GIL LEIVA, I, RODRÍGUEZ MUÑOZ, J.V. Tendencias en los sistemas de indización automática. Estudio evolutivo. *Revista Española de Documentación Científica*, 1996, vol. 19, nº 3, p 273-291.

TENDENCIAS EN LOS SISTEMAS DE INDIZACIÓN AUTOMÁTICA. ESTUDIO EVOLUTIVO

Isidoro Gil Leiva
José Vicente Rodríguez Muñoz
Departamento de Información y Documentación
Universidad de Murcia

Resumen: Se presenta la evolución de la indización automática y asistida por ordenador desde que se inician las primeras investigaciones, a finales de los años cincuenta, hasta la actualidad. Los sistemas descritos se han dividido en cuatro apartados: aquellos que utilizan métodos no lingüísticos, los que emplean criterios lingüísticos, los sistemas comercializados, y las últimas tendencias en indización automática de información multimedia. Para el estudio de esta evolución se realiza una amplia revisión bibliográfica que abarca cuatro décadas de indagaciones que han dado lugar a numerosos proyectos de indización automática o asistida por ordenador de los cuales, se muestran más de veinticinco. Finalmente se muestran las conclusiones.

Palabras clave: Indización automática, indización asistida por ordenador, sistemas de indización automática, tratamiento automático de textos, Estadística, Revisión bibliográfica

1 Introducción

La indización es catalogada como una operación compleja, pero esta dificultad se torna doble cuando se intenta obtener de forma automática. La complejidad se duplica porque, como veremos en los párrafos siguientes, en los intentos por lograr una indización automática intervienen en algunos casos disciplinas diversas como la Estadística, la Probabilidad, la Lingüística, la Programación, y por supuesto la Documentación. De esta última, se ha reconocido su interdisciplinariedad, pero sin lugar a dudas, a la indización automática (en adelante, Ia) también hay que inscribirla como una técnica tremendamente interdisciplinaria.

Antes de introducirnos más en la Ia creemos recomendable hacer referencia al debate recurrente que se inició a principios de los años sesenta y que aún se mantiene con total actualidad. Estamos hablando acerca de la preferencia o no de la indización automática. Entonces había y todavía hay, investigadores y profesionales de la Biblioteconomía y la Documentación que consideran que una máquina es incapaz de realizar convenientemente la labor de indizar, ya que no puede llegar a captar todos los matices conceptuales como puede hacerlo un indizador humano. Algunos alegan que para qué, servirse de la indización automática si los términos que proponen algunos

sistemas que existen actualmente, los debe validar el indizador humano, y hay quien desecha estos modos de análisis porque de momento está restringido a áreas del conocimiento muy concretas. Por contra, los defensores de la Ia alegan mayor economía tanto de proceso como de presupuesto, una mayor objetividad puesto que se aplicarían siempre los mismos parámetros, así como una disminución de los errores que son un claro inconveniente en el momento de la recuperación de la información en una base de datos.

Estas diferencias entre profesionales e investigadores hemos tenido oportunidad de observarlas recientemente, gracias a una interrogación que lanzamos a la lista de correo electrónico Iwetel. A pesar de que no intervino un gran número de abonados, si fue un debate interesante donde se pudieron comprobar opiniones encontradas con respecto a la idoneidad o no de la Ia.

Dejando de lado esta polémica, lo que si resulta evidente que éste es un campo de investigación con gran tradición en países como Francia o incluso Brasil, aunque principalmente, Estados Unidos lleva la iniciativa. Lo que presentamos en estas líneas es parte de la revisión bibliográfica practicada para iniciar el trabajo de investigación en el que estamos inmersos, que no es otro que buscar unas bases conceptuales-metodológicas para un sistema de indización automática para el español, que tendremos la oportunidad de presentar en otro trabajo. El presente estudio abarca casi cuatro décadas de indagatorias con el objetivo de sistematizar, siguiendo en la medida de lo posible un orden cronológico, la evolución ocurrida en la Ia para ofrecer, de un modo general, las diferentes corrientes seguidas. Sin embargo, debido a lo ambicioso de este repaso, somos conscientes que algunos de los aspectos a los que nos acercaremos se tratarán de un modo conciso pero no encontramos otra opción para poder abarcar tanto tiempo de indagaciones en este campo. Asimismo, esperemos que esta aproximación a la Ia sirva como punto de partida para la iniciación de ejercicios en esta dirección para la lengua española, puesto que se observa un marcado vacío con respecto a otros países.

Entrando ya en materia se puede anticipar que a finales de los años cincuenta y sesenta se produjo un incremento exponencial de la información científica disponible en todos los campos del saber, aunque preferentemente en las áreas de ciencias experimentales. Por estos motivos se fueron ideando sistemas de información cada vez más operativos, a la vez que aumentó el número de investigaciones sobre el tratamiento de la información con la finalidad última de atender de forma más eficaz y rápida las necesidades de información de los científicos. Paulatinamente, se fue generalizando la idea de que el ordenador constituía una herramienta muy útil para el procesamiento de textos y en especial para la indización, dado que se consideraba al ordenador objetivo en las operaciones repetitivas. De esta forma, se pretendía evitar que una persona pudiera indizar un documento de forma diferente en momentos distintos o que dos indizadores representaran un documento con términos desiguales. Además, una máquina es generalmente exacta y precisa en las operaciones, por lo que consideraban que se podrían minimizar los errores en la selección de términos para la indización.

Por tanto, el análisis automático de textos se convirtió en un arduo tema de investigación. Algunas de las causas que favorecieron este auge fueron la disponibilidad de máquinas capaces del procesamiento de dígitos alfanuméricos, esto es, tanto caracteres como números; y por otro lado, el alumbramiento de un nuevo campo de estudio llamado Lingüística Computacional, que era la aplicación de la ciencia de la

computación a la estructura y significado del lenguaje, dirigido básicamente por Noam Chomsky.

A mitad de la década de los sesenta M.E. Stevens (1) realizó una disertación donde revisó los criterios que aplicaban los ordenadores a la tarea de indizar, y definió la indización automática como el uso de máquinas para extraer o asignar términos de indización sin intervención humana una vez que se han establecido programas o normas relativas al procedimiento. En estos primeros momentos el acercamiento a la Ia se hacía desde dos concepciones distintas pero en ocasiones complementarias, es decir, hallamos los métodos no lingüísticos que agrupa substancialmente, a los estadísticos, la atribución de pesos, los probrabilísticos, y los de clustering, y por otro lado, aquellos en los que se ejecutaban ciertos análisis lingüísticos de los textos, que con el paso del tiempo han estado más presentes en este tipo de sistemas debido al procesamiento del lenguaje natural (PLN), disciplina de la cual creemos necesario incorporar para la Ia, los avances dados en los distintos niveles del análisis lingüístico.

2 Métodos no lingüísticos

Paulatinamente fueron florecieron formas de indización automáticas que contribuyeron a alimentar la idea de que las técnicas tradicionales de indización iban a cambiar. En la mayoría de los casos, se quedaban en meras experimentaciones o cabía la posibilidad de que las aplicaran en centros de documentación bien de organismos oficiales o privados, donde habían sido desarrolladas, pero rara vez tenían la intención de ser comercializadas.

