De la indización humana a la indización automática

Isidoro Gil Leiva
José Vicente Rodríguez Muñoz
Universidad de Murcia

0.1. Resumen

Se presentan en primer lugar, algunos aspectos de la indización humana para pasar a continuación a tratar la evolución de la indización automática. Esta abarca desde finales de los años cincuenta hasta la actualidad, con la aplicación en los primeros momentos de métodos estadísticos, y posteriormente, la incorporación de medios lingüísticos. Por último, además de presentarse varios programas de indización automática recientemente aparecidos en el mercado, se apuntan algunas líneas actuales de investigación, que ya no tienen como objeto de análisis la tradicional información alfanumérica sino la cada vez más presente información multimedia.

Palabras clave: Indización automática. Evolución. Líneas de investigación actuales.

0.2. Abstract

The evolution of automatic indexing is traced, starting with a sketch of human indexing characteristics. The history of automatic indexing begins in the fifties, with the application of statistical methods, and, thereafter, linguistic processing. Finally, some automatic indexing packets are presented, and some current research lines are pointed, which overcome the traditional focus into textual information, to take care of multimedia information.

Keywords: Automatic indexing. Evolution. Current research lines.

1. Introducción

La indización es captar y representar el contenido de un documento, para lo cual en el proceso de indización humana se siguen dos pasos: seleccionar aquellos conceptos en lengua natural, que van a representar el contenido del documento y, posteriormente trasladar esos conceptos a un lenguaje controlado, aun-

que últimamente se está tratando de emplear más frecuentemente el lenguaje natural tanto para el almacenamiento como en la recuperación de información.

Al ser la indización una de las operaciones más complejas de realizar en el análisis de la información se puede observar que no siempre los investigadores que han teorizado sobre esta técnica o incluso los mismos indizadores, están completamente de acuerdo en muchos aspectos, y se está lejos de una normalización total en la ejecución de esta tarea.

Algunas de las razones por las cuales es complicado alcanzar la normalización son diversas y muy diferentes, en primer lugar, se ha debatido bastante y se continuará realizando, sobre qué conocimientos debe poseer el indizador, si una formación como documentalista o debe ser un especialista en la materia que indiza. Pero además, surgen claras disparidades (como hemos podido constatar) cuando se les pregunta a los indizadores sobre cuestiones como: ¿cuáles son las partes de los documentos de donde se debe extraer la información?; ¿cuántos son los términos, que de promedio serían los adecuados para representar un documento?; ¿está relacionado el tamaño del documento que se analiza con el número de conceptos que se seleccionan?; ¿qué tiempo se debe dedicar a cada documento?; ¿todos los indizadores cuando realizan esta tarea tienen en cuenta de qué manera solicitará el usuario ese material?. Estas diferencias en el modo de realizar la misma operación hace que sea muy complicado alcanzar una completa normalización.

Otra controversia importante, y a la que se pretendía llegar, es la que comenzó a principios de los años sesenta acerca de la preferencia o no de la indización automática. Entonces había y todavía hay, investigadores y profesionales que consideran que una máquina es incapaz de realizar convenientemente la labor de indizar, ya que no pueden llegar a captar todos los matices conceptuales como puede hacerlo un indizador humano. Otros alegan que para qué emplear la indización automática si los términos que proponen algunos sistemas que existen actualmente los debe validar el indizador humano. Algunos desechan este modo de análisis porque de momento está restringido a áreas del conocimiento muy concretas como puede ser la medicina, aunque como veremos los sistemas actuales se aplican a cualquier tema. También hay quien no descarta la indización automática siempre que sea para aquellos documentos sin demasiada importancia, a los cuales, el indizador humano no puede dedicar su tiempo y esfuerzo.

Dejando de la lado la idoneidad o no del uso de la indización automática, lo que sí es cierto es que se lleva trabajando décadas para conseguir que un algoritmo sea capaz de ejecutar la misma labor que una persona. A continuación vamos a introducir algunas de las líneas por las que se comenzó a investigar en este campo.

2. Indización automática

A finales de los cincuenta y sesenta se produjo un crecimiento exponencial de la información científica disponible en todos los campos del saber, aunque principalmente en las áreas de ciencias. Por estos motivos se fueron ideando sistemas de información cada vez más operativos, así como aumentó el número de investigaciones sobre el tratamiento de la información, con la finalidad última de atender de forma cada vez más eficaz y rápida, las necesidades de información de los científicos.

Por estos años, se fue generalizando paulatinamente la idea de que el ordenador constituía una herramienta muy útil para el análisis de la información y en especial, a través del procesamiento de textos para la indización. 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 y florecieron poco a poco, formas de indización automáticas que contribuyeron a alimentar la idea –en ciertos entornos– que las técnicas tradicionales de indización iban a cambiar. En la mayoría de los casos, estos proyectos se quedaban en meras experimentaciones o cabía la posibilidad que los aplicaran en centros de documentación bien de organismos oficiales o privados, donde habían sido desarrollados, pero rara vez tenían la intención de comercializarlos.

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. Esta se ha venido aplicando tradicionalmente a tres campos principalmente: en la traducción automática, en la recuperación de información, campo más conocido para nosotros y, por último en interfaces hombre-máquina, buscando una mayor comunicación con los sistemas interactivos (1)

Entrando en materia hay que señalar que ya en 1965, M.E. 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)

A lo largo de las últimas décadas, los intentos de diseño e implementación de algoritmos capaces de obtener términos para la indización, una vez analizado un

documento, han venido desde los campos de la Estadística y del Procesamiento del Lenguaje Natural (PLN). Los medios estadísticos fueron los primeros en aplicarse y casi de forma generalizada, desde finales de los años cincuenta hasta hace poco más de una década. En cambio, el acercamiento a la indización automática desde la lingüística computacional comenzó en los sesenta y aún en la actualidad, se continúa trabajando en ello. Por estas razones, nos detendremos más en estos últimos métodos que en los primeros.

2.1 Métodos estadísticos

Luhn 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 (3).

Posteriormente, surgieron sistemas que se basaron en la probabilidad, como un estudio que examinó varios de los sistemas ya existentes con el propósito de predecir posibles términos de indización; otros se fundamentaron en el análisis de las clases de las palabras para averiguar las que proporcionaban información temática de las que no, determinando su agrupamiento (clustering) a través de un cálculo estadístico; y otros sistemas se apoyaron en el valor de discriminación de los términos. Estas distintas direcciones que perseguían proporcionar nuevos elementos en la consecución de la indización automática, se estuvieron aplicando hasta aproximadamente principios de los ochenta.

El uso de medios estadísticos para la indización automática significaba por tanto, que 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 (4).

En general, la mayor parte de las técnicas estadísticas que se emplearon

requerían una efectividad superior, aunque realmente, los procedimientos más precisos no eran computacionalmente rentables. Esto nos induce a creer que posiblemente se había llegado a las cotas más altas de efectividad utilizando los medios estadísticos para lograr una adