Relevancia de las recuperaciones con texto completo

Miguel-Ángel López Alonso
Universidad Carlos III de Madrid

0.1. Resumen

El problema de la conceptualización de la relevancia de las recuperaciones con texto completo, ha tomado nueva importancia en los Sistemas de Gestión de la Información que interactúan en entornos multidimensionales de redes distribuidas tipo Internet. Dado que la valoración de la relevancia de las recuperaciones hipertextuales tiene poco que ver con el modelo experimental de Cranfield, se ha aceptado la superación de las medidas cuantitativas tradicionales imperantes en los años 80.

En este trabajo se analiza el estado de la cuestión, partiendo del análisis de la inconsistencia en la representación de la información, y del fracaso de los intentos de relacionar la profundidad de la indización con la relevancia de las recuperaciones, desde el llamado Paradigma Sistémico.

A continuación, se estudia la relación de dicha inconsistencia con la indeterminación del Lenguaje Natural para organizar la información, y los intentos de medir la relevancia desde un enfoque cualitativo centrado en el usuario, como sujeto fundamental del proceso de recuperación de la información. Se proponen varios procedimientos para tratar de paliar dicha indeterminación, destacando la integración de los Tesauros Conceptuales en los interfases hombre-máquina, diseñados específicamente como herramientas de recuperación desde el Paradigma Conceptual.

Finalmente, se apuesta por una fusión de los enfoques sistémico y subjetivo, de manera que permita la redefinición de la relevancia como un concepto multidimensional, dinámico y complejo, pero, medible dentro de un nuevo Modelo Conceptual Integrado de la Información. Este deberá apoyarse en potentes estructuras del conocimiento, capaces de representar al mismo tiempo las complejas relaciones sintagmáticas del razonamiento analógico, sus participantes y sus correspondientes actuaciones. (Autor)

Palabras clave: Relevancia. Tesauros conceptuales. Modelo cognitivo.

0.2. Abstract

The problem of the conceptualization of full-text retrieval relevance has taken new importance in the Information Systems that interact in multidimensional environments of distributed nets, type Internet. Since the evaluation of hypertext retrieval relevance has little to do with the experimental pattern of Cranfield, the traditional quantitative measures prevailing in the eighties have been surpassed.

This work analyzes the state of the question, leaving the analysis of the inconsistency in the representation of the information, and the disappointed intent of relating the indexing depth with the retrieval relevance, from the call Systemic Paradigm.

Next, it study the relationship of this inconsistency with the indetermination of the Natural Language to organize the information, and the intents of measuring the relevance from a qualitative focus centered in the user, acting as fundamental part of the process of retrieval information. It propose several procedures to try to palliate this indetermination, highlighting the integration of Conceptual Thesauri in the interfaces men-machine, designed specifically as retrieval tools from the Conceptual Paradigm.

We bet for a coalition of Systemic and Subjective Focuses, so that it allows the redefinition of the relevance as a multidimensional, dynamic and complex concept, but, appreciable inside a new Integrated Conceptual Model of the Information. This will lean on potent structures of knowledge, able to represent the complex syntagmatic and analogical reasoning, their participants and their corresponding performances, simultaneously. (Author)

Keywords: Relevance retrieval. Conceptual thesauri. Cognitive model.

1. Introducción

Desde que Bradford utilizara por primera vez el término "relevancia" en la década de los años 40, la consideración de la colección de documentos de un Sistema de Gestión de la Información como un todo, dividido en documentos relevantes y no relevantes y/o en documentos recuperados y no recuperados, ha ido evolucionando.

Los primeros experimentos de Cleverdon establecieron como estándares de medida de la relevancia de las recuperaciones en los Sistemas Convencionales de Recuperación de la Información: la precisión, como cociente entre el número de documentos relevantes y el número total de documentos recuperados, y la exhaustividad, como cociente entre el número de documentos relevantes recuperados y el número total de documentos relevantes existentes en la colección de documentos (1) .

Sucesivas revisiones de los experimentos de Cranfield realizadas en los años ochenta, coincidieron en que:

> Los efectos no deseados de las búsquedas en Lenguaje Natural no son el resultado de la mala indización de los textos buscados, sino de la falta de precisión en la ecuación de búsqueda planteada por el usuario (2).

Los precursores en la medición experimental de la relevancia en grandes Bases de Datos Jurídicas con texto completo Blair y Maron, observaron en sus tests: una exhaustividad de alrededor del 20% y una precisión no mayor del 79% en el Sistema STAIRS de IBM; esto les indujo a pensar que o bien los documentos no estaban indizados adecuadamente o que las búsquedas no habían sido bien realizadas (3).

Los recientes experimentos TREC, que abarcan extensas colecciones de textos de gran tamaño y fomentan la interacción entre los distintos grupos de investigadores, están permitiendo realizar un amplio espectro de investigaciones relacionadas con la recuperación de la información. Sin embargo, las conclusiones alcanzadas por TREC-1, con respecto a la evaluación de las recuperaciones, no están demasiado claras. Posteriormente, los tests que inicialmente utilizaban el Modelo Convencional de Recuperación de la Información han sido reconvertidos en los nuevos "TREC interactivos", que permiten la validación del Modelo Experimental.

Los últimos estudios de Soërgel (1994), han demostrado que:

> La precisión de las recuperaciones en línea no depende únicamente de las características de la previa indización, sino también de la eficacia del sistema de búsqueda utilizado y de la competencia del usuario para adaptarse a éste y reajustar las búsquedas mediante el uso de los términos asociados sugeridos por tesauros especializados (4).