A continuación se mencionan brevemente algunos ensayos en Ia que tuvieron como base los criterios no lingüísticos, y a través de ellos se puede obtener una visión global acerca de los trabajos emprendidos en las últimas cuatro décadas.

2.1 Frecuencia. H. P. Luhn (2) fue el primero en sugerir que la frecuencia de aparición de los términos en una colección tiene que ver con la utilidad de éstos para la indización. Los términos de frecuencia muy alta (aquellos que se manifiestan en bastantes documentos) serían demasiado generales y producirían menos precisión en una búsqueda; mientras que aquellos de frecuencia muy baja (los adjudicados a muy pocos documentos) serían muy específicos y provocarían una baja exhaustividad. Para Luhn los mejores términos eran los que tenían una frecuencia media, es decir los que no se presentaban ni en muchos ni pocos documentos (3). F.J. Damerau pensó que los métodos propuestos hasta ese momento para reemplazar el esfuerzo manual (seleccionando frases nominales, usando listas de autoridades seleccionadas previamente de forma manual, y extrayendo vocablos no comunes) no eran lo suficientemente eficaces por lo que propuso, la elección de una colección grande de documentos sobre un tema específico para acumular la frecuencia de aparición de las palabras y de estas estadísticas se obtendría para cada vocablo su frecuencia relativa. Posteriormente, para indizar un documento concreto hallaría la frecuencia de aparición de las palabras en ese documento y se compararía con la frecuencia esperada (es decir, con la frecuencia obtenida de analizar la colección sobre el tema específico), y se seleccionarían como términos de indización aquellos cuya reiteración fuera estadísticamente más significativa que la esperada (4).

2.2 Probabilidad. Casi paralelamente a la ejecución de estos proyectos, se fueron conformando otra serie de experiencias encaminadas a examinar varios de los sistemas ya existentes con el propósito de predecir los posibles términos de indización. Un ejemplo es el acometido por Victor Rosenberg (5) que tras la evaluación de algunos sistemas confeccionó una lista que contenía términos de indización clasificados según estimaciones de probabilidad. Consideraba que los datos sobre la co-aparición de términos en una colección ya indizada, podría ser útil en la representación de nuevos documentos. Después, se servía de un procedimiento automático para cosechar una enumeración de los vocablos asociados a cada término desde un vocabulario restringido de descriptores. Cuando el indizador, por medios convencionales elegía los términos asociados para representar el contenido de los documentos tenía la posibilidad de ayudarse de la lista confeccionada previamente, en la cual, se mostraban términos que podían ser "sugerencias" con el fin de que algunos términos no pasaran desapercibidos.

En definitiva, Rosenberg sugirió que este proceso se podría considerar como un primer paso para el desarrollo de un sistema de indización asistido por ordenador, puesto que una organización interactiva permitiría al profesional recomendar términos de indización. Además, un sistema de este tipo gozaría de una doble ventaja: ayudaría a eliminar errores evitando la omisión de términos importantes, e incluso darían flexibilidad al indizador para hallar nuevos conceptos yendo más allá de las recomendaciones del programa.

2.3 Análisis de clases de palabras. Clustering. En otros trabajos como por ejemplo en (6) se estudiaron las apariciones de las palabras con la finalidad de establecer normas formales para identificar aquellos vocablos capaces de transmitir el tema de un documento y por tanto, serían los más adecuados para emplearlos como términos de indización. Para ello, tras el examen de las palabras, distinguían las que proporcionaban información temática de las que no, determinando su agrupamiento (clustering) con un análisis estadístico.

Lo mencionado en los últimos párrafos se puede resumir con palabras de Susan Artandi (7) cuando declara que en contraste con la indización manual, en la automática es un algoritmo el que toma la posición del indizador y se aplica repetidamente a cada documento. El algoritmo examina los textos como una secuencia de símbolos, pudiendo establecer las palabras del texto por la identificación de series de caracteres separadas por espacios. Ahora bien, dado que las limitaciones para incorporar algoritmos que fueran capaces de interpretar los textos era extremadamente limitada, y no se podía consecuentemente simular las decisiones intuitivas del indizador humano, los métodos de indización automática se basaban en la capacidad de la máquina para reconocer signos y secuencias de signos. Por tanto, los sistemas ideados trataron de extraer del texto la frecuencia de aparición de vocablos, partes de palabras o frases, así como la co-aparición y posición relativa en las oraciones. Y el producto final no era más que una lista de unidades lingüísticas extraídas del texto y reorganizada de varias maneras.

2.4 Modelo del valor de discriminación. A mitad de los setenta un grupo de investigadores liderados por Gerard Salton (8) presentaron una nueva técnica denominada el valor de discriminación, por la que clasificaban los vocablos de un texto según la capacidad de éstos para discriminar unos documentos de otros en una colección, es decir, el valor de un término depende de cómo varía la separación media

entre los documentos cuando a un término se le fijaba una identificación de contenido. Por tanto, las mejores palabras son aquellas que consiguen la mayor distancia. El análisis del valor de discriminación consignaba una función específica en el análisis de contenido a las palabras simples, a las yuxtapuestas, a las frases, y a grupos de palabras.

Consideraban además los autores que si los términos para identificar un documento eran más de tres, se podía recurrir al vector espacial para representar una colección. Partiendo de aquí, idearon un sistema de indización, conocido como el *modelo de valor de discriminación*, que atribuye el peso o valor más alto a aquellos términos que causan la máxima separación posible entre los documentos de una colección. El valor de discriminación de un término lo definían como la medida de los cambios en la separación espacial, que se manifiesta cuando una palabra cualquiera es asignada a una colección como término de indización para representar mejor las diferencias que pueda haber entre los documentos. Precisamente, la asignación disminuye la densidad espacial de éstos, y al contrario, un discriminador pobre incrementa la densidad del espacio. De este modo, si calculaban primero las densidades espaciales y se las atribuían a cada término, era posible especificar los términos en orden decreciente por sus valores de discriminación.

2.5 Relevancia de los términos. En algunos de los sistemas proyectados en la segunda mitad de los años setenta se incorporó al ya estudiado cálculo de frecuencia las propiedades de relevancia de los términos. Esta teoría de la relevancia de un término (3) introdujo distinciones entre las apariciones de éstos en un documento relevante, y su presencia en un documento no relevante. De esta distinción hicieron ministerio tanto el valor de precisión basado en consideraciones probabilísticas, como el valor de utilidad de los términos. No obstante, para disfrutar de una visión más amplia de las importantes aportaciones proporcionadas por G. Salton, en más de tres décadas de investigación, tanto en el análisis automático de la información como en su almacenamiento y recuperación ver (9). Además, algunos de estos aspectos se han recogido en trabajos publicados en español (10) y (11).

2.6 Imitación de la indización humana. A principios de los ochenta se propusieron algunos métodos que trataban de imitar a los indizadores humanos fundamentándose de nuevo en la probabilística. Estas propuestas teóricas iban encaminadas a que el sistema aprendiera qué términos atribuiría un indizador humano a un determinado documento. Para ello se requería un conjunto de documentos previamente indizados por profesionales, y a la hora de operar el programa manejaría este grupo de términos para calcular normas de asociación o coeficientes de adhesión entre éstos y los términos aparecidos en los nuevos documentos a indizar (12).

