classifying
8 Indexing
9. *Information Retrieval*.
10 Knowledge Organization
11 Knowledge Organization Systems
12 Ontology
13 Organization of Knowledge
14 Subject Headings List
15 Thesaurus

b) el motor de búsqueda, identificado por su inicial:

A – Altavista
G – Google
Y – Yahoo

c) la posición en el ranking, numerada de 1 a 10.

Para entender mejor la codificación presentamos el siguiente ejemplo:

309

1A1
W4: Automatic classification of WAIS databases

... WAIS/World Wide Web Project Subproject: Automatic classification of WAIS databases Anders Ardö, ... ***Automatic*** *detection of new WAIS* ***databases Automatic classification*** *according to UDC, English medium.. www.ub2.lu.se/autoclass.html • Páginas relacionadas • Traduzir Mais páginas de www.ub2.lu.se*

En donde la anotación a la izquierda, en la parte superior, el primer dígito 1 se refiere al código del término –Automatic Classification-, el segundo dígito A representa la inicial del

[1] http://www.altavista.com
[2] http://www.google.com
[3] http://www.yahoo.com
[4] Considerados por The best Search Engines los más utilizados en 19/05/2003.

motor de búsqueda – Altavista- y, el tercer dígito 1, indica la posición de la referencia en el raking.

Las referencias fueron organizadas en orden alfabético, por autor o título, el formato (pdf, ppt etc.) del documento no se tuvo en consideración y las referencias fueron numeradas secuencialmente. Después les atribuimos otra codificación referente a su posición en la base creada por todas las referencias que encontramos a partir de la recuperación de la información basándonos en las 3 herramientas. Esta codificación fue colocada en la parte superior del lado derecho de la ficha, en el ejemplo que presentamos éste equivale al 309.

A continuación fueron elaboradas seis planillas para registrar los datos del número total de documentos recuperados por cada herramienta de búsqueda en relación a los términos; de los documentos pertinentes y relevantes con sus respectivos rankings; las ocurrencias de los términos en relación a los mecanismos de búsqueda en el documento y ocurrencias de los términos en los documentos a partir de las herramientas de búsqueda.

Posteriormente se llevó a cabo un análisis de las búsquedas realizadas en los sitios en lenguaje natural por medio de la aplicación de técnicas de pooling –adoptadas en el proceso de evaluación de la Text Retrieval Conference (TREC)-, que se trata de la evaluación de las respuestas obtenidas en las búsquedas realizadas, haciendo una comparación de los resultados relevantes y los no relevantes.

Al fundamentar esta investigación utilizamos algunos principios y métodos de las teorías ya consolidadas en el área de la Ciencia de la Información, a los cuales le añadimos la teoría del rizoma, como un posible valor añadido a considerar.

Con base en los principios de la teoría de la indexación, analizamos el contenido de los documentos que quedaron en las diez primeras posiciones del ranking de los resultados con el objetivo de poder identificar los conceptos relevantes, su pertinencia y cobertura. Para que nos fuese permitido analizar la efectiva, precisa y correcta recuperación de la información, utilizamos principios y medidas de evaluación de la indexación/recuperación de la información, tales como: a) temática –sobre lo que trata el documento-; b) precisión –la extensión con se consideran relevantes o pertinentes los puntos recuperados durante una búsqueda en una base de datos-; c) relevancia –se refiere a la relación entre la necesidad de información formulada por una persona y las fuentes potenciales de información-; d) cobertura –extensión abarcada por el asunto en una base de datos o en toda la base de datos-; e) pertinencia –relación que existe entre una fuente de información y la necesidad de información de una determinada persona en un determinado momento- (Lancaster, 1993).

Para que tratemos sobre este último punto, introducimos la teoría del rizoma. La imagen del rizoma de Deleuze y Guatarri (1995) es una respuesta a la metáfora de la raíz que se bifurca, lo que representa la lógica clásica y los procedimientos binarios y dicotómicos. El método tipo rizoma, es un campo de experiencias y de posibilidades, toda vez que no se limita solamente a un análisis por descomposiciones internas. Este método analiza el lenguaje a partir de una descentralización sobre otras dimensiones y registros.

La idea de introducirse el rizoma para el estudio de las relaciones conceptuales surgió por la asociación de éste con la idea de red, considerando que en la Web ocurren relaciones conceptuales diferentes de aquellas de los SRI tradicionales. La topología de las redes se

configura a medida que son trazadas tramas conceptuales en la estructura cognitiva del hombre, de la misma manera los conceptos que en ella se encuentran almacenados se asocian de alguna forma a aquellos que están por detrás de las informaciones buscadas en cualquier entorno de información, estos conceptos pueden o no estar incluidos en dominios de conocimiento genéricos y/o específicos. Nos queda claro que este fenómeno está relacionado con el rizoma de Deleuze y Gatarri y con el árbol baniano de Ranganathan.

