Revista Interamericana de Bibliotecología Medellín Vol. 19 No. 1. Enero-Junio de 1996

TENDENCIAS EN LOS SISTEMAS DE INDIZACION AUTOMATICA: ESTUDIO EVOLUTIVO

Isidoro Gil Leyva*
José Vicente Rodríguez Muñoz*

RESUMEN

Se presenta una evolución de la indización automática desde que se inician las primeras investigaciones, a finales de los años cincuenta hasta la actualidad. Los sistemas estudiados se han dividido en aquellos que utilizan métodos estadísticos, los cuales se han descrito mediante términos que identifican sus características, y los que emplean criterios lingüísticos. Se examinan entre otros, sistemas tales como SMART, INDEXD, FASIT o CLARIT. Por último, se apuntan algunas de las tendencias más recientes en indización automática, que ya no tienen como base sólo la información alfanumérica sino que tratan de indizar documentos con información multimedia.

INTRODUCCION

A finales de los 50 y 60 proliferaron las investigaciones sobre el tratamiento de la información, siendo la llamada «explosión de información» uno de los principales motivos que alentaban estos estudios.

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

* Depto. Información y Documentación Universidad de Murcia

repetitivas. De esta forma se evitaba 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, producto de 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 principalmente por Noam Chomsky[1].

Ya en 1965 Stevens 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»[2].

Generalizando podemos dividir los métodos de indización automática, en estas primeras etapas, en estadísticos y aquellos en los que se ejecutaba un cierto análisis lingüístico de los textos.

METODOS ESTADISTICOS Y PROBABILISTICOS

Luhn[3] 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. Para éste 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 asignados 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[4].

Algunos ensayos en indización automática

Florecieron paulatinamente 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 través de algunos experimentos que tuvieron como base los criterios estadísticos se puede dar una visión de la evolución de estos sistemas.

Frecuencia. F.J. Damerau[5] pensaba que los métodos propuestos hasta ese momento para reemplazar el esfuerzo manual (a través de seleccionar frases nominales, el uso de listas de autoridades seleccionadas previamente de forma manual, así como la extracción de 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 frecuencia esperada.

Sistema SMART. En 1961 se inició el proyecto SMART[6], que era un sistema de análisis automático y de recuperación de textos. Intentaba diseñar e implementar sobre un ordenador un sistema capaz de procesar documentos de forma automática y posteriormente, atender peticiones de búsqueda. En aquellos años representó lo más avanzado en el análisis de documentos. Fue un gran esfuerzo para sustituir la indización convencional por procedimientos sofisticados mediante ordenadores.

SMART aunque no era puramente estadístico incorporó un procedimiento para hallar los coeficientes de semejanza basado en las co-apariciones de conceptos en los textos; además, se llevó a cabo mediante cálculos el emparejamiento de frases (del documento y del sistema), utilizando un diccionario elaborado previamente para identificar las frases significativas[7].

Probabilidad. Casi paralelamente a la ejecución de estos proyectos, se fueron conformando una serie de experiencias encaminadas a examinar varios de los sistemas ya existentes, que se fundamentaban en las frecuencias de las palabras, con el propósito de predecir los posibles términos de indización. Un ejemplo es el estudio que acometió Victor Rosenberg[8] que tras la evaluación de algunos sistemas, confeccionó una lista que contenía términos de indización clasificados según estimaciones de probabilidad. Elaboró esta lista porque 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 obtener 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 en cada etapa. Además, un sistema de este tipo gozaría de una doble ventaja: ayudaría a eliminar errores evitando la omisión de términos importantes, e incluso daría flexibilidad al indizador para hallar nuevos conceptos yendo más allá de las recomendaciones del programa.

Análisis de clases de palabras. Clustering. En otros trabajos[9] se estudiaron las apariciones de las palabras con la finalidad de establecer normas formales, y así 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) a través de un análisis estadístico.

De lo visto hasta ahora se puede señalar[10] que, en contraste con la indización manual, la automática es un algoritmo 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 identifica-

ción 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.

No obstante, no hay la menor duda de que la labor acometida por los indizadores es bastante más compleja que entresacar vocablos del texto para que representen su contenido. En Salton[11] se señala que la mayor parte de las técnicas estadísticas que se venían empleando requerían más efectividad, aunque realmente, los procedimientos más precisos no eran computacionalmente efectivos.