2.7 INDEXD. En la Universidad de Louisiana investigadores de varios Departamentos han desarrollado este sistema que está basado en métodos estadísticos pero apoyado sin embargo, en cierto componente lingüístico que se verá ampliado, según los autores, en futuros trabajos. Se trata de dos programas complementarios llamados el primero INDEX, capaz de localizar frases repetidas en un documento, proporcionar información estadística a cerca de ellas y clasificarlas según su valor como frases de indización. El otro programa es INDEXD extensión del primero donde la lista de palabras vacías de

aquél pasa a ser un diccionario de raíces de vocablos con capacidad para preasignar pesos a las palabras por su capacidad para distinguir entre un área u otra, con capacidad para descomponer las palabras en sus raíces, y con la opción de validar el léxico (13).

En definitiva, los métodos de indización automática no lingüísticos que acabamos de presentar, se pueden resumir señalando que de un mero recuento de las palabras con unos umbrales, se pasó al estudio de la probabilidad de aparición de los términos de indización y al de las palabras para averiguar su capacidad para comunicar el tema de los documentos. A partir de la mitad de los setenta se puede observar una mayor complejidad en los intentos de obtención de la Ia. En primer lugar, con las propuestas de G. Salton concretadas en el modelo de valor de discriminación y, posteriormente, con el estudio de la relevancia de los términos ya entramos en los ochenta, donde en sus primeros años se pretendió una vez más, desplegar métodos para imitar a los indizadores humanos, que fueron conviviendo a mitad y finales de esta década, con sistemas basados cada vez más en los análisis lingüísticos. Por último, cabe reseñar que si se desea una visión más completa, en cuanto al número de ensayos basados en métodos no lingüísticos, se puede consultar el trabajo de la brasileña Bastos Vieira (14)

Seguidamente, vamos a describir dos sistemas que no se apoyan tanto en cálculos numéricos sino más bien en la utilización de tesauros, pero como veremos, son procedimientos diseñados para dominios específicos del conocimiento como es el caso de la Medicina y la documentación Aeroespacial.

2.8 SAPHIRE. Una vez más este sistema procede de una Universidad norteamericana, y más concretamente de Porland, en Oregón. El núcleo de este sistema de indización y recuperación automática para el dominio de la Biomedicina es un algoritmo que toma un texto o una frase de búsqueda y obtiene una lista ordenada de todos los conceptos hallados tras su comparación con un tesauro. Posteriormente, a cada concepto se le atribuye un peso según su frecuencia tanto en la base de datos, como en el documento analizado. El programa consagra un metatesauro[1] perteneciente a la Biblioteca Nacional de Medicina de Estados Unidos, llamado Meta-1, que contiene un gran número de sinónimos lo que le permite detectarlos en el momento del análisis ya sea a nivel de la palabra como por ejemplo "alto" y "elevado" o de conceptos como "hipertensión" y "tensión alta".

Vamos a referirnos en las próximas secuencias a algunas de las pruebas que se realizaron con objeto de evaluar este sistema desde distintas perspectivas. En primer lugar, se indizaron 200 resúmenes tanto por expertos como por el sistema SAPHIRE (tomando para ello títulos y resúmenes). El sistema automático asignó un total de 4552 conceptos, con una media de 22,8 por resumen, por el contrario, los indizadores humanos asignaron un total de 1966 términos con la MeSH con una media de 9,8. El nuevo examen de los resúmenes proporcionó una reevaluación de 535 concesiones inapropiadas por parte de SAPHIRE. Posteriormente, tras seleccionar una serie de preguntas se acometieron búsquedas booleanas con palabras en el título y resumen empleando términos de MeSH.

La conclusión a la que se llegó es que con SAPHIRE se produjo una menor exhaustividad y precisión en la fase de recuperación con respecto a los métodos tradicionales, tanto de indización como de recuperación ejecutados en MEDLINE. Las causas de estas diferencias son debidas, según los autores, a las lagunas de sinonimia del tesauro Meta-1, por lo que consideran que con una mejora substancial en el mismo

repercutirá fructuosamente tanto en la indización como en la posterior recuperación (15).

2.9 Experiencia en la NASA. Hace más de una década buscando una mayor rapidez y reducción de costes en la indización se inició en el Centro de Información Aeroespacial de la NASA un proyecto para diseñar un sistema de indización asistida por ordenador, proyecto, que según sus autores, sigue ampliándose y perfeccionándose cada año. El sistema a grandes rasgos está compuesto por tres módulos. En el primero se realizan diversas funciones entre las que destacan: la identificación de las fuentes que van a ser procesadas, la limitación de las series de palabras del texto (título y resumen), y llamadas al segundo módulo para ejecutar ciertas operaciones. Además, a este primer módulo llegan finalmente los términos de indización propuestos para su validación. En el segundo módulo se efectúa la búsqueda de frases significativas del texto con un máximo de cinco palabras, para lo cual, recurren a la denominada base del conocimiento o red conceptual, que es ya el tercer módulo, y que contiene más de 115.000 entradas que pueden convertirse en términos de indización. Asimismo, desde esta base del conocimiento se establecen las posibles relaciones existentes entre los términos desde el punto de vista jerárquico o incluso, la desambigüación de los mismos. Según los autores, cuando los resúmenes con los que trabaja el sistema están bien realizados el 60 % de los términos propuestos por el sistema se aceptan.

En lo referente al impacto del sistema en el centro de documentación comentan que el número de indizadores en la institución es menor que antes de valerse del sistema debido, no sólo a su incorporación como herramienta de trabajo, sino también porque la consulta del tesauro por parte de los indizadores se hace en línea desde sus puestos de trabajo, así como la validación de términos de modo automático (16).

3 Métodos lingüísticos

A partir de los años cincuenta se comenzó a bregar en el PLN y desde el primer momento, estas investigaciones estuvieron íntimamente relacionadas con disciplinas como la lingüística formal y las ciencias de la computación entre otras. Surgían en estos años, distintos caminos de estudio. Por un lado, ensayos con un objetivo práctico encaminados a la traducción automática, y por otro lado, trabajos teóricos dirigidos por N. Chomsky sobre formalización del lenguaje, y paralelamente a estas dos direcciones, el comienzo de actividades en Inteligencia Artificial que incluían aspectos del procesamiento del lenguaje natural. Posteriormente, a finales de los sesenta, se planteó la necesidad de entrar de lleno en la comprensión del lenguaje natural, que fue sustituida años más tarde por un fuerte avance en el tratamiento de la sintaxis, en términos de formalismos y de algoritmos de análisis. Si bien la teoría lingüística y la práctica computacional pocas veces convergieron, hasta aproximadamente la década de los ochenta (17).

A principios de los sesenta es cuando comienzan a incorporse tímidamente a la indización automática aspectos del PLN ya que, algunos investigadores intuían que la aplicación de medios lingüísticos era necesaria y se podía combinar con los métodos no lingüísticos, hasta entonces utilizados casi de forma exclusiva.

Al igual que en el apartado anterior vamos a presentar algunos sistemas para descubrir el desarrollo producido en la Ia tomando como base principios lingüísticos.

Pero antes de proseguir es necesario realizar la siguiente aclaración. Como es conocido el sistema SMART está fundamentado claramente en principios estadísticos, puesto que no en vano su creador es un defensor a ultranza de estos métodos, pero lo vamos a incluir en este apartado debido a que fue uno de los primeros sistemas que introdujo ciertas consideraciones lingüísticas en cuanto a la morfología de las palabras o sintaxis de las frases.