Experimentos al respecto, de Al-Hawamdeh et al. y de Savoy (5) en sistemas hipertextuales, parecen confirmar dichas afirmaciones.

Ellis propone una nueva visión de la relevancia de las recuperaciones, contraria a la tradicionalmente fundada en la cuantificación, que se apoye en el reconocimiento de los métodos cualitativos relacionados con el conocimiento, los juicios de valor y los aspectos afectivos de la interacción sistema-usuario (6).

Igualmente, Marchionini y Barlow apoyan la necesidad de nuevos ratios o medidas cuantitativas para la evaluación de las recuperaciones en los Sistemas de Gestión de la Información en línea (7), junto con la necesidad de un seguimiento que evalúe la relevancia de las búsquedas en el entorno de trabajo del usuario (8). En esta línea, Schamber propone dos criterios de medición directa por los usuarios (9): el desarrollado por Wang sobre la base de once premisas para la selección de los documentos (10), y el propuesto por Janes para la evaluación de los

"criterios de relevancia" a partir del utilizado previamente en la medición del "desarrollo instrumental" (11).

Recientemente se han abierto camino los nuevos Sistemas de Recuperación de la Información en línea que utilizan mecanismos de navegación hipertextual y adoptan técnicas cualitativas de medición de la eficiencia de las recuperaciones muy próximas a las hipótesis de Swanson (12) sobre un "nuevo tipo de relevancia" que toma en consideración las opiniones del usuario en cuanto a los documentos recuperados. Estas miden la eficiencia del sistema hipertextual por medio de dos nuevas medidas cualitativas: la exactitud (que recuerda la precisión de los sistemas convencionales) y la integridad (que evoca la exhaustividad de los sistemas convencionales).

Sin embargo, con la navegación de los recursos documentales en redes han surgido serias dudas al mantenimiento de las medidas cuantitativas tradicionales de la relevancia: la exhaustividad y la precisión. Parece que los problemas que afectan a la valoración de la relevancia hipertextual tienen poco que ver con el modelo experimental de Cranfield, pensado para medir la relevancia de elementos aislados en una reducida base documental. La necesidad de desarrollo de "paradigmas diferentes" para la evaluación y diseño de los nuevos instrumentos de recuperación de la información parece innegable.

2. La inconsistencia en la representación de la información

Si partimos del hecho conocido que los Gestores de la Información tradicionales no recuperan la información pedida, sino una representación aproximada de ella: su indización. Podemos afirmar que la consistencia de ésta ultima caracterizará la relevancia de la recuperación, y la relacionará directamente con los descriptores que describan el contexto del documento: autor, título, fecha, longitud, tipo, editor, etc., o sus materias cubiertas.

Además, esta selección de conceptos por el indizador varía en función de:

- El área temática de los documentos a indizar.
- Su nivel de entrenamiento y experiencia.
- Si se usa o no un tesauro y su nivel de desarrollo.
- La "carga emocional" que ponga en la argumentación sobre los temas de los documentos indizados, (13) y de ...
- Su conocimiento del dominio científico a indizar, etc.

Todos estos aspectos deberán influir en las estrategias de búsqueda a aplicar, aunque no las defina totalmente.

2.1. Fracaso de los intentos de relacionar la profundidad de la indización con la relevancia de las recuperaciones

Durante los últimos treinta años se ha insistido en la relación inversa existente entre las dos medidas de la relevancia más utilizadas: la exhaustividad y la precisión. Unida a ella se ha seguido otra discusión subyacente en la que se cuestiona si el aumento o reducción en el número de términos de indización asignados a cada documento o profundidad de la indización, producía algún tipo de efecto conjunto sobre ambos ratios?.

Para mediar en la anterior discusión, se han realizado recientes revisiones de los experimentos de Cleverdon por Seely (14) , Svenonius (15) , etc., en el ámbito de las Bases de Datos Bibliográficas de distribución comercial. Aunque no se ha podido disponer de ningún experimento, lo suficiente sofisticado y extenso, que controle todas las variables (16) , se comentan algunos de los parciales más conocidos:

1) Funk y Reid, en su estudio de consistencia en la base de datos MEDLINE, han señalado algunas de las limitaciones de los estudios anteriores:
   a) El pequeño número de documentos, lo que les hace perder la representatividad de la indización total.
   b) El deseo de los indizadores por ser evaluados, lo que diferencia los tests de sus habituales condiciones de trabajo, y ...
   c) La medición de solo un pequeño número de términos de indización (17) .
2) Boyce y McLain, durante su investigación de la base de datos ONTAP ERIC en Dialog, confirmaron que la disminución en la profundidad de indización en las Bases de Datos comerciales producía la ya conocida disminución de los documentos recuperados, junto con un incremento en la precisión de éstos últimos (18).
3) Sievert y Andrews (19) , con su estudio de la consistencia de indización en la Base de Datos "Information Science Abstracts", demostraron que según aumentaba la profundidad de la indización disminuía su consistencia.

Se han confirmado los resultados equívocos de estos tests parciales, adelantado ya por Svenonius, y la necesidad de continuar con investigaciones más profundas que demuestren plenamente el fracaso de la relación directa entre la profundidad de la indización y la efectividad de las recuperaciones.