Modelo del valor de discriminación. En el trabajo mencionado anteriormente de Gerard Salton se presentó una nueva técnica, denominada como el valor de discriminación. Clasificaba 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 asigna una identificación de contenido. Las mejores palabras son aquellas que consiguen la mayor distancia.

El análisis del valor de discriminación era computacionalmente sencillo, asignaba una función específica en el análisis de contenido a las palabras simples, a las yuxtapuestas, a las frases, así como a grupos de palabras.

Consideraban además los autores que si los términos asignados 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 una 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; por tanto, 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.

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[12] introdujo distinciones entre las apariciones de éstos en un documento relevante, y su presencia en un documento no relevante. De esta distinción hicieron uso tanto el valor de precisión basado en consideraciones probabilísticas, como el valor de utilidad de los términos.

Imitación de la indización humana. A principios de los ochenta se diseñaron métodos fundamentados en imitar a los indizadores humanos[13]. Los sistemas automáticos trataban de averiguar qué clase de términos aplicaría un indizador a un determinado documento. Para ello se requería un conjunto de documentos previamente indizado por profesionales, y posteriormente la máquina manejaría este grupo para calcular normas de asociación o coeficientes de adhesión[14].

Programas INDEX - INDEXD- En la Universidad de Louisiana, a mitad de los ochenta, implementaron dos programas llamados INDEX e INDEXD[15] que eran capaces de localizar frases repetidas en un documento y reunir información estadística acerca de ellas, así como clasificarlas según su valor como frases de indización.

INDEXD es una extensión de INDEX que incorpora un diccionario de raíces de vocablos, la ponderación de palabras y una validación del léxico. Este repertorio posibilitaba acometer tanto un análisis estadístico como cierta capacidad para un análisis sintáctico.

La intención de los autores era combinar INDEXD con técnicas de inteligencia artificial que permitieran la inclusión de un tesauro con conocimiento específico acerca de los conceptos empleados en un área determinada.

En definitiva, podemos señalar a modo de reflexión que en los últimos sistemas de indización automática presentados y fundamentados en bases estadísticas, no incorporan grandes cambios con respecto a los ya ideados en la década de los setenta. Esto nos condiciona a pensar que posiblemente resulte difícil avanzar más siguiendo caminos estadísticos. En el siguiente apartado tratamos la implicación de la lingüística en este campo.

METODOS LINGÜISTICOS

A partir de los cincuenta se comenzó a trabajar en el procesamiento del lenguaje natural 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 Chomsky sobre formalización del lenguaje, y paralelamente a estas dos direcciones, el comienzo de actividades en Inteligencia Artificial que incluían aspectos de 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.

Es ya a principios de los sesenta cuando se incorporan a la indización automática aspectos del procesamiento del lenguaje ya que algunos investigadores intuían que la aplicación de medios lingüísticos era necesaria y se podía combinar con los métodos estadísticos, hasta entonces utilizados casi de forma exclusiva.

Antes de introducirnos en este tipo de métodos debemos indicar que, cuando hablamos de criterios lingüísticos, nos referimos a un análisis morfológico, sintáctico

o semántico, por lo que a continuación se señala brevemente en qué consiste cada uno de estos análisis.

Análisis morfológico:

En este tipo de análisis se trata de realizar una segmentación de la palabra ortográfica para obtener la palabra gramatical, y determinar su estructura y propiedades, esto es, en cada palabra, como primer paso, se determina su raíz, considerando para ello posibles composiciones tales como prefijos y/o sufijos.

Un algoritmo reciente que trata de agrupar vocablos derivados de una raíz común bajo una sola lo ha elaborado J. Savoy[16].

Análisis morfosintáctico

Aquí se lleva a cabo una revisión del resultado obtenido por el estudio morfológico. Analiza algunas estructuras tales como los tiempos compuestos de los verbos y las formas comparativas y superlativas de los adjetivos, que son tratadas como palabras separadas durante la etapa anterior; y por otro lado, simplifica el trabajo sintáctico verificando la concordancia en género y número, artículos, nombres, adjetivos, etc. Este análisis lo incluyen algunos investigadores como una parte del morfológico[17].