3.1 Sistema SMART. En 1961 G. Salton emprendió el proyecto SMART (18), que era un sistema de análisis automático y de recuperación de textos, y cabe destacar que en la actualidad se continúa trabajando en él como se puede comprobar en (19), pero en este caso para mejorar la fase de recuperación de documentos. G. Salton intentó diseñar e implementar sobre un ordenador un sistema capaz de procesar documentos de forma automática para posteriormente, atender peticiones de búsqueda. En los sesenta representó lo más avanzado en el análisis de documentos por lo que fue un gran impulso en el intento de sustituir la indización convencional por procedimientos sofisticados mediante ordenadores. La base fundamental de este sistema eran los cálculos estadísticos pero le incorporaron otras herramientas tales como: un método para extraer las raíces de las palabras, un diccionario de sinónimos, un análisis sintáctico, y métodos de comparación de vocablos que hacían posible parangonar los documentos ya analizados con peticiones de búsqueda explotando un léxico para identificar oraciones significativas del texto (20). El glosario estaba compuesto por un gran número de estructuras semánticamente equivalentes, pero construidas de modo diferente desde el punto de vista sintáctico.

La obtención de las raíces y sufijos se lograba aprovechándose de un diccionario compuesto de dos partes: una con raíces de palabras ordenadas alfabéticamente que contenía por ejemplo "ecom-", y otra con sufijos como "ist", "ists", "ical", que se aplicaba para la descomposición de vocablos como "economist", "economists", o "economical". Se introdujo también la posibilidad de que fuera capaz de reconocer como equivalentes una voz bien en singular o plural ("location" y "locations"), las cuales tendrían un único código de identificación. Por otro lado, el glosario de raíces se constituyó para que las palabras con la misma raíz también fueran tratadas como semejantes, como por ejemplo "automaton", "automation" o "automatic" (18).

Gerard Salton realizó una comparación entre la indización automática obtenida por SMART y la manual por medio de MEDLARS (Medical Literature Analisys and Retrieval System) adoptado en la Biblioteca Nacional de Medicina de los Estados Unidos, en la cual, a finales de los sesenta ya se indizaban regularmente unas 2400 revistas científicas. Para comparar los dos sistemas hizo un ensayo con dieciocho preguntas, y de las respuestas ofrecidas se dedujo que con SMART la exhaustividad había sido ligeramente superior, si bien con MEDLARS fue algo mejor la precisión. Pero en cualquier caso, la mejora potencial empleando el sistema SMART osciló entre el 10 y el 15% (19).

A continuación veremos varios sistemas de indización automática donde el componente lingüístico es mucho mayor de lo mostrado hasta ahora, pero antes conviene apuntar que la mayoría de los procesadores del lenguaje natural incorporan en su base lexical, bien un significado conceptual (o profundo) que es el contenido cognoscitivo de las palabras, o un significado superficial, que ofrece las asociaciones entre las palabras o clases de vocablos. Por otro lado, en cuanto al nivel semántico, uno

de los superiores del análisis lingüístico, cabe manifestar que la adquisición de conocimiento semántico de forma sistemática es una tarea verdaderamente no exenta de dificultad, y en los últimos años se han presentado algunos métodos que ayudan a conquistar este conocimiento, aunque la mayoría de éstos utilizan diccionarios on-line como fuente de datos. Otra forma es empleando corpus, puesto que proporcionan el manejo de las palabras, sus asociaciones y los fenómenos del lenguaje (22).

Por estos y otros motivos, algunos investigadores consideran que mientras el disfrute metodologías lingüísticas sea tan complicado computacionalmente, requiera tanto espacio de almacenamiento y la disponibilidad de aplicaciones sea menor que en la estadística, seguirán recomendando ésta última, puesto que ante resultados parecidos se debe elegir la más simple (23).

3.2 Sistema CLARIT. (Computational-Linguistic Approaches to Indexing and Retrieval of Text). Es otro acercamiento a la Ia que trata de solucionar dos problemas tradicionales en este tema: capturar la estructura lingüística de los textos o identificación de los conceptos, y seleccionar aquellos que reflejan el contenido de un documento. En particular, el sistema encuentra sintagmas nominales y los convierte en candidatos para la indización tras un proceso morfológico. Después, éstos se compararán con un tesauro, y se clasifican como términos exactos, generales o nuevos.

Este sistema partiendo de un texto ejecutaba tres pasos: formateado, procesamiento del lenguaje, y filtrado. En el formateado se añaden símbolos de demarcación del texto, como los comienzos y finales de las oraciones y párrafos, algo análogo a preparar el texto para el análisis posterior. El PLN implica dos etapas: el análisis morfológico y el sintáctico junto con una operación opcional de desambigüación lexical. Se trata de encontrar un conjunto de palabras candidatas que en la siguiente etapa de filtrado se conviertan en un grupo de términos ponderados. El filtrado de los términos de indización se consuma en tres pasos: en el primero de ellos a los términos candidatos, producto del procesamiento del lenguaje, se les asocia un valor basado en las características de distribución de las palabras. Posteriormente, se comparan los candidatos con un conjunto de términos ya normalizado. Y en el último estadio, todos los términos se dividen y clasifican en tres categorías: los que coinciden con los del tesauro se retienen como exactos; aquellos que no están entre los exactos se definen como generales; y los que sobrepasan un determinado umbral se conservan como un conjunto de términos nuevos (24).

3.3 Proyecto SIMPR. (Structured Information Management: Processing and Retrieval). Se trata de un prototipo diseñado por un grupo interdisciplinar compuesto primordialmente por lingüistas computacionales, documentalistas e informáticos de Finlandia, Escocia y Alemania respectivamente. En este sistema realizan un análisis del lenguaje valiéndose de una nueva técnica basada en la explotación de contrastes. Lista todas las posibles interpretaciones léxicas y sintácticas de una palabra, y entonces utiliza información interna para contrastar estas interpretaciones, eliminando aquellas que no son las adecuadas para el contexto de la palabra analizada. El fin es desechar todas excepto una, la correcta.

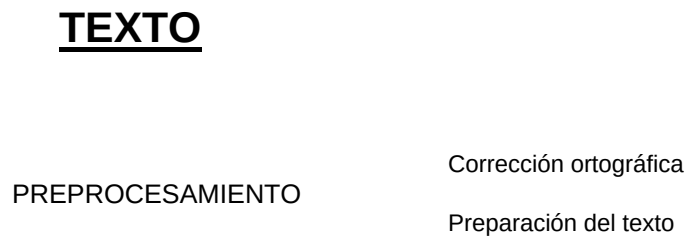
En líneas generales la indización en este sistema se realiza de la siguiente manera. Se efectúa un examen del texto para rechazar aquellas partes que no son indizables. Seguidamente lleva a cabo un análisis de carácter morfo-sintáctico que se descompone

en subanálisis, con el fin de simplificar y clarificar computacionalmente los problemas, pero cada uno de ellos se constituye como elemento individualizado del procedimiento general. Y por último, el resultado del proceso anterior se introduce en el llamado módulo de indización (MIDAS, Módulo de Identificación de Analíticas -términos de indización-). En éste se identifican las partes del texto que son potencialmente útiles para obtener los términos de indización, y estas secuencias de palabras significativas sufren entre otros un proceso de normalización (25).