Podemos afirmar, pues, que el complejo fenómeno de la relevancia abarca relaciones a muy distintos niveles:

1) Jerárquicas, estructurales o de equivalencia, típicas del desarrollo de los lenguajes documentales.
2) Entre el lenguaje del texto y la capacidad lingüística del usuario, o entre el contenido sustancial del texto y la amplitud del conocimiento básico del usuario, o ...
3) Más complejas, entre el tema del pasaje del texto y el tema demandado por el usuario ("topic matching relations"), diferentes de las habituales relaciones de equivalencia temática desde el punto de vista del sistema.

2.2. Intentos de hacer corresponder el concepto de relevancia con la satisfacción de los usuarios

Ya Farradane, en su Comunicación de 1970 a la 44ª Conferencia de Aslib en la Universidad de Aberdeen, propugnaba que los resultados del almacenamiento y recuperación integrada de la información debían ser medidos desde un enfoque tripartito:

1) El de la representación del conocimiento con exactitud y su fácil recuperación (indización).
2) El de la organización de los registros de información grabados manual o mecánicamente, mediante codificación u otros procedimientos de transcripción (clasificación), y ...
3) El de la interpretación y satisfacción de los requerimientos de los usuarios (evaluación) (21).

Transcurridos más de veinticinco años, vemos que en las dos primeras áreas se ha avanzado enormemente, y esto nos lleva a la duda sobre si el relativo estancamiento en el tema de la medición de los resultados podrá derivarse del olvido del principal beneficiario de ellos, el usuario final, que es quien, como principal actor y protagonista, precisa controlar la abrumadora cantidad de información a su alcance.

Para apoyar las anteriores afirmaciones recordemos las conclusiones de:

1) Un reciente estudio llevado a cabo por T. Su, con la finalidad de determinar la efectividad de veinte mediciones de la exactitud de diversas Recuperaciones de Información en Lenguaje Natural, entre las que se han incluido (como referencia) la exhaustividad y la precisión. Tomando la opinión de los usuarios (22) como medida con la que las otras veinte debían correlacionarse, llegaron a la conclusión principal de que "no es

la precisión la medida que mejor se correlaciona con los juicios de los usuarios sobre los resultados obtenidos sino la exhaustividad total" (23).

2) Las de Schamber, que confirman la preferencia de los usuarios por la calidad más que por la cantidad de las referencias. Y que, los sistemas que miden los resultados obtenidos desde la perspectiva del usuario (incertidumbre y ansiedad) son los que mejor afinan la relevancia de las recuperaciones; en vez de los que ponen el énfasis en plantear aquellos tipos de búsquedas que mejor encajan con la representación de los textos depositados en ellos, desde la perspectiva del sistema (certidumbre y orden) (24).

Los sistemas que se basan en la interacción con el usuario centran el estudio de la relevancia de las recuperaciones en:

a) La evaluación por el usuario de la utilidad de la información obtenida para la resolución de su problema, y ...

b) La integración de los resultados, en las sucesivas y más depuradas ecuaciones de búsqueda del usuario, con vistas a obtener la mayor sinergia sistema-usuario.

2.3. Aplicación de la heurística como medio de incrementar la exhaustividad y la precisión de las recuperaciones

Los modelos que representan la opinión de los usuarios (25) son bastante más complejos que los modelos centrados en el sistema, ya que deben asimilar la información de las distintas fuentes para realizar su elección a través de la correspondencia de sus tres campos de actividad:

- El físico, que trata de la acción realizada.
- El afectivo, que trata de las opiniones expresadas, y ...
- El cognitivo, que trata de los conceptos del proceso y del contenido (26).

Mientras que los aspectos relacionados con el campo físico se investigan dentro del campo más amplio de las relaciones sistema-usuario, que tratan los interfases interactivos. Los otros dos campos: el afectivo y el cognitivo, deben concretarse dentro de las investigaciones sobre Relevancia Psicológica (27) que desarrollan los diferentes Sistemas de Gestión de la Información.

Harter y Rogers investigaron la falta de reglas de actuación en la recuperación de la información en línea, y propusieron como alternativa la potenciación del razonamiento inductivo tomado de la experiencia del profesional especializado en búsquedas, al que llamaron "Relevancia Psicológica".(28)

Su mayor importancia radica en que, ésta es la línea seguida por los prototipos de Agentes Expertos una década después al adoptar las distintas tácticas

(acciones, operaciones mentales, deseos y actitudes) del trabajo habitual de varios profesionales en un área concreta, para transformarlas en las reglas de la Base de Conocimientos de un Sistema Experto.

A partir de la heurística, dos serán las actuaciones con las que se mejorará la relevancia de las recuperaciones:

1) La del alcance del tema buscado o precisión de las búsquedas:
   — Sea modificando la especificidad dentro de las facetas existentes (variando el número de términos que representan un determinado concepto buscado).
   — Sea modificando el número de combinaciones booleanas de las facetas (añadiendo o anulando facetas completas o cambiando la lógica de las búsquedas).
2) La del método de búsqueda de los ficheros invertidos o exhaustividad de las búsquedas, que implica una pequeña pérdida de precisión:
   — Sea variando cómo los términos deben ser buscados dentro de los campos o la lógica de la ecuación de búsqueda (redefiniendo los campos en que se debe buscar).
   — Sea modificando los criterios de recuperación sin cambiar los términos mismos (con revisión la adyacencia de los operadores y limitación del idioma de los documentos a recuperar, la publicación a buscar o el autor, etc.), o ...
   — Sea incrementando el truncamiento de los términos para utilizar otros términos alternativos.