Análisis sintáctico

Este se realiza para comprobar si las palabras del texto están bien coordinadas y unidas, en definitiva, para averiguar si las oraciones son gramaticalmente correctas. En esta etapa se pretende también resolver otros problemas no solucionados por los análisis anteriores, como por ejemplo la homografía.

Análisis semántico

En este análisis se trata de conocer el significado de una oración, pero dada la ambigüedad del lenguaje natural surgen los problemas de interpretación. Estos

problemas a veces, se podrán resolver a nivel local pero otras, será necesario utilizar información contextual. Esto requerirá que el sistema cuente con un corpus que contenga información detallada del sentido de las palabras.

Con objeto de aproximarnos a la evolución que se ha ido experimentando en el campo de la indización automática, debido a la incorporación de los métodos lingüísticos, examinamos a continuación una serie de sistemas basados en este conjunto de procesos.

Sistema SMART. Como se ha señalado anteriormente, el proyecto SMART constituyó en los años 60-70 uno de los sistemas más avanzados en el análisis de texto de forma automática. Este sistema añadía a las herramientas utilizadas en los cálculos estadísticos, otras tales como: a) un método para extraer las raíces de las palabras inglesas, b) un diccionario de sinónimos, c) un análisis sintáctico y métodos de comparación de vocablos que hacían posible parangonar los documentos ya analizados con peticiones de búsqueda. El diccionario estaba compuesto por un gran número de estructuras semánticamente equivalentes, pero construidas de modo diferente desde el punto de vista sintáctico; un ejemplo puede ser «recuperación de información» y «la recuperación de la información», ambas construcciones tendrían una misma entrada para su identificación en el sistema.

Por último, se le incorporó un método de confrontación de frases, que operaba de forma semejante al procedimiento de análisis sintáctico, puesto que utilizaba un diccionario para identificar oraciones significativas del texto[18].

La obtención de las raíces y sufijos se realizaba por medio de un diccionario compuesto de dos partes: una con raíces de palabras ordenadas alfabéticamente que contenía por ejemplo «econom-», y otra con sufijos como «ist», «ists», «ical», que se aplicaba para la descomposición de palabras como «economist», «economists», o «economical».

Se introdujo también la posibilidad de que fuera capaz de reconocer como equivalentes una palabra bien en singular o plural («location» y «locations»), las cuales tendrían un único código de identificación. Por otro lado, el diccionario de raíces se constituyó para que las palabras con la misma raíz también fueran tratadas como equivalentes, como por ejemplo «automaton», «automation» o «automatic»[19].

Gerard Salton realizó una comparación entre la indización automática obtenida por SMART y la manual por medio del sistema MEDLARS, utilizado en la Biblioteca Nacional de Medicina de los Estados Unidos[20].

MEDLARS (Medical Literature Anlisys and Retrieval System) tenía diversos objetivos aunque el primero era la producción del Index Medicus. En este Centro se indizaban ya, a finales de los sesenta, regularmente unas 2400 revistas científicas por lo que disponían de completos medios de indización manual.

Para comparar los dos sistemas se 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%.

Salton señaló que una buena indización manual dependía de lo completo y rico que fuera el lenguaje de indización empleado, así como de la rigurosidad y exactitud con que se efectuara la operación. En cambio, en el análisis de texto automático era difícil que los términos resultantes no representaran bien el contenido de los documentos. También se observó que era más efectivo analizar el texto completo de un resumen que su título con el fin de obtener mayor exhaustividad.

Borko[21] indicó a mitad de los setenta que, ya una década antes, se reconocían los beneficios potenciales del análisis tanto sintáctico como semántico, aunque indudablemente había investigadores que planteaban sus reservas en cuanto al poder de resolución que podían aportar éstos. Este es el caso de Sparck-Jones[22] defensor de que las descripciones sintácticas sólo podían ser de valor en contextos específicos, pero de todos modos reconoció que en un futuro este tipo de análisis proporcionaría mayores resultados.