Para analizar las relaciones entre los términos/conceptos utilizados en la búsqueda de información en esta investigación, adoptamos los principios de la teoría del concepto y los principios de la teoría de la terminología. Los principios de la Teoría del Concepto nos permiten identificar cualquier objeto en el universo empírico (referente), atribuyéndole un conjunto de características con el objetivo de construir preguntas verdaderas acerca de tal objeto que en el futuro recibirá un nombre, y así llegamos a la definición de un concepto. Este modelo de formación de conceptos también nos da la posibilidad de identificar características semejantes o diferentes entre los objetos, lo que permite establecer las relaciones conceptuales existentes. Por otro lado, los principios de la Teoría de la Terminología establecen los conceptos y términos, así como la relación entre los conceptos, los sistemas de conceptos y definiciones, de donde destacamos el principio de la univocidad (Campos, 2001, p. 72), que se refiere a la correspondencia única entre denominación y concepto y el principio de la monoreferencialidad, en donde un significante terminológico, aunque sea complejo, representa en el espíritu de un especialista del área un conjunto conceptual único (Rondeau apud Campos, 2001 p. 73).

4. Análisis de los datos e interpretación de los resultados

En esta sección presentamos el análisis de los datos y la interpretación de los resultados que se consiguieron de acuerdo con las siguientes variables: los documentos en cada resultado de búsqueda; las ocurrencias de los documentos en los resultados de cada buscador; los términos utilizados para realizar las búsquedas; el ranking de las referencias de los documentos presentados en los resultados de búsqueda obtenidos.

4.1 Documentos

A continuación analizamos el rendimiento de los motores de búsqueda teniendo en consideración la temática y la relevancia. Con las tres herramientas de búsqueda seleccionadas fue posible recuperar 67.171.603 documentos.

De acuerdo con la cantidad de ítems recuperados en el Web mediante motores de búsqueda, se nota la diferencia que existe entre los resultados obtenidos, lo que demuestra que las respuestas varían de asunto en asunto y de buscador en buscador. Considerando que utilizamos los 10 primeros resultados en los motores de búsqueda a partir de los 15 términos anteriormente seleccionados, trabajamos con 450 (10x3x15 = 450) ítems.

Basándonos en nuestros recursos de conocimiento en el área de Organización del Conocimiento, o sea, la lista de los documentos recuperados en relación a los términos de búsqueda, verificamos si los ítems recuperados eran pertinentes o no con respecto a las ideas almacenadas en nuestra estructura cognitiva.

A partir de aquí el análisis se realizó con los diez primeros ítems recuperados.

a) Temática: Por lo menos pareció que la recuperación de la información, mediante los términos 3, 7, 9 y 12, utilizando los tres mecanismos de búsqueda, fue total. Es importante destacar que de los 15 términos solamente 4 permitieron el 100% de recuperación de la información, y que la media de la suma más elevada fue la de Google. Los datos conseguidos nos permiten evaluar el desempeño de los motores de búsqueda por medio de la cantidad de ítems relevantes que de hecho fueron recuperados y la cantidad que debería haber sido recuperada.

b) Relevancia: Así como mencionamos en el inicio de esta sección, la relevancia de los documentos recuperados fue considerada mediante la técnica de pooling. El número de documentos relevantes en relación al total de documentos recuperados por los tres motores de búsqueda, nos proporcionó un universo de 17 a 27 documentos, por cada término. Consideramos que un documento es relevante para un usuario si le resuelve la necesidad de información o la situación que está representada por la pregunta planteada. Esta relación es subjetiva si consideramos que diferentes personas tomaran diferentes decisiones al respecto de cuales son los ítems relevantes para las determinadas preguntas o en qué medida son relevantes para aquella pregunta en aquél determinado momento. Una forma especial de relevancia – relevancia para la necesidad de una información– es denominada de pertinencia. Luego analizamos los ítems recuperados considerando la pertinencia, la precisión y la cobertura.