3. La indeterminación del Lenguaje Natural para la organización de la información documental

En cuanto a la mejora de la relevancia de las recuperaciones de documentos en los Sistemas de Gestión de la Información, la inconsistencia en la representación de la información aparece relacionada con la indeterminación del Lenguaje Natural para la representación del conocimiento, que según Maniez tiene como explicación la relación de los individuos con su contexto sociocultural:

> La envoltura sensible de la palabra oral o escrita no establece una relación directa entre las palabras y las cosas, sino entre las palabras y nuestras representaciones mentales del mundo, procedentes de un desglose de la realidad sociocultural propio de una época, sociedad o cultura, transportado por la lengua (29).

Pero, mientras que en la comunicación lingüística las palabras se clarifican entre sí según el contexto, en las búsquedas documentales esta ambigüedad es un obstáculo que conduce al usuario a encontrar documentos distintos de los busca-

dos. Esto obliga a establecer los términos de indización de la forma más unívoca posible, para evitar las respuestas "parásitas" o ruido documental en las recuperaciones. Los Lenguajes Controlados buscan un acercamiento entre los conceptos documentales y los lingüísticos que mejore el análisis de la información almacenada en las Bases de Datos Documentales, para su exacta recuperación.

Aunque los estudios de Markey, Atherton y Newton (30) ya habían señalado que las recuperaciones con Lenguaje Natural producían mayor exhaustividad y menor precisión que las realizadas con un vocabulario controlado, será Boyce quien afirmará rotundamente que:

> Dado que cada documento recuperado mediante una búsqueda, a partir de un término descriptor, es uno de los documentos recuperados por otra búsqueda basada en sus términos componentes en el índice inverso primitivo, será imposible que la primera búsqueda, con un vocabulario controlado, recupere más documentos que la segunda con el lenguaje libre.

Ésto le induce a afirmar que:

> El vocabulario controlado, usado en los grandes Sistemas Comerciales de Gestión de la Información, constituye un elemento de precisión en virtud de la propia estructura de sus registros informáticos y nunca un elemento de exhaustividad (31).

En esta línea, podemos afirmar que la menor exhaustividad y mayor precisión inherente a las recuperaciones realizadas con los Lenguajes Controlados, dependen del hecho intrínseco que las caracteriza: ser una reducción en la cantidad de términos utilizados para la búsqueda. Y, subsidiariamente, de la calidad de los vocabularios controlados utilizados, por ejemplo: del número de términos asignados a cada descriptor, de su capacidad para evitar los homónimos, etc. (32).

Bates considera que la indeterminación del Lenguaje Natural está arraigada en el caos de naturaleza de la mente humana, según el Principio de Incertidumbre de Heisenberg (de la Mecánica Cuántica), aplicado por Rosenberg a algunos comportamientos humanos relacionados con la Ciencia de la Documentación e Información (33). Como consecuencia, sugiere que con la aplicación del Principio de Variedad de Requisitos de Ashby a los Sistemas de Gestión de la Información se obtendrá un acoplamiento perfecto entre las descripciones de los documentos y las preguntas de los usuarios (34).

3.1. Propuestas para evitar dicha indeterminación del Lenguaje Natural

A pesar de las teorías de Bates y de los experimentos de Blair y Maron (35), siguen sin dominarse las causas profundas de la indeterminación del Lenguaje Natural que provocan la inconsistencia en la Organización de la Información.

Para mejorar su efectividad práctica, proponemos las líneas de acción que se han convertido en un estándar de dichos Sistemas:

1) Reducir la diversidad en la descripción de los documentos, mediante el uso en la indización de tesauros especializados, y ...
2) Aumentar la variedad de las ecuaciones de búsqueda de los usuarios, con la potenciación de las Bases de Conocimientos Terminológicos y sus correspondientes Tesauros Conceptuales.

En los apartados siguientes analizaremos algunos de los procedimientos propuestos para evitar las causas de esta indeterminación. Partiremos de estimar que se derivan, tanto de la naturaleza de la mente humana (propuesta de Bates) como de la inadecuada aplicación de la heurística y el análisis inferencial al proceso de recuperación de la información (propuesta de Harter).

3.2. El análisis del contenido textual

A pesar de los avances logrados en la mejora de los Ratios de Efectividad en la Recuperación de la Información, desde una perspectiva de interacción entre el usuario y el sistema. Algunos autores como Bates (36) o Beghtol (37) insisten en la necesidad de reestudiar la información misma: su estructura y organización, dados los problemas de indeterminación derivados de la gran flexibilidad y creatividad del Lenguaje Natural. De la misma opinión se muestran Molto (38) o Moreiro (39) al propugnar el perfeccionamiento en la selección conceptual de los temas objeto de estudio de cada documento, mediante el análisis semántico y discursivo del contenido textual.

Estamos plenamente de acuerdo con ellos en que, para evitar la indeterminación intrínseca a dicho lenguaje, debería ahondarse en el Análisis del Contenido de los documentos; tanto desde un nivel microestructural o inductivo (de abajo-arriba) como desde un nivel global o deductivo (de arriba-abajo) que abarque la macroestructura o tema general sobre el que trata cada texto. Al analizar la superestructura o macroestructuras parciales de cada texto, se avanzará en una comprensión profunda del texto, como conocimiento más especializado que deberá confirmar la macroestructura global. Según Beghtol (40), tendrá lugar durante la lectura profunda del texto y, será un proceso como de "controlado olvido", gobernado por Macroreglas que "expulsen" provisionalmente la información menos importante, con objeto de que aparezca claramente identificado el tema principal del texto.