Sistema FASIT. A principios de los ochenta los análisis lingüísticos continuaban consumiendo gran cantidad de recursos computacionales porque no estaban lo suficientemente desarrollados. Esto contribuyó a que se ideara el proyecto FASIT[23] que era un sistema de indización automática mediante un análisis sintáctico pero sin ser completo, ya que no utilizaba criterios semánticos. Perseguía la indización a través de tres grandes etapas: la primera era el etiquetado de las palabras de acuerdo

a categorías sintácticas, es decir, se pondría etiqueta a todas las categorías gramaticales. En la segunda, se efectuaba la selección de conceptos basada en un conjunto de pautas predefinidas, y por último, se llevaba a cabo la agrupación de conceptos en clases que reducían los conceptos seleccionados a su raíz, de tal modo que las palabras con la misma base se consideran sinónimas.

Continuando con la evolución habida en los sistemas de indización que empleaban criterios lingüísticos, hay que señalar que en un siguiente paso en el análisis sintáctico se introdujeron nuevos mecanismos que pretendían extraer unidades complejas tales como sintagmas nominales o sintagmas preposicionales para arrancar de éstos las palabras que representan el contenido del texto. Generalmente, estos sistemas a finales de los ochenta, realizaban análisis estructurales operando sobre voluminosos corpus lingüísticos con varios cientos de normas gramaticales, no consiguiendo éstas, a pesar de todo, eliminar las ambigüedades y proporcionar términos de indización adecuados sin apoyarse en el contexto y otras consideraciones semánticas.

Por estos motivos, investigadores dedicados al estudio de estas técnicas siguen pensando que mientras la utilización de metodologías sintácticas sea tan complicada computacionalmente, requiera tanto espacio de almacenamiento y la disponibilidad de aplicaciones sea menor que en la metodología estadística, seguirán recomendando ésta última, puesto que ante resultados parecidos se debe elegir la más simple[24]. Efectivamente, hay informáticos que reconocen que a pesar de haber diseñado programas que tratan la sintaxis y la semántica de frases completas, sólo sirven para contextos limitados, pero incluso en estas situaciones restringidas, los programas son muy complejos.

Algunos autores sostienen que la causa por la cual la indización automática a finales de los ochenta y principios de los noventa, no haya alcanzado las expectativas esperadas, viene dada porque no se han superado entre otros aspectos, la sinonimia y la polisemia que pueden surgir en los textos. Además, podemos achacar la falta de solución de estos problemas a tres grandes factores: el primero, es que el sentido de los términos de indización elegidos es incompleto, puesto que los términos empleados para describir un documento sólo son una fracción de los que utilizarían los usuarios.

El segundo factor, es la falta de un método automático adecuado para resolver la polisemia. Una aproximación común es el uso de vocabulario controlado y la intervención humana para actos de interpretación semántica, aunque esta solución es extremadamente cara y poco efectiva. Otro intento, es la coordinación de unos términos con otros para evitar la ambigüedad en su significado.

El tercer factor es debido a que generalmente, en los sistemas de indización automática, cada tipo de palabras es tratado como independiente de los demás. Así en la equiparación de dos términos que casi siempre aparezcan juntos, en la mayoría de los casos, se toma como si se hubieran encontrado casualmente en el mismo documento. De este modo, en las búsquedas no se aseguraría una concordancia semántica precisa[25].

Sistema ARIOSTO. Este sistema no ha sido diseñado con la finalidad de servir únicamente como método de indización automática, sino que el resultado obtenido de sus distintos análisis puede utilizarse además, para la traducción automática y la definición de hipertextos inteligentes. Este, parte de un grupo de investigadores de universidades italianas y está dedicado principalmente a la adquisición automática de conocimiento semántico desde un corpus.

El núcleo del conocimiento que se adquiere es un grupo de asociaciones de palabras incrementando con marcadores sintácticos y semánticos. Estos datos se extraen a través de cuatro etapas usando técnicas de procesamiento del lenguaje natural (PLN). En este sistema no se realiza un análisis sintáctico profundo porque lo que importa es la detección de las relaciones binarias y ternarias entre las palabras, dado que los resultados de un examen sintáctico en profundidad no son justificables frente a la complejidad y magnitud computacional.

En ARIOSTO las herramientas estadísticas se aplican para extraer información sintáctica de los corpus, procesados en este caso por un analizador simple basado en gramáticas discontinuas (que es capaz de detectar las ya mencionadas relaciones sintácticas binarias y ternarias).