c) Pertinencia: Los datos trabajados en porcentajes, nos dan el índice de pertinencia[5] lo que nos permitió verificar que sólo con los términos “Automatic Indexing”, “Classifying”, “Information Retrieval” y “Ontologies” hubo 100% de pertinencia. Observamos que muchas veces el documento era pertinente, pero no era recuperado por el buscador por el mismo término. “A framework for understanding and classifying ontology” por ejemplo, fue recuperado por el término “Classifying” en Google y Yahoo!, pero en Altavista sólo fue recuperado por el término “Automatic Classifying”. Con relación al número de documentos pertinentes recuperados por los tres motores de búsqueda, teniendo como universo los diez primeros documentos recuperados por término, verificamos que de media, para los términos “Automatic Indexing”, “Classifying” y “Ontologies”, el índice presentó un excelente funcionamiento (100%). Por otro lado, para “Automatic Classification”, “Automatic Classifying”, “Classifying Systems” e “Information Retrieval” presentó un muy correcto rendimiento (96,9%), así como los términos “Classification”, “Indexing”, “Subject Headings Lists” y “Thesaurus” (93,3%). Con relación al término “Knowledege Organization Systems” verificamos un rendimiento muy bueno (90%), y para el término “Knowledge Organization” un buen funcionamiento (83,3%). En contrapartida, para los términos “Organization of Knowledge” y “Classification Research Group”, las tres herramientas presentaron un rendimiento razonable (73,3%) e insuficiente (29,2%), respectivamente.

d) Precisión: Para verificar el índice de precisión utilizamos la fórmula $\frac{100 \text{ x } R}{L}$ establecida por Cleverdon en 1950 (Piedade, 1983), donde R = nº. de documentos relevantes recuperados y L = nº. total de documentos recuperados. La fórmula fue aplicada para el análisis del índice de recuperación y la información. La precisión, en este caso, está relacionada con los primeros diez resultados de búsqueda obtenidos en relación al

[5] 100% - Excelente; 95% a 99,9% - Muy muy bueno; 90% a 94,9%- Muy bueno; 80% a 89,9% -Bueno; 60% a 79,9% - Razonable; 30% a 59,9% - Pésimo; 0 a 29,9%-Insuficiente.

conjunto de documentos relevantes recuperados por los tres motores de búsqueda, mediante el análisis de la temática, o sea, de lo que trata el documento. Observamos que el índice de recuperación de la información superó las expectativas, lo que significa que muchos documentos relevantes en la Web, a partir de los términos utilizados para la búsqueda de información, dejaron de ser recuperados. Ello demuestra que los mecanismos utilizados en la búsqueda de información tuvieron un escaso papel en la recuperación del conjunto total de documentos de cada uno de los motores de búsqueda.

e) Cobertura: La tasa de cobertura fue medida por la proporción de los ítems encontrados utilizando cada buscador, en relación al total de ítems recuperados y considerados relevantes por los tres motores de búsqueda:

$$\text{Tx Co} = \frac{\text{nº. de puntos por mecanismo}}{\text{nº. total de puntos por los tres mecanismos}} \times 100$$

Aún utilizando las tres herramientas para la realización de la búsqueda de información en la Web, no recuperamos los treinta documentos, a partir de lo que escogimos - trabajar únicamente con los diez primeros resultados de búsqueda -. Lo máximo que se recuperó fueron 26 ítems, en el caso del término Automatic Classification y 19 ítems con el término Thesaurus. Los resultados indican que cada mecanismo de búsqueda, por sí solo no es suficiente para recuperar los ítems que respondan de manera completa a una necesidad de información sobre un determinado asunto.

4.2 Términos

El lenguaje natural por si solo ya promueve el fenómeno de dispersión, considerando que no existe control de sinónimos, homónimos, acrónimos, etc.

Es importante destacar que cuando el término es escrito de forma incorrecta en la caja de búsqueda o cuando el usuario no utiliza una combinación de términos, el sistema considera su *default*[6]. No sabemos si utilizan la lógica booleana o la lógica difusa.

Observamos en algunos documentos recuperados en *view source* que, aun los que eran de autoría del NordicW4 y del Desire, que son proyectos de representación, organización y recuperación de la información en la Web, muchas veces no presentan los términos que representan los asuntos debidamente indexados. Presentan elementos como autor, título y otra descripción, o sea, fueron indexados con html o metadatos, pero no se preocuparon de la representación correcta, para que se pudiera obtener una futura recuperación por asunto.

Puede verificarse que en algunos de los casos el documento era pertinente, pero no fue recuperado por el motor de búsqueda por el hecho de que la palabra no fue indexada.

Verificamos que cuando había proximidad de palabras, la palabra pasaba a ser considerada término a medida que este término estaba constituido por un grupo de palabras y los softwares realizaban la lectura como expresiones, pues identificaban la ocurrencia de palabras.