Este Análisis de Contenido ayuda a profundizar en la selección de los conceptos y en su encuadramiento dentro de las diferentes áreas del conocimiento, para proporcionar información útil que ayude a superar los puntos débiles del Lenguaje Natural, tanto al indizador como al recuperador e, incluso, a los diseñadores de los Sistemas de Gestión de la Información.

Llevando a sus últimos efectos la tesis anterior, Sillince ha propuesto que, para evitar la inconsistencia de la indización basada en contenidos concretos (semánticos, etc.) se la combine con otras aproximaciones más complejas, tales como:

- El uso de Metareglas, del tipo de las propuestas por Fugmann (41).
- La discriminación entre las premisas y las hipótesis, y ...
- La construcción de la argumentación con opiniones personales y culturales, que ayuden al indizador a concentrarse en el establecimiento de los objetivos de la argumentación (42)

3.4. Incorporación en el sistema de un Tesauro diseñado para la recuperación, como elemento de precisión del Lenguaje Natural

Otro de los procedimientos propuestos para evitar los equívocos del Lenguaje Natural, consiste en incorporar los conceptos obtenidos del Análisis Documental de contenido a un Tesauro Conceptual, que pueda ser utilizado para las recuperaciones desde el interfase sistema-usuario.

Este procedimiento incrementa la variedad de las ecuaciones de búsqueda del usuario, o al menos le muestra la variedad de las existentes; lo que, de acuerdo con el Principio de Variedad de Requisitos de Ashby (propuesto por Bates en la Recuperación de la Información), incrementa simultáneamente la cantidad de respuestas del sistema consultado.

Los Tesauros Conceptuales, diseñados específicamente para ayudar en la enunciación de las preguntas en la fase de Recuperación de la Información, propuestos por autores como Bates (43), Lancaster (44) o Schmitz-Esser (45), palian la indeterminación de las búsquedas en Lenguaje Natural, especialmente en aquellas Bases de Datos cuyos documentos con texto completo no hayan sido previamente indizados con ningún otro tesauro.

Deberán diferenciarse de los Tesauros Documentales usados para la indización, según con las siguientes características:

a) Listar todos los términos no "vacíos", usados en cualquier momento en el Catálogo o en la Base de Datos.

b) Distinguir cuidadosamente los términos realmente usados de los no usados.

c) Añadir notas de alcance que aclaren las dudas a los posibles usuarios, incluso aportando algunas definiciones.

d) Contener equivalencias autoexplicativas de los términos y/o sus relaciones.

e) Aportar un extenso vocabulario que sea superior al número de términos controlados.

f) Incluir los términos coloquiales, variaciones de los términos reconocidos e incluso truncamientos.

Diversos experimentos en que los usuarios son apoyados, en la enunciación de sus ecuaciones de búsqueda, con términos adicionales extraídos de un tesauro diseñado específicamente para la recuperación con Lenguaje Natural en grandes Bases de Datos; han aportado avances significativos en el conocimiento intrínseco de la relevancia de las recuperaciones (46) , y se ha llegado incluso a doblar la precisión en el número de documentos recuperados si el usuario selecciona y usa los términos sugeridos como adicionales a sus propios términos (47).

Según los experimentos de Jones, que tratan de medir los efectos de la ampliación de las búsquedas mediante este tipo de tesauros, los mejores resultados se obtienen cuando los usuarios:

a) Disponen de una gran cantidad de términos para elegir del campo específico tratado, sean suyos, de la Base de Datos o incluso de indizaciones previas, y ...

b) Controlan interactivamente el proceso de navegación por el tesauro, en vez de utilizar procedimientos automáticos (48).

Recientemente, Harter y Cheng han propuesto una nueva técnica, denominada de "descriptores vinculados", que permite avanzar automáticamente a partir dos o más términos de búsqueda hacia otros términos que mejoren los originales en cuanto a calidad de las recuperaciones. El uso de esta técnica aporta un fuerte argumento para la compilación de más ricos y complejos Tesauros Conceptuales que reflejen la mayor cantidad de vínculos posibles entre descriptores (49).

3.5. Técnicas de parcial equivalencia entre los términos de indización y los de recuperación

3.5.1. Corriente conceptual

De entre las dos corrientes principales que han investigado el uso del Lenguaje Natural para la recuperación de la información textual, hemos analizado brevemente dos de los procedimientos de la denominada corriente "conceptual" que tratan de evitar los equívocos de dicho lenguaje mediante el procesamiento semántico de los textos y su acceso intelectual.

Esta corriente "conceptual" considera la información por encima del mero significado y la hace depender tanto del contexto como de la intencionalidad, para llegar a definirla como

> Un suplemento contextual que, añadido a un objeto en forma de entidad semántica o estructura cognitiva, percibe su alcance en el ámbito individualizado. (Van Rijsvergen y Lalmas (50) o Ingwersen (51)).

Requiere, para su desarrollo práctico, que el sistema entienda la pregunta y cada uno de los textos de la Base de Datos.

Para Blair (52):

> "Su superioridad estriba en la posibilidad de recuperar documentos o representaciones de documentos que equivalgan al "concepto" intelectual representado por los términos de la ecuación de búsqueda, pudiendo no contener ninguno de los términos especificados en dicha ecuación siempre que tengan el contenido relacionado cognitivamente con ellos".