El sistema Ariosto realiza el procesamiento de los textos en estas fases: a) un análisis morfológico, b) una segmentación del texto, c) un análisis sintáctico poco profundo, y d) un etiquetado semántico. Posteriormente el resultado de estos análisis se podrá utilizar en las distintas aplicaciones señaladas anteriormente[26].

Llegados a este punto se puede señalar que muchos 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 palabras.

Por otro lado, la adquisición de conocimiento semántico de forma sistemática es una tarea muy compleja, y en los últimos años se han presentado algunos métodos que ayudan a la obtención de este conocimiento, aunque la mayoría de éstos utilizan diccionarios on-line como fuente de datos. Otra forma es empleando corpus, puesto que proporcionan el uso de las palabras, sus asociaciones, así como fenómenos del lenguaje[27].

Sistema CLARIT. Computacional-Linguist Approaches to Indexing and Retrieval of Text, es otro acercamiento a la indización automática que trata de solucionar los 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 a través de un proceso morfológico. Después, éstos se comparan con un tesauro, y a continuación se clasifican como términos exactos, generales o nuevos.

Este sistema partiendo de un texto realiza tres pasos: formateado, procesamiento del lenguaje, y filtrado.

En el formateado se añaden símbolos de demarcación del texto, así como los comienzos y finales de las oraciones y secciones, algo así como preparar el texto para el análisis posterior.

El PLN implica dos etapas, el análisis morfológico y el sintáctico junto con una operación opcional de desambiguación 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 ejecuta en tres pasos: en el primero de ellos a los términos candidatos, producto del PLN, se les asocia un valor basado en

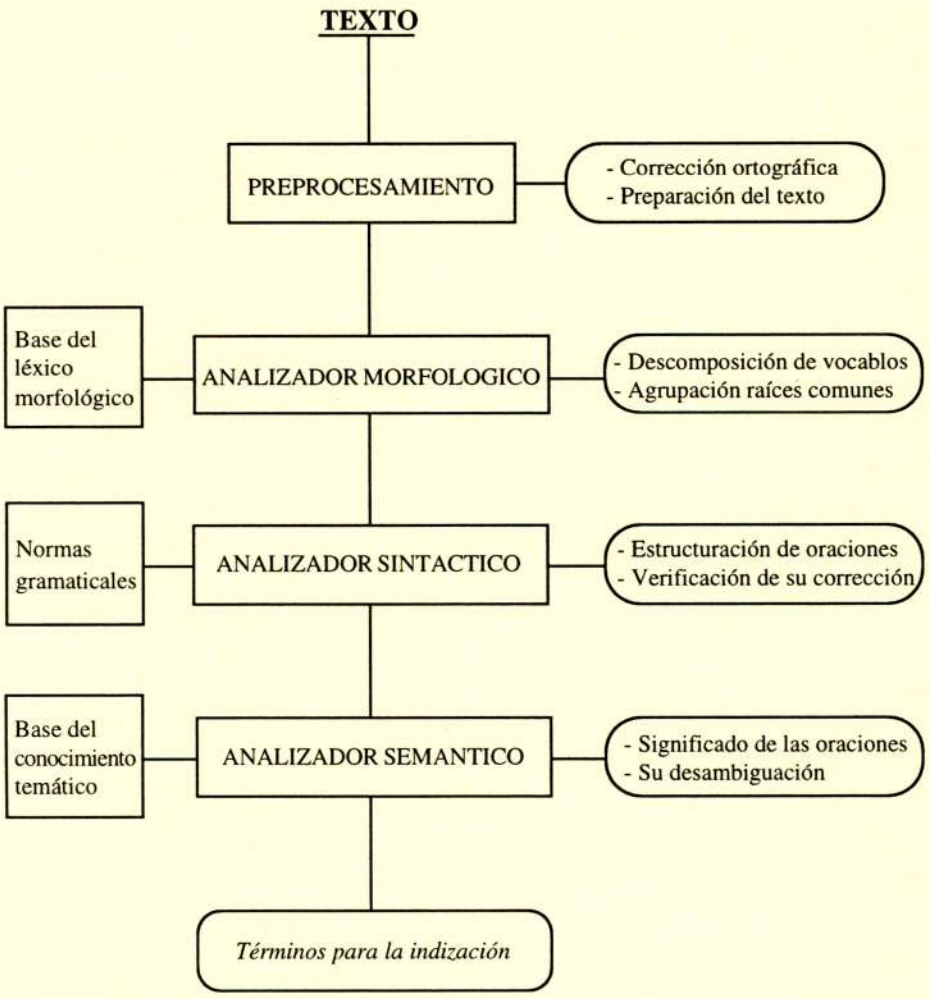
las características de distribución de las palabras. Posteriormente, se comparan los términos candidatos con un conjunto ya normalizado. Y en el último estadio, todos los términos se dividen y clasifican en tres categorías: los que coinciden con los del tesauro se retienen como exacto; aquellos que no están entre los exactos se definen como generales; y los que sobrepasan un determinado umbral se conservan como un conjunto de términos nuevos[28].