[6] Decisión que el software del mecanismo de búsqueda adopta en la ausencia de decisión por parte del usuario en el momento de la combinación de términos para la búsqueda de información en la Web.

Partiendo de las definiciones y de la relación de términos/conceptos – a la luz de la Teoría General de la Terminología -y la Teoría del Concepto, y del conjunto de términos/sistema de conceptos – a la luz de la Teoría de la Clasificación Facetada-, podemos dar sentido a la elaboración de sistemas de conceptos cuando ocurre la representación del conocimiento en la Web.

Es importante destacar que a medida que la selección de un término era realizada, y hacíamos la búsqueda de la información, éste ya era asociado a otros términos en nuestras mentes. A partir de esta asociación se configuraba una verdadera red de conceptos. Por tanto, observamos que este fenómeno estaba relacionado con el rizoma de Deleuze y Gatarri y con el árbol baniano de Ranganathan.

4.3 Ranking

Como se ha señalado anteriormente, la búsqueda de la información en la Web depende de un software que rastrea Internet buscando respuestas para satisfacer las necesidades de información de los usuarios, de un software que busca en Internet las informaciones relevantes y de un software que indexa los resultados.

Los criterios de relevancia para ordenar estos resultados son establecidos mediante algoritmos y, dentro de los criterios utilizados por estos algoritmos, están la localización y la frecuencia de palabras. Otros criterios utilizados son *metatags*, popularidad de los links, Direct Hit, inclusión de los sitios en el directorio, conceptos, pago y *Spam*. Estos criterios permiten que automáticamente se haga un ranking de los resultados obtenidos en la búsqueda.

Verificamos en esta investigación que la dispersión de la información tiene lugar cuando los mecanismos de búsqueda no recuperan los documentos que están distribuidos a lo largo de las ramas y que el propio mecanismo de búsqueda puede que no recupere, de hecho, un documento relevante para el usuario.

Los metadados y el HTML todavía no realizan un buen ranking de las páginas o de los documentos en la Web. Creemos que sea porque la utilización de la clasificación y la indexación automáticas todavía no han sido potenciadas. Los documentos, sitios y páginas disponibles en la Web cuando se presentan de manera organizada, permiten a los buscadores identificar los términos de forma adecuada y asociarlos de inmediato a las URL.

Los resultados que aquí se obtuvieron nos hicieron reflexionar sobre la configuración de una “topografía del conocimiento” con base en las relaciones conceptuales de forma que se permita la construcción de redes de conceptos que creemos sean el camino que posibilitará la representación del conocimiento para el perfeccionamiento de la recuperación de la información en entornos virtuales.

5 Consideraciones finales

Ya se sabe que la clasificación y la indexación automática, así como los metadatos han sido utilizados para la representación del conocimiento en entornos virtuales. A pesar de todo, es necesario conceder más atención al tratamiento de la información en estos ámbitos para que la búsqueda y recuperación de la información sean realizadas correctamente por los usuarios.

Los resultados que obtuvimos en el análisis empírico indican que la recuperación de la información en entornos virtuales puede ser mejorada con la utilización del corpus de conocimiento teórico y metodológico para que se puedan establecer relaciones conceptuales. Por eso, destacamos la importancia de estas relaciones para la representación del conocimiento en la Web.

Diversos estudios han sido realizados con el objetivo de plantear la discusión teórica y la ampliación de los estándares que faciliten y certifiquen al usuario la recuperación y visualización de la información, así como que permitan que no se recupere información indebida o que no fue solicitada.

Las dificultades que están relacionadas con las formas de representación en entornos actuales y virtuales constituyen desafíos para la racionalidad humana. El Hombre piensa en representación si ésta puede darle una expresión de racionalidad, a medida que relaciona principios, categorías, procedimientos y normas, en la medida exacta que construye modelos y algoritmos con la finalidad de que todos esos aspectos puedan estabilizar la tarea de representar la realidad de una manera comprensible.

Para mejorar la búsqueda y recuperación de la información en la Web es necesario construir herramientas que ofrezcan precisión y buena cobertura, y que posean un funcionamiento transparente para el usuario, permitiendo cierta flexibilidad en la interrogación. Los motores de búsqueda, en general, presentan una cobertura razonable, pero una baja precisión, poca transparencia y flexibilidad.