Finalmente, comentaremos dos experiencias no excesivamente complicadas sobre Ia practicadas por investigadores españoles pero reveladoras porque son los primeros ejercicios en este rumbo perpetrados en España. En (30) se describe someramente un sistema de indización y coordinación de descriptores de modo automático partiendo de los títulos de los documentos. Para lo cual se buscan en el título términos que estuvieran contenidos en una lista autorizada sobre metalurgia. Cuando se tienen los unitérminos (siempre sustantivos) tras un análisis morfológico se procede a buscar en los títulos alguna de estas estructuras sintagmáticas: sustantivo + adjetivo; sustantivo + participio; sustantivo + sustantivo. Cuando se han conseguido dichas estructuras si alguna de las palabras que las constituyen aparecen recogidas en la lista de términos autorizados se convierte automáticamente en descriptor compuesto, y si no viene recogido queda como descriptor el término simple.

El otro trabajo al que nos hemos referido anteriormente es el de Simón Granda y E. de Lema (31), cuyas propuestas para un sistema de indización asistida por ordenador, a grandes rasgos, las dividimos, según nuestro criterio, en siete etapas: una segmentación del texto en unidades inferiores de la oración (frases comprendidas entre los signos de puntuación); verticalización de las frases comprendidas entre los signos de puntuación; eliminación de todas las palabras vacías tras su comparación con un fichero; análisis morfológico del resto de palabras para adjudicar las posibles categorías gramaticales a cada una de ellas, y ordenación alfabética de todos los términos candidatos a descriptores. Posteriormente, un nuevo módulo del programa se encarga de eliminar los términos repetidos y establece con los restantes, una jerarquía de más general a más específica. La última etapa es la validación de los términos tras una fase de postedición.

Desde el comienzo de esta revisión se han realizado sucesivas referencias al procesamiento del lenguaje natural por lo que en la figura 1 se muestran, de un modo esquemático, los principales análisis que puede sufrir un texto tratado de modo automático.



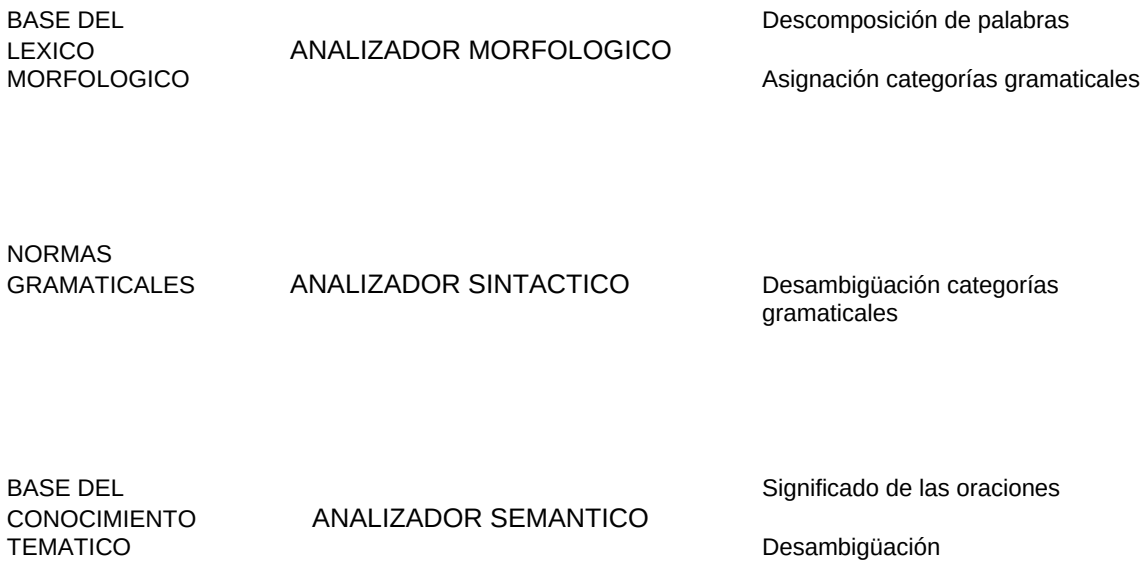


Figura. 1

4 Sistemas de Indización Automática comercializados

En los párrafos siguientes vamos a examinar una serie de programas de Ia disponibles en el mercado francés o en regiones de habla francófona. Como se podrá comprobar SPIRIT (comercializado desde 1980), DARWIN (desde 1987) y ALEHT (desde 1988) son sistemas de análisis, almacenamiento y recuperación de información, y en general, se puede señalar que se apoyan en diversas herramientas de análisis. El morfológico es un instrumento constante mientras que un tratamiento sintáctico o semántico es menos común. Por otro lado, SPIRIT y DARWIN no recurren a un tesauro para obtener los términos de indización lo que sugiere que son sistemas, en principio, más industrializables y por tanto, puede ser más fácil procesar cualquier área del conocimiento. Se observará también la capacidad para actuar en distintas lenguas, así como la posibilidad de realizar la búsqueda de los documentos en la base de datos en lenguaje natural (SPIRIT y ALETH), lo que conlleva a que el usuario no necesite conocer un lenguaje específico de recuperación y por tanto, habrá menos riesgo de error que. al plantear las ecuaciones de búsqueda.

4.1 SPIRIT. (Sistema Probabilista de Indización y de Recuperación de Información Textual). Es un programa de Gestión Documental que permite la adquisición, indización, almacenamiento, búsqueda y difusión de la información. Es el resultado de más de 20 años de investigaciones teóricas y aplicadas realizadas en Francia en el seno de la Comisaría de la Energía Atómica (C.E.A) y la Facultad de Lingüística d'Orsay. La Ia reposa sobre un análisis lingüístico y estadístico. Puede analizar textos jurídicos, científicos, técnicos, comerciales, informáticos, en idiomas como el francés, inglés y alemán. Además, llevan varios años esforzándose para que la interrogación en la base de datos, que se formaliza en lenguaje natural, se pueda realizar en estos mismos idiomas.

Según Orban de Xivry, director de la empresa que comercializa el producto en Benelux (26), los tratamientos lingüísticos permiten eliminar ambigüedades del lenguaje, retener sólo las palabras significativas y sus dependencias contextuales, y alcanzar una representación normalizada del contenido del texto. En cambio, la aplicación estadística permite la clasificación de los documentos por orden decreciente de pertinencia.

El tratamiento lingüístico de un texto se hace en estas fases: recorte de las palabras del mismo; análisis morfológico con la consulta a un diccionario de 500.000 palabras ya estén en mayúsculas o minúsculas; reconocimiento de locuciones por medio de un diccionario; análisis sintáctico por el que se detectan las ambigüedades homográficas y las resuelve; búsqueda de palabras compuestas desde criterios esencialmente sintácticos y algorítmicos; y la normalización de algunas de las palabras que el sistema considere. En cuanto al módulo estadístico apunta que el cálculo del peso de cada concepto, ya sea simple o compuesto, se marca en relación al conjunto de los conceptos contenidos en la base de datos. De este modo, el sistema asigna mayor peso cuanto menos común sea el concepto, y el ejemplo al que alude es que en una base de datos sobre Economía, se atribuirá mayor peso a "país en vías de desarrollo" que a "economía" puesto que efectivamente, esta última palabra estará contenida en casi todos los documentos.