Proyecto SIMPR. (Structured Information Management: Processing and Retrieval). Realiza un análisis del lenguaje por medio de una nueva técnica basada en el uso 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. A continuación lleva a cabo un análisis de carácter morfo-sintáctico que se descompone en subanálisis, con el fin de simplificar y clarificar computacionalmente los problemas, y cada uno de los mismos se constituye como elementos individualizados 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 tratado que son potencialmente útiles para obtener términos de indización, y estas secuencias de palabras significativas sufren entre otros un proceso de normalización[29].

A modo de resumen esquemático y para que sirva para ejemplificar genéricamente las operaciones que debe sufrir un texto en las sucesivas etapas constitutivas de los sistemas de indización automática, se presenta el siguiente organigrama:


TEXTO
PREPROCESAMIENTO
- Corrección ortográfica
- Preparación del texto
Base del léxico morfológico
ANALIZADOR MORFOLOGICO
- Descomposición de vocablos
- Agrupación raíces comunes
Normas gramaticales
ANALIZADOR SINTACTICO
- Estructuración de oraciones
- Verificación de su corrección
Base del conocimiento temático
ANALIZADOR SEMANTICO
- Significado de las oraciones
- Su desambiguación
Términos para la indización

OTRAS INVESTIGACIONES EN INDIZACION AUTOMATICA

La indización automática que hemos visto hasta ahora, ya tuviera como base la estadística o la lingüística, se ha venido aplicando a documentos con información alfanumérica, esto es, textual. En la última década, aunque principalmente a finales de los ochenta, han aflorado nuevos y variados caminos de investigación. Así por ejemplo, surgieron estudios para conseguir una interpretación automática del contenido de imágenes y gráficos. Esto es fruto del desarrollo que ha experimentado el procesamiento digital de imágenes (PDI) por la incorporación de soluciones desde áreas tan diversas como la inteligencia artificial, la teoría de ondas y la morfología matemática entre otras. Hoy en día el PDI ocupa un espacio fundamental en campos como las telecomunicaciones o la robótica[30].

En la actualidad, los sistemas de recuperación de información de entornos multimedia realizan una distinción funcional entre los datos gráficos y los textuales, ya que la parte textual ha sido la que tradicionalmente se ha utilizado 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. Esta 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[31].

F. Rabitti y P. Savino[32] tratando de facilitar el acceso a las bases de datos de imágenes, han desarrollado un nuevo sistema de indización automática. Estos consideran que un aspecto clave en el proceso de indización es tener presente la composición de los objetos, así como las diferentes interpretaciones y niveles de reconocimiento.

En el sistema desarrollado, el análisis de las imágenes lo realizan en dos etapas. En la primera, se recuperan los objetos simples comenzando por los elementos pictórico/gráficos básicos; en la segunda fase, se reconocen los objetos complejos por composición de los básicos, y se generan diferentes interpretaciones.

Para estos autores el problema del almacenamiento y recuperación eficientes, en el ámbito de las bases de datos de imágenes, juega un papel muy importante, pero aún lo es más el problema que deriva 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.

Otro campo al que se está dirigiendo la indización automática es al del sonido. Con la ya señalada expansión de la información multimedia se está incrementando el número de bases de datos que contienen sonido. Y un ejemplo para facilitar tanto el acceso como el tiempo y esfuerzo para seleccionar este tipo de información es un ensayo que se ha realizado utilizando redes neuronales[33].