Sería bueno contextualizar las búsquedas, no sólo mediante jerarquías de conceptos en áreas del conocimiento, como en algunos sistemas de organización del conocimiento conclusivo, ilustrados por la arborescencia, sino también por medio de otros sistemas de organización del conocimiento que permitan la configuración de otras relaciones conceptuales en determinados dominios de conocimiento, así como un rizoma, en una configuración reticular, como ocurre con el procesamiento de la información en la mente humana.

De esta manera, mediante las búsquedas conceptuales, solucionaríamos los problemas de adecuación temática, pertinencia, precisión, relevancia, cobertura, transparencia y flexibilidad. La información debería ser tratada en los entornos virtuales, a la luz de principios y teorías que permitieran la representación y la organización del conocimiento con el objetivo de recuperar la información utilizada en algún momento en los entornos actuales.

Concluimos que a partir de un sistema de conceptos construido de esta manera, se consigue situar los términos de búsqueda en un área de conocimiento, identificando sus relaciones arborescentes, del tipo género/especie, y rizomáticas, del tipo asociativa. Ello obedece a que la búsqueda de la información en la Web se realiza mediante la utilización de palabras como si fueran símbolos, es decir, es la palabra por la palabra – cuestiones sintéticas-, y no una palabra que representa un significado – cuestiones semánticas-. Cuando se trabaja con la palabra, se opera con el lenguaje natural, y cuando se trabaja con los conceptos se opera con el lenguaje artificial, esto es, con el lenguaje construido para fines de representación del conocimiento y recuperación de la información, o más concretamente, con un sistema de organización del conocimiento.

Con base en los análisis realizados, podemos afirmar que para mejorar la búsqueda y la recuperación de la información en entornos virtuales sería necesaria la elaboración de una propuesta de un sistema de navegación conceptual a partir de un modelo de sistema de organización del conocimiento, donde el usuario a partir del procesamiento de información podría establecer las relaciones conceptuales (re)configurando su rizoma mental, permitiendo recuperar la información deseable para completar su necesidad informativa de manera más correcta.

Pero el conocimiento debe ser representado y organizado mucho más allá de las relaciones arborescentes, pues las conexiones cognitivas surgen naturalmente, conforme a la necesidad del individuo de adquirir y producir conocimiento en determinados dominios, lo que provoca relaciones reticulares, aquí denominadas relaciones rizomáticas.

Bibliografía citada

BINWAL, J. C. Ranganathan and the universe of knowledge. *Int. Classif*., 1992, vol. 19, n. 4, p.195-200.

BRIET, S. *Qu'est-ce que la documentation?* Paris: Edit, 1951.

CAMPOS, M. L. de A. *A organização de unidades do conhecimento em hiperdocumentos: o modelo conceitual como um espaço comunicacional para realização da autoria. 2001*. Tese (Doutorado em Ciência da Informação)– Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2001.

DAHLBERG, I. Knowledge organization: its scope and possibilities. *Knowl*. Org., 1993, vol. 20, n. 4, p. 211-222.

DELEUZE, G.; GUATTARI, F. Introdução: Rizoma. En: *Mil platôs: capitalismo e esquizofrenia.* Tradução Aurélio Guerra Neto e Celia Pinto Costa. Rio de Janeiro: Ed. 34, 1995. (Coleção Trans). Vol. 1, p. 11-37.

GOMES, H. E. Elaboração *de tesauro documentário: aspectos teóricos e práticos*. 1996.

HODGE, G. *Systems of knowledge organization for digital libraries: beyond traditional authority files. Washington*, D.C.: Clir Publication, 2000.

LANCASTER, F. W. *Indexação e resumos: teoria e prática*. 2ª. ed. Brasília, DF: Briquet de Lemos/Livros, 2004.

LÉVY, P. *As tecnologias da inteligência: o futuro do pensamento na era da informática*. Tradução Carlos Irineu da Costa. Rio de Janeiro: Ed. 34, 1993.

MIRANDA, M. L. C. de. A Organização do Conhecimento e seus paradigmas científicos: uma abordagem epistemológica. *Informare - Cad. Prog. Pós-Grad. Ci. Inf*., jul.-dez. 1999, vol. 5, n. 2, p. 64-77.

PIEDADE, M. A. R. *Introdução à teoria da classificação*. 2ª ed. rev. y ampl. Rio de Janeiro: Interciência, 1983.

VICKERY, B. C. Knowledge representation: a brief review. *J. Doc*., Sept. 1986, vol. 42, n. 3, p.145-159.

WERSIG, G.; NEVELING, U (Comp.). Terminology of documentation= Terminologie de la documentation: a selection of 1,200 basic terms published in English, Grench, German, Russian and Spanish. Paris: The Unesco Press, 1976.