La crítica a esta corriente "conceptual" se basa en el reconocimiento de la complejidad de la tarea, dada la necesidad de disponer previamente de grandes cantidades de conocimientos terminológicos de la materia correspondiente, para poder procesar las ecuaciones de los usuarios (53).

3.5.2. Corriente lingüística

La otra gran corriente, la llamada "lingüística", que usa el Lenguaje Natural para la recuperación de la información mediante acceso físico, precisa para su desarrollo práctico que el sistema entienda donde están colocados los documentos que tratan de similares "conceptos", sin necesitar entender el concepto en sí. Una vez recuperados físicamente los documentos, será el usuario quien deberá juzgar cual son los que mejor responden a su pregunta.

Estos sistemas concentran su análisis en el nivel morfo-sintáctico del lenguaje para obtener la mayor independencia posible del área de conocimientos tratada, evitando la necesidad de disponer de información específica previa — Salton (54) o Fagan (55)—.

De las diferentes aproximaciones dentro de esta segunda corriente, se ha destacado la encabezada desde los años setenta por Metzler, la cual, a partir de la expansión de las búsquedas en línea en grandes Bases de Datos con texto completo, está siendo utilizada en el software de algunos rastreadores de Internet. Se apoya en las Técnicas de Equivalencia entre las frases de los textos y de las preguntas, para generar un valor medible en porcentaje (ponderación) que clasifique dichas frases según su proximidad ponderada. Esta aproximación basa su hipótesis en que: "En la composición semántica de grandes entidades lingüísticas, el aspecto (de la descripción sintáctica) que más destaca, es su estructura jerárquica."

Metzler construye estructuras arbóreas de clasificación jerárquica para representar las relaciones de dependencia entre los términos representativos del texto

y los de las preguntas, utilizando en la recuperación algoritmos que buscan la equivalencia entre ambas estructuras.

Sheridan y Smeaton comparten esta aproximación, y afirman que:

> La ambigüedad del texto en Lenguaje Natural puede ser codificada dentro de las estructuras jerárquicas de las entidades lingüísticas para ser utilizadas en el proceso de búsqueda de equivalencias y porque, además, puede proporcionar una equivalencia medible de estructuras que aporte algún tipo de indicación de la equivalencia entre las entradas del índice (57).

Blair la critica, por el contrario, y estima que la fortaleza de esta corriente "lingüística" se apoya en la dificultad que tienen las distintas aproximaciones "conceptuales" para medir sus avances, al no existir ratios estandarizados que comparen la efectividad de ambas corrientes en la recuperación de la información.

4. Conclusiones sobre el fenómeno de la relevancia

El estudio de la relevancia, a pesar de ser un tema dominante desde los años setenta, se ha enfocado desde dos orientaciones bien diferentes: la centrada en el sistema o la centrada en el usuario, sin que todavía en la actualidad se haya llegado a un consenso definitivo.

Dado que las Técnicas de Recuperación de la Información tratan de obtener resultados comparables a los que la localización y el examen de los documentos proporcionan en el espacio físico. Pensamos con Dominich (58) que unas medidas de relevancia ampliamente aceptadas pasan por la consideración de los documentos y de las ecuaciones de búsqueda como "unidades interrelacionadas", con el objetivo final puesto en que las recuperaciones sean útiles al usuario.

Para Dominich, los modelos que se apoyan en los puntos de vista del usuario propugnan la naturaleza subjetiva de la relevancia. Por ello, demanda el desarrollo de un nuevo Modelo Conceptual Integrado de la Información que, tome como punto de partida la "interacción total" y considere los documentos y las preguntas como entidades del mismo tipo, pero, que no considere a la relevancia como una entidad subjetiva (59), sino resultante de la interacción total entre dichas entidades.

Dado que la valoración de la relevancia es imprescindible para la completa comprensión del proceso cognitivo del comportamiento humano en la Organización de la Información, la profundización deberá partir de la consideración de su naturaleza como un concepto multidimensional, dinámico y complejo, pero, medible.

Green propone un tercer enfoque, que soslaye los planteamientos de los enfoques sistémico y subjetivo y los aúne, para lo que:

a) Define la relevancia como "la propiedad de un texto que puede ser potencialmente útil al usuario en la resolución de una necesidad", pero "restringida por la capacidad cognitiva del usuario para utilizar dicho texto dentro de los límites de un esfuerzo razonable", y ...

b) Afirma su carácter como noción teórica y no subjetiva (siguiendo a Dominich) y la imposibilidad de que sea evaluada únicamente por una de las partes implicadas en el proceso de recuperación de la información: "A pesar de que sea el usuario quien generalmente puede evaluar mejor la relevancia, a veces no es el perfecto juez del valor que un determinado documento puede tener para resolver su necesidad."

c) Discrepa en los factores que deben tenerse en cuenta para su evaluación, dado su carácter multidimensional (siguiendo a Schamber), y matiza que:

   1) Aunque, desde el punto de vista del usuario, deba tomarse como parámetro más importante la materia de la que el texto habla; teniendo en cuenta que:

      a) Este parámetro se extiende por encima del tema a su perspecti va: ideológica, metodológica, etc., y que ...

      b) El texto es solo una parte del documento completo y, por tan to, el tema tratado se circunscribe a ese área del documento.

   2) hay otros parámetros centrados en el usuario que deben tenerse en cuenta, como por ej.: el razonamiento analógico derivado de las relaciones sintagmáticas, etc.