Finalmente cabe señalar que desde la Universidad de California (Berkeley) dos investigadores[34] han tratado de indizar un tipo de información muy concreta, ya que han diseñado un algoritmo que extrae automáticamente términos para facilitar tanto la indización como la recuperación de documentos georeferenciados. Mediante este algoritmo se extraen palabras y frases que contienen nombres de lugares geográficos o características de éstos que serán empleadas como posibles términos de indización.

La finalidad del sistema es atender a un grupo amplio de usuarios que busca un acceso a las colecciones de documentos que contengan una orientación geográfica. Entre los usuarios se destacan gestores de recursos naturales, cuyas peticiones pueden ser por información pertinente sobre áreas específicas, o científicos que necesitan localizar publicaciones que traten sobre ciertas zonas.

CONCLUSIONES

La indización automática desde sus primeros momentos ha pretendido extraer de los textos los conceptos que mejor representaran el contenido de los mismos. Como hemos ido observando, primero se intentó utilizando medios estadísticos, puesto que se consideraba que dependiendo del número de veces que apareciera una palabra en un documento, podría ser empleada o no en la indización.

Posteriormente, surgieron sistemas que se basaron en la probabilidad, en la comparación de frases o en el valor de discriminación de los términos. Estos métodos se estuvieron aplicando principalmente hasta principios de los ochenta. A partir de entonces la mayor parte de los intentos de indización automática adoptaron

como fundamento el estudio lingüístico y más concretamente, el procesamiento del lenguaje natural.

Como se ha podido comprobar a lo largo de este estudio no existe una única corriente de cómo debe ser un sistema de indización automática, ya que por un lado, hay investigadores que propugnan sistemas que no llevan a cabo un análisis sintáctico completo, mientras que otros prefieren no realizar el semántico. En cualquier caso, lo que resulta evidente es que en el análisis lingüístico hay que salvar distintos obstáculos en cada una de sus etapas, pero de forma general los principales escollos surgen a la hora de averiguar el significado semántico de una oración, que es en definitiva lo que proporciona el conocimiento.

Y dado que la imagen y el sonido son percepciones íntimamente ligadas al hombre, que como el texto aportan conocimiento, se convierten en elementos susceptibles de ser tratados en el campo de la indización automática, y se han puesto ejemplos de aplicaciones que así lo demuestran.

Por tanto nos encontramos en un punto en el cual, tras una fuerte evolución del análisis de información textual, todavía seguimos en una etapa bastante primigenia de la «solución final» del problema del procesamiento del lenguaje natural. Pues bien, con todo ello se comienzan a plantear tratamientos con un tipo de información que se ha venido en denominar multimedia, que contempla no sólo el texto sino que incorpora la imagen y el sonido.

Esto nos lleva a pensar que el problema comienza a alcanzar una magnitud que es en la actualidad difícil de medir, pero todo apunta a que el futuro nos traerá sistemas donde no sólo se almacenará información (entendida ésta como multimedia), sino que mediante particulares mecanismos, almacenaremos conocimiento, y los sistemas de recuperación serán transparentes al tipo de consultas que se realicen.

REFERENCIAS BIBLIOGRÁFICAS

1 CHOMSKY, N. Syntactic structures. La Haya: Mouton and Co., 1957

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

3 LUHN, H.P. A statical approach to mechanized encoding and searching of literary information. ¦ En: *IBM Journal of Research and Development.* Vol. 1, no. 4 (1975); p. 309-317

4 SALTON, G., H, WU, C.T. YU. The measurement of term importance in automatic indexing. ¦ En: *Journal of the American Society for Information Science.* (May 1981); p. 175-186

5 DAMERAU, F.J. An experiment in automatic indexing. ¦ En: *American Documentation,* Vol.16, no. 4, (1965); p. 283-289

6 SALTON, G. The SMART system 1961-1976: experiments in dymanic document processing. ¦ En: *Encyclopedia of library and information science,* Vol. 28 (1980); p. 1-28

7 SALTON, G. The evaluation of automatic retrieval procedures. Selected test results using the SMART system. ¦ En: *American Documentation.* Vol.16, no 3 (1965); p. 209-222

8 ROSENBERG, V. A study of statistical measures for predicting terms used to index documents. ¦ En: *Journal of the American Society for Information Science,* (1971); p. 41-50