Asimismo, cabe mencionar la capacidad para realizar enlaces de hipertexto a partir de una parte del texto como punto de partida, estableciendo un enlace con todos los que tratan del mismo tema. Para obtener estos resultados, el sistema calcula el grado de proximidad semántica entre el texto de la pregunta y todos los existentes en la base de datos. De tal manera, que para cada documento y página del texto seleccionado se calcula un parámetro de proximidad semántica, dando la opción de clasificar los documentos seleccionados para la visualización en orden jerárquico basado en su pertinencia.

En cuanto a los aspectos tecnológicos hay que apuntar que está disponible para puestos de trabajo Windows 3.1 (autónomo o en red local); en arquitectura cliente/servidor, como servidor Unix, Windows NT y como cliente Windows, Macintosh, Unix y OS2.

4.2 GOLEM. La entidad SIEMENS, en su vertiente de productos informáticos, puso en el mercado un sistema de almacenamiento y recuperación documental llamado Golem. El módulo de este sistema que materializa la Ia se conoce como PASSAT. Como no se ha encontrado información más actualizada sobre el sistema se remite al lector a (27) donde se hace un repaso completo del mismo.

4.3 ALETH. El programa ALETH pertenece a la empresa francesa ERLI, el cual, según J.P. Taravella, invierte varios componentes para llevar a cabo la Ia como un tesauro con términos excluidos y descriptores unitérminos, con una serie de códigos numéricos que ayudan a constituir descriptores compuestos, a establecer relaciones entre distintos términos, y a remitir a los documentos indizados. Igualmente, dispone de un diccionario lingüístico que contiene aspectos sintácticos que incluyen posibles composiciones de verbos, adjetivos o adverbios, mientras que en el plano semántico, existen marcadores del tipo "humano", o "acción". Se puede considerar que estos datos constituyen una importante base de conocimiento.

En cuanto al proceso de indización se distinguen dos fases. La primera se denomina morfo-léxica en donde se intenta detectar qué palabras de las presentes en el texto se encuentran tanto en el tesauro como en el diccionario, aplicando para ello un análisis morfológico que incluye la descomposición de la frase en vocablos, un reconocimiento de las expresiones y la asignación de las categorias gramaticales. En la segunda fase conocida como sintáctico-semántica el objetivo es la búsqueda de las palabras pertinentes adjudicando en este proceso reglas de desambigüación y de normalización. Finalmente, el programa propondrá una serie de términos para que el documentalista los valide (28).

4.4 DARWIN. A principios de los noventa este programa estaba disponible para el francés e inglés. El sistema cuenta con un diccionario gramatical con el que lleva a cabo la desambigüación morfológica, es decir, averiguar si la palabra francesa "couvent" es sustantivo, masculino y singular (convento) o es un verbo en plural (empollan). Estas reglas gramaticales sirven a la vez para desambigüar y hallar los grupos nominales que figuran en el texto analizado. Posteriormente, tras un análisis sintáctico DARWIN extrae las palabras o expresiones significativas, y finalmente, indiza el texto por conceptos o más precisamente, por sintagmas nominales. Además, cuenta con un módulo complementario que permite tratar las abreviaturas o siglas. Por último, cabe apuntar que el sistema de CORA ofrece la posibilidad de interrogar al sistema en lenguaje natural.

4.5 SINTEX Y **ALEXDOC.** Se trata de dos sistemas de indización asistida por ordenador, el primero con una orientación probabilística y el segundo lingüística. Como en el caso del sistema PASSAT no se ha encontrado documentación reciente, por tanto, se remite al lector a (27 , p. 139-151) donde se comentan de modo detallado.

4.6 INDEXICON. Esta herramienta la ha diseñado la entidad norteamericana ICONOVEX. Se trata de un software explotable en estos momentos en su versión 2.0 para el inglés que funciona bajo los procesadores de textos como Word 6.0 para Windows y WordPerfect 6.1. En otoño de 1996 está prevista la versión para Macintosh.

El sistema lee los documentos, localiza los términos y frases significativas y genera un listado con los términos de indización. Para recabar este resultado se sigue un proceso que comienza con la lectura del documento y el etiquetado de los términos indizables, pasando en segundo término, a llevar a efecto un análisis semántico y sintáctico con la finalidad de desambigüar el lenguaje empleado, en el caso de que fuera necesario; permitiendo distinguir por ejemplo, la palabra inglesa "lead" (plomo) de "lead" con el sentido de guiar o influenciar, tomando como base el contexto en el que éstas aparecen. También saca partido de un diccionario compuesto por cincuenta y cinco mil palabras, y un conjunto de normas para determinar las partes de los vocablos y expresiones, y de este modo, analizar cada oración[2]. Si se desea obtener más información acerca de los niveles de aplicación, requisitos técnicos o incluso, una Beta para probar el programa se pueden consultar las páginas html de Iconovex en (29).

Como hemos podido comprobar en los apartados precedentes no existe una única corriente de cómo debe ser un sistema de Ia, ya que por un lado, hay investigadores que propugnan sistemas basados en métodos no lingüísticos, los hay quienes defienden sistemas híbridos, y los que prefieren los métodos lingüísticos. Pero además, tras optar

por uno de estos procedimientos también se observan diferencias importantes a la hora de perseguir el mismo fin. Y con la intención de clarificar las diversas herramientas que destinan los sistemas de indización mencionados hasta el momento, se ha elaborado la siguiente tabla, en donde al referirnos a Análisis lingüísticos englobamos los sistemas que utilizan alguno de los principales niveles, es decir, morfológico, sintáctico y semántico, si bien es cierto que en la mayoría de los casos se emplea el morfológico y el sintáctico para proceder a la desambiguación de las categorías gramaticales propuestas en el morfológico.

HERRAMIENTAS	SISTEMAS
Análisis Lingüístico	SMART, INDEXD, CLARIT, SIMPR, SPIRIT, PASSAT, ALETH, DARWIN, ALEXDOC, INDEXICON, Valle Bracero, Simón Granda
Análisis Estadístico	H.P. Luhn, F.J. Damerau, SMART, INDEXD, SHAPIRE, CLARIT, SPIRIT, PASSAT, SINTEX
Análisis Probabilístico	V. Rosenberg, S.E. Robertson, SPIRIT, SINTEX, Simón Granda
Vocabulario Controlado	Sistema NASA, CLARIT, PASSAT, ALETH, SINTEX, ALEXDOC, Valle Bracero
Fichero Palabras Vacías	Todos
Fichero Expres. Idiomáticas	SPIRIT, ALEXDOC
Fichero de Siglas	DARWIN
Normalización de Términos	SIMPR, SPIRIT, Valle Bracero
Autoreenvío de Conceptos	SMART, SHAPIRE, SINTEX, ALEXDOC,
Validación de Términos	V. Rosenberg, Sistema NASA, ALETH, INDEXICON, SINTEX, ALEXDOC, Simón Granda

5. Otras investigaciones en Indización Automática

La indización asistida o automática que hemos visto hasta ahora, bien tuviera como base los métodos no lingüístico o lingüísticos (o combinando los dos), se ha venido aplicando a documentos con información alfanumérica, esto es, textual. En la última

década, aunque esencialmente a finales de los ochenta, han aflorado nuevos y variados caminos de investigación. Surgieron estudios para conseguir una interpretación automática del sonido, puesto que debido a la expansión de la información multimedia se está incrementando el número de bases de datos que contiene este tipo de información. Y un ejemplo para facilitar tanto el acceso como el tiempo y esfuerzo para seleccionar un sonido, o conjunto de sonidos, de una base de datos, es el ensayo realizado lucrándose de los avances en redes neuronales como se puede ver en (32).