Dado que los medios empleados por la mayoría de los Sistemas de Gestión de la Información no cumplen los estándares de sistematización y complejidad que permitan aprovechar los beneficios derivados del análisis conceptual de las relaciones sintagmáticas, urge el desarrollo de un Modelo Conceptual Integrado de la Información (del tipo demandado por Ingwersen) que se apoye en potentes estructuras del conocimiento (*gestalt structures*) capaces de representar al mismo tiempo estas relaciones, sus participantes y sus interacciones correspondientes.

5. Notas

(1) En su primer Proyecto de Cranfield (1962).

(2) Svenonius, E. (1986). Unanswered Questions in the Design of Controlled Vocabularies. // Journal of the ASIS. 37 : 5 (1986) 331-340.

(3) Blair, D.C. ; Maron, M.E. An evaluation of retrieval effectiveness for a full-text document retrieval system. Communications of the ACM: 1985. 28 : 3 (1985) 289-299. Con el estudio STAIRS levantaron una fuerte polémica que ha marcado las discusiones entre los diferentes investigadores en los últimos años, alentada por la opi-

nión contraria de Salton. Salton, G. Automatic Text Processing. New York: Addison-Wesley, 1989, 540 p.

(4) Soergel, D. (1994). Indexing and Retrieval Performance: The Logical Evidence. // Journal of ASIS. 45 : 8 (1994) 589-599.

(5) Al-Hawamdeh, S. et al. Paragraph-based access to full-text documents using a hypertext system. // PROGRAM 25(1991) 119-131. Savoy, J. Effectiveness of information retrieval systems used in a hypertext environment. // HYPERMEDIA. 5 (1993) 23-46.

(6) Ellis, D. (1996). The Dilemma of Measurement in Information Retrieval Research. // Journal of the ASIS. 47 : 1 (1996) 23-36.

(7) Marchionini, G. ; Barlow, D. (1994). Extending Retrieval Strategies to Networked Environments: Old Ways, News Ways and a Critical Look at WAIS. Journal of ASIS. 45 : 8 (1994) 561-564.

(8) Siempre desde la perspectiva de la relación entre la relevancia y la satisfacción del usuario que para M. Gluck son conceptos totalmente diferenciados. Gluck, M. Exploring the relationship between user satisfactionand relevance in information systems. Information Processing & Management. 32 :1 (1996) 89-104.

(9) Schamber, L. ; Bateman, J. (1996). User criteria in relevance evaluation: toward development of a measurement scale. // ASIS Conference Procedings, octubre 19-24, 1996.

(10) Wang, P. A. (1994). Cognitive Model of Document Selection of Real Users of IR Systems. Univ. of Maryland, tesis doctoral sin publicar, 1994.

(11) Janes, J.W. (1991). Relevance Judgments and Incremental Presentation of Document Representations, Information. // Processing Management. 27 : 6 (1991) 629-646.

(12) Swanson, D.R. (1986). Subjetive versus objective relevance in bibliographic retrieval systems. // Library Quarterly. 56 (1986) 389-398.

(13) Según los estudios de Van Dijk sobre manipulación de la memoria en el proceso de conceptualización de la información.

(14) Seeley, B. J. (1982). Indexing Depth and Retrieval Effectiveness. // Library Quarterly. 8 (1982) 201-208.

(15) Svenonius, E. Unanswered Questions in the Design of Controlled Vocabularies. Ibid.

(16) Revelándose como uno de los mayores problemas para la comparación de resultados el no poder encontrar facilmente documentos indizados con distinto grado de profundidad.

(17) Funk, M. ; Reid, C. A. ; Mcgoogar, L. S. Indexing Consistency in MEDLINE. // Bulletin of the Medical Library Association. 71 (1983) 176-183.

(18) Boyce, B. R. ; Mclain, J. P. Entry Point Depth and Online Searching using a controlled vocabulary. // Journal of the ASIS. 40 : 4 (1989) 273.

(19) Sievert, M. E. y Andrews, M. J.Indexing Consistency in Information Science Abstracts. // Journal of the ASIS. 42 : 1 (1991) 1-6.

(20) Paice, Chris D. (1991). A Thesaural Model of Information Retrieval. // Information Processing Management. 27 : 5 (1991) 435.

(21) Farradane, J. E. L. (1970). Analysis and organization of knowledge for retrieval. Aslib Proceedings. 22 : 12 (1970) 607.

(22) Sus objetivos, sus expectativas, su opinión de la relevancia de las referencias obtenidas y del tiempo utlizado para ello.

(23) T. Su, L. (1994). The Relevance of Recall and Precision in User Evaluation. // Journal of the ASIS. 45 : 3 (1994) 216.

(24) Schamber, L. (1994). Relevance and Information Behavior. // ARIST. 29 (1994) 3-48.

(25) Por la que éstos forman su punto de vista particular moviéndose desde el estado inicial de información hacia el estado final de resolución.

(26) Kuhlthu, C. C(1991). Inside the Search Process: Information Seeking from the User's Perspective. // Journal of the ASIS. 42 : 5 (1991) 362.

(27) Harter, S.P. (1992). Psychological Relevance and Information Science. // Journal of ASIS. 43 : 9 (1992) 602-615.

(28) Independiente del sistema de búsqueda, de la base de datos utilizada y de la búsqueda en sí misma y alternativo a los distintos algoritmos matemáticos de recuperación propuestos por los teóricos de la recuperación automatizada. Harter, S. P. ; Rogers P., A. Heuristics for online information retrieval: a typology and preliminary listing, Online Review.9 : 5 (1985) 407-424.

(29) Maniez, J. (1993). Los lenguajes documentales y de clasificación. Madrid : Pirámide, 1993. p. 199.

(30) Markey, K. ; Atherton, P. ; Newton, C. (1980). An Analysis of Controlled Vocabulary and Free Text Search statements in online searches. // Online Review. 4 (1980) 225-236.

(31) Boyce, B. R. ; Mclain, J. P. Entry Point Depth and Online Searching using a controlled vocabulary. Ibid.

(32) Soergel, D. Indexing and Retrieval Performance: The Logical Evidence. Ibid.

(33) Rosenberg, V. (1974) . The scientific promises of information science. // Journal of the ASIS. 25 : 4 (1974) 263-269.

(34) Solo la variedad puede destruir la variedad. Para que un sistema (máquina u organismo) funcione adecuadamente, éste debe generar tal variedad de respuestas hacia su entorno como de preguntas al sistema. Ashby, W. R. An Introduction to Cybernetics.London: Methven, 1973. p. 202-212. La variedad puede venir dada tanto en la forma de alteraciones físicas como de información. Bates, M. J. Subject Access in Online Catalogs: A Design Model. // Journal of the ASIS. 37 : 6 (1986) 361.

(35) Blair, D.C. ; Maron, M.E. An evaluation of retrieval effectiveness for a full-text document retrieval system. Ibid.

(36) Ibid.

(37) Beghtol, C. (1985). Facets as interdisciplinary undiscovered public knowledge. // Journal of Documentation. 51 : 3 (1985) 194-224.

(38) Molto, M. (1993). Improving full text search performance through textual analysis. Information Processing Management. (1993). 615-632.

(39) Establece la recuperación de la información desde indizaciones terminológicas y mediante combinaciones de lógica matemática. Moreiro, J. A. Aplicación de las Ciencias del Texto al Resumen Documental. Madrid: U. Carlos III-BOE, 1993.

(40) Ibid.

(41) Fugmann, R. (1985). The five-axiom theory of indexing and information supply. // Journal of the ASIS. 36 : 2 (1985) 116-129.

(42) Sillince, J. A. A. Argumentation-based indexing for information retrieval from learned articles. // Journal of Documentation. 48 : 4 (1992) 391.

(43) Ibid.

(44) Lancaster, F.W. Vocabulary control for information retrieval. Information Resource Press, 1986, XVII ed., 270 p.

(45) Schmitz-Esser, W. (1991). New Approaches in Thesaurus Application. // International Classification. 18 : 3 (1991) 143-147.

(46) Croft, W. B. ; Das, R. Experiments with query acquisition and use in document retrieval systems. // Proceeedings of the 13th Conference on Research and Development in Information Retrieval : Brussels, Belgium, 1990.

Kristensen, J. (1993). Expanded End-user's Query statements for free text searching with a search-aid thesaurus. // Information Processing & Management. 29 : 6 (1993) 733-744.

(47) Ekmekcioglu, F. C. ; Robertson, A. M. ; Willet, P. (1992). Effectiveness of query expansion in ranked-output document retrieval systems. Journal of Information Science. 18 (1992) 139-147.

(48) Jones, Susan ; et al. (1995). Interactive Thesaurus Navigation: Intelligence Rules Ok?. // Journal of ASIS. 46 : 1 (1995) p. 58.

(49) Harter, S.P.; Cheng, Y-R. (1996). Colinked Descriptors: Improving Vocabulary Selection for End-User Searching. // Journal of ASIS. 47 : 4 (1996) 311-325.

(50) Van Rijsbergen, C. ; Lalmas, M. An information calculus for information retrieval. // Journal of the ASIS, versión previa a la impresión, 1996.

(51) Ingwersen, P. (1996) . Cognitive Perspectives of Information Retrieval Interactions: elements of a cognitive IR Theory. // Journal of Documentation. 52 : 1 (1996) 34.

(52) Blair, D.C. (1996) . Stairs Redux: Thoughts on the STAIRS Evaluation, Ten Years After. // Journal of the ASIS. 47 : 1 (1996) 5.

(53) Ellis, D. (1992). The Physical and Cognitive Paradigms in Information Retrieval Research. // Journal of Documentation. 48 : 1 (1992) 45-64.

(54) Ibid.

(55) Fagan, J. L. (1987). Experiments in automatic phrase indexing for document retrieval: A comparaison of syntactic and non-syntactic methods. Doctoral Dissertation, 1987. Cornell U., Ithaca, NY. Technical Report 87-868.

(56) Metzler, D. P. ; et al.(1989). Constituent Object Parsing for IR and Similar Text Processing Problems. // Journal of the ASIS. 40 : 6 (1989) 398.

(57) Sheridan, P. ; Smeaton, A. F. (1992). The application of morpho-syntactic language processing to effective phrase matching. // Information Processing and Management. 28 : 3 (1992) 354.

(58) Dominich, S. (1994) . Interaction Information Retrieval. // Journal of Documentation. 50 : 3 (1994) 197-212.

(59) Dominich, S. Interaction Information Retrieval. Ibid.

(60) Schamber, L. Relevance and Information Behavior. Ibid.

(61) Green, R. (1995). Topical Relevance Relationships. Why Topic Matching Fails?. // Journal of the ASIS. 46 : 9 (1995) 648.

(62) Capaces de lograr la precisión en la recuperación de la información cuando todos sus elementos están integrados indisolublemente y se utiliza la mínima cantidad de energía para su procesamiento. Ibid.