9 BOOKSTEIN, A., D.R. SWANSON. Probabilistic models for automatic indexing. ¦ En: *Journal of the American Society for Information Science.* Vol. 25, no 5. (1974), p. 312-318

10 ARTANDI, S. Machine indexing: lingüistic and semiotic implications. ¦ En: *Journal of the American Society for Information Science.* (1976); p. 235-239

11 SALTON, G., C.S. YANG, C.T. YU. A theory of term importance in automatic text analysis. ¦ En: *Journal of the American Society for Information Science.* Vol. 26, no 1. (1975); p. 33-44

12 SALTON, G., H. WU, C.T., YU. The measurement of term...op., cit.

[13] ROBERTSON, S.E., P. HARDING. Probabilistic automatic indexing by learning from human indexers. ¦ En: *Journal of Documentation.* Vol. 40, no. 4 (1984); p. 264-270

[14] HARDING, P. Automatic indexing and classification for mechanised information retrieval. London: British Library Research and Development Department, 1982 (informe n° 5723)

[15] JONES, L.P. (et al.). INDEX: The statistical basis for an automatic conceptual phrase-indexing system. ¦ En: *Journal of the American Society for Information Science.* Vol. 41, no 2 (1990); p. 87-97

[16] SAVOY, J. Stemming of french words based on grammatical categories. ¦ En: *Journal of the American Society for Information Science.* Vol. 44 no 1. (1993); p. 1-9

[17] ANTONACCI, F. (et al.). Representation and control strategies for large knoledge domains: A application to NLP. ¦ En: *Applied artificial intelligence.* Vol. 2 (1988); p. 213-249

[18] SALTON, G. The evaluation of automatic...op., cit.

[19] SALTON, G. The SMART system 1961-1976...op., cit.

[20] SALTON, G. A comparison between manual and automatic indexing methods. ¦ En: *American Documentation.* Vol. 20, no 1 (1969); p. 61-71

[21] BORKO, Harold. Indexing concepts and methods. New York: Academic Press, 1978

[22] SPARCK-JONES, K. Automatic indexing: A state of the art review. Cambridge University, 1974

[23] DILLON, M., MCDONALD, L.K. Fully automatic book indexing. ¦ En: *Journal of Documentation.* Vol. 39, no 1 (1983); p. 135-154

[24] SALTON, G. (et al.). On the application of syntactic methodologies in automatic text analisys. ¦ En: *Information processing & management.* Vol. 26, no 1 (1990); p. 73-92

[25] DEERWESTER, S. (et al.). Indexing by latent semantic analysis. ¦ En: *Journal of the American Society for Information Science.* Vol. 41, no 6 (1990); p. 391-407

[26] PAZIENZA, M.T. Extraction of semantic knowledge from text: a goal or a starting point?. *Actas X Congreso SEPLEN.* 1994, p. 1-23

[27] VELARDI, P. How to encode semantic knowledge: a method for meaning representation and computer-aide acquisition. ¦ En: *Association for Computational Linguistics.* Vol 17, no 2 (1991); p. 153-170

[28] EVANS, D.A. Automatic indexing of abstracts via natural- language processing using a simple thesaurus. ¦ En: *Med Decis Making.* No 11 (1991); p. 108-115

[29] KARETNYK, D. Knolewledge-based indexing of morpho-syntactically analysed language. ¦ En: *Expert systems for information management.* Vol 4, no 1 (1991); p. 1-29

[30] RAMIREZ, J. <jrp @esga.es>. Información aparecida en un grupo de noticias de *INTERNET,* junio 1995

[31] BORDOGNA, G. et al. Pictorial indexing for an integrated pictorial and textual IR enviroment. ¦ En: *Journal of information science.* No 16 (1990); p. 165.173.

[32] RABITTI, F., P. SAVINO. Automatic image indexation to support content-based retrieval. ¦ En: *Information processing & management.* Vol. 28, no 5 (1992); p. 547-565

[33] FEITEN, B., S. GUNZEL. Automatic indexing of a sound database using self-organizing neural nets. ¦ En: *Computer music journal.* Vol 18, no 3 (1994); p. 53-65

[34] GYLE WOODRUFF, A., PLAUNT, C. GIPSY: Automated geographic of text documents. ¦ En: *Journal of the American Society for Information Science.* Vol 45, no 9 (1994); p. 645-655