Otra propuesta de Ia totalmente distinta de las indicadas hasta el momento la constituye el prototipo GIPSY (33), que persigue la indización a través de atributos y características geográficas. Este programa lo han desarrollado en la Universidad de Berkeley profesores de Biblioteconomía y Documentación e informática respectivamente. Mediante un algoritmo se extraen palabras y frases que contienen nombres de lugares geográficos o características de éstos acudiendo a un tesauro. También se utiliza un léxico con información espacial como por ejemplo "adyacente a la costa" o "entre el río y la carretera". Posteriormente, en una segunda etapa se busca por el texto información relacionada con las palabras y frases extraídas del tipo: nombre, tamaño y localización de ciudades, estados, etc. o nombre y localización de especies en peligro de extinción. Por otro lado, el programa también debe identificar las localizaciones espaciales más cercanas a los términos geográficos que extrajo el algoritmo.

En definitiva, se trata de un sistema que debe albergar una gran base del conocimiento cuya finalidad última es poder recuperar, del mismo modo con el que se han indizado los documentos, la información contenida en la base de datos, es decir, con nombres o descripciones geográficas. El sistema pretende atender a un grupo amplio de usuarios que busca acceso a las colecciones de documentos que contengan una orientación geográfica. Entre los usuarios destacan gestores de recursos naturales, cuyas peticiones pueden ser información pertinente sobre áreas específicas, o científicos que necesitan localizar publicaciones que traten sobre ciertas zonas. Por último, cabe señalar que este trabajo forma parte de un proyecto más amplio que lleva desarrollándose desde principios de los noventa denominado Sequoia 2000, del que se puede procurar información en (34).

Por otro lado, se han iniciado trabajos para analizar el contenido de imágenes y gráficos, ya que tradicionalmente los sistemas de recuperación de información en entornos multimedia han cuajado una distinción funcional entre los datos gráficos y los textuales, recurriendo a éstos últimos en las operaciones de recuperación. Y ahora con el uso de datos gráficos o imágenes se genera un nuevo contexto para la aplicación de la indización, la cual tendría su núcleo en un sistema capaz de analizar una imagen, localizar las formas que están asociadas a estructuras de interés y describirlas, así como evaluar sus propiedades (35).

En (36) tratando de facilitar el acceso a las bases de datos de imágenes han desarrollado un sistema de Ia para esta clase de información. Consideran que un aspecto clave en el proceso de indización es tener presente la composición de los objetos, las diferentes interpretaciones y los niveles de reconocimiento. En el sistema implementado, lo primero que se acomete es la definición del dominio de aplicación, es decir, la especificación de los tipos de objetos que se van a tratar, bien simples (definidos en términos de las características de los elementos -bitmap o gráfico-, al punto que posibles relaciones entre estos elementos) o bien complejos (definidos por

normas donde se especifican las posibles composiciones). Sería como establecer el área o materia, y una vez alimentado el sistema ya está en condiciones de iniciar la segunda etapa que es el proceso de análisis.

El análisis de imágenes puede ser una tarea enmarañada y consumir bastante tiempo, cuando son imágenes compuestas íntegramente por bitmaps, por lo que se divide esta etapa en dos subfases. En una se lleva a cabo un análisis de bajo nivel en el cual, se capta una imagen compuesta por un bitmap o un conjunto de primitivas gráficas para reconocer los objetos básicos y sus posiciones relativas. Y la otra subfase, llamada de alto nivel, comienza desde los objetos básicos reconocidos en la subfase anterior, y aplica un conjunto de normas para el reconocimiento recursivo de los objetos más complejos contenidos en la imagen. En definitiva, se logra una interpretación de la imagen en términos de los objetos que contiene y su grado de reconocimiento.

En la tercera etapa se generan las estructuras que permitirán el acceso a esa imagen desde de su contenido, en el momento de la recuperación con un lenguaje de interrogación, que permite al usuario expresar condiciones sobre la representación simbólica de las imágenes que se van a recuperar. Las cláusulas convenidas son de este tipo:

ENCONTRAR EN DOMINIO X
QUE CONTENGA
OBJETO (O_{IS})
AND
OBJETO (O_I)

donde DOMINIO es el área o materia de los documentos o simplemente el nombre que recibe un grupo de documentos; QUE CONTENGA, significa que los documentos a buscar deben tener los objetos previamente indizados como (O_{IS}) y (O_I) sirviéndose para ello de los operadores booleanos.

Sin lugar a dudas, una cuestión esencial en el ámbito de las bases de datos de imágenes es obtener un almacenamiento y recuperación eficientes, pero no menos importantes son los escollos que debe salvar este tipo de sistemas, cuando se enfrenta a los problemas que derivan de la dificultad de definir e interpretar exactamente el contenido de las imágenes. Estas pueden ser muy ricas en aspectos semánticos, lo que conlleva a distintas interpretaciones según las perspectivas de la persona. Además, por otro lado, también es complicado determinar y representar las relaciones comunes entre los objetos ya que forman estructuras que varían enormemente de una imagen a otra.

6 Conclusiones

1. La indización automática es una técnica interdisciplinaria donde intervienen entre otras, la Lingüística, la Informática, la Estadística y la Documentación, pero a la vez no exenta de polémica debido a la existencia de profesionales e investigadores en Biblioteconomía y Documentación que defienden su desarrollo, mientras que otros niegan su capacidad para indizar convenientemente los documentos.

2. Las primeras investigaciones en este sentido se ejecutaron, a finales de los cincuenta y durante los sesenta, fundamentándose substancialmente en principios estadísticos y probabilísticos, para dejar paso en los setenta a métodos más complejos dominados por

la corriente de G. Salton con el Modelo de valor de discriminación y la Relevancia de los términos. En la década de los ochenta los sistemas propuestos muestran una mayor presencia de criterios lingüísticos a la hora de efectuar el análisis de los textos, si bien en algunas ocasiones se conjugan los dos métodos dando como resultado sistemas híbridos. Y cuando aún no están completamente consolidados ni reconocidas las capacidades -de forma generalizada- de los sistemas de indización automática o asistida por ordenador que procesan información alfanumérica, se iniciaron a finales de los ochenta, pero principalmente durante los noventa, investigaciones encaminadas a indizar de modo automático información multimedia (imágenes y sonido).

3. Los sistemas comercializados de Ia no acometen la exclusiva misión del análisis de los documentos de cara a la indización de los mismos, sino son gestores documentales, que además del procesamiento tienen la capacidad del almacenamieto y recuperación de los documentos.

4. Por lo observado en este estudio no existe una única teoría de los principios y herramientas que se deben adoptar para diseñar e implementar un sistema de Ia.

5. Se percibe una escasa presencia de investigadores del área de Biblioteconomía y Documentación en los intentos por lograr un sistema de indización automática, donde por el contrario, destacan informáticos, lingüístas computacionales, o incluso propuestas procedentes de investigadores de otras ciencias. Asimismo, existe una exigua tradición de exploración de esta técnica en España.

Notas

[1] Meta-1 es un metatesauro producto de un proyecto emprendido en la Biblioteca Nacional de Medicina de Estados Unidos a partir de 1986 con la finalidad de obtener una herramienta que enlazara un gran número de vocabulario médico donde se unificaron vocabularios como el MeSH (manejado para indizar MEDLINE), DSM-III (American Psychiatry Association), SNOMED (American College of Pathologists), ICD-9 (World Health Organization), y LCSH (Library of Congress). La versión empleada en SHAPIRE contiene 28.423 conceptos, 78.244 sinónimos, y 28.603 raíces de palabras.

[2] En realidad se trata de un simple programa que extrae palabras o frases del texto utilizando para ello distintos niveles lingüísticos y las lista por orden alfabético, pero se ha hecho referencia a este programa principalmente, porque el lector puede agenciarse una Beta en la dirección que se señala, y de este modo, probar el programa con textos en inglés.

Bibliografía

1. STEVENS, M.E. Automatic indexing: a state of the art report, Monograph 91, National Bureau of Standars, Washington, D.C., 1965

2. LUHN, H.P. A statistical approach to mechanized enconding and searching of literary information. *IBM Journal of Research and Development,* 1975, 1 (4), p. 309-317
3. SALTON, G., H.WU, C.T.YU. The measurement of term importance in automatic indexing. *Journal of the American Society for Information Science,* 1982, may, p. 175-186
4. DAMERAU, F.J. An experiment in automatic indexing. *American Documentation*, 1965, 16 (4), p. 283-289
5. ROSENBERG, V. A study of statistical measures for predicting terms used to index documents. *Journal of the American Society for Information Science,* 1971, 22 (1) p. 41-50
6. BOOKSTEIN, A., D.R. SWANSON. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 1974, 25 (5), p. 312-318
7. ARTANDI, S. Machine indexing: linguistic and semiotic implications. *Journal of the American Society for Information Science*, July-August, 1976, p. 235-239
8. SALTON, G., C.S.YANG, C.T.YU. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science,* 1975, 26 (1), p. 33-44
9. SALTON, G. Automatic text processing: The transformation, analysis, and retrieval of information by computer, Massachussets: Addison-Wesley, 1989
10. CODINA, LL. Teoría de recuperación de información: modelos fundamentales y aplicaciones a la gestión documental. *Information World en Español,* 38, octubre, 1995, p. 18-22
11. MOYA ANEGÓN, F. Los sistemas integrados de gestión bibliotecaria: Estructuras de datos y recuperación de información, Madrid: ANABAD, 1995, p. 156-208
12. ROBERTSON, S.E., P.HARDING. Probabilistic automatic indexing by learning from human indexers. *Journal of Documentation*, 1984, 40 (4), p. 264-270
13. JONES, L.P. (et al.). INDEX: The statistical basis for an automatic conceptual phrase-indexing system. *Journal of the American Society for Information Science*, 1990, 41 (2), p. 87-97
14. BASTOS VIEIRA, S. Indexaçao automática e manual: Revisao de literatura. *Ciência da Informao*, 1988, 17 (1), p. 43-57
15. HERSH, W.R. (et al.). A comparison of retrieval effectivenes for three methods of indexing medical literature. *The American Journal of the Medical Sciences*, 1992, 303 (5), 292-300
16. SILVESTER, J.P, M.T. GENUARDI, P.H. KLINGBIEL. Machine-aided indexing at NASA. *Information Processing & Management,* 1994, 30 (5), p. 631-645
17. VERDEJO MAILLO, M.F. Comprensión del lenguaje natural: Avances, aplicaciones y tendencias. *Procesamiento del lenguaje natural: Fundamentos y aplicaciones*, 1994, p. 5-29
18. SALTON, G. The SMART system 1961-1976: Experiments in dynamic document processing. *Encyclopedia of Library and Information Science,* vol. 28, 1980, p. 1-28
19. BUCKLEY, C. J. ALLAND, G. SALTON. Automatic routing and retrieval using SMART: TREC-2. *Information Processing & Management*, 1995, 31 (3), p. 315-326
20. SALTON, G. The evaluation of automatic retrieval procedures. Selected test results using theSMART system. *American Documentation*, 1965, 16 (3), p. 209-222

21. SALTON, G. A comparison between manual and automatic indexing methods. *American Documentation*, 1969, 20 (1), p. 61-71
22. VELARDI, P. How to encode semantic knowledge: A method for meaning representation and computer-aide acquisition. *Association for Computational Linguistics*, 1991, 17 (2), p. 153-170
23. SALTON, G. (et al.). On the application of syntactic methodologies in automatic text analysis. *Information Processing & Management*, 1990, 26 (1), p. 73-92
24. EVANS, D.A. Automatic indexing of abstracts via natural-language processing using a simple thesaurus. *Med Decis Making*, 1991, 11, p. 108-115
25. KARETNYK, D. Knolewledge-based indexing of morpho-syntactically analysed language. *Expert Systems for Information Management*, 1991, 4 (1), p. 1-29
26. XIVRY, O. Le traitement de l'information textuelle utilisation du systeme "SPIRIT": (Système Probabiliste d'indexation et de Recherche d'Informations Textuelles). *Cahiers de la Documentacion*, 1, 1993, p. 15-23
27. SLYPE, G. Los lenguajes de indización: Concepción, construcción y utilización en los sistemas documentales. Madrid: F.G.S.R., 1991, p. 129-133
28. TARAVELLA, J.P. L'indexation automatique en France: Etat de la recherche, problèmes rencontrés et analyse de produits disponibles. *Mémoire présenté pour le DESS*. Institut d'Etudes Politiques de Paris, 1990
29. URL: http://www.iconovex.com. (19-6-1996)
30. VALLE BRACERO, A., J.A. FERNÁNDEZ GARCÍA. Automatización de la indización y coordinación de descriptores. *Revista Española de Documentación Científica*, 1983, 6 (1), p. 9- 16
31. SIMÓN GRANDA, J., E. de LEMA GARZÓN. Primeras experiencias sobre el análisis de textos en castellano aplicado a la indexación automática de información. *Terceras Jornadas Españolas de Documentación Automatizada*, 1990, p. 1255-1270
32. FEITE, B. S. GUNZEL. Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 1994, 18 (3), p. 53-65
33. GYLE WOODRUFF, A. C. PLAUNT. GIPSY: Automated geographic of text documents. *Journal of the American Society for Information Science*, 1994, 45 (9), p. 645-655
34. Larson, R.R. (et al.). The Sequoia 2000 electronic repository.
http://bliss.SIMS.Berkeley.EDU:80/papers/decpaper/decpaper.html (16-5-1996)
35. BORDOGNA, G. (et al.). Pictorial indexing for an integrated pictorial and textual IR environment. *Journal of the Information Science*, 1990, 16, p. 165-173
36. RABITTI, F., P. SAVINO. Automatic image indexation to support content-based retrieval. *Information Processing & Management*, 1992, 28 (5), p. 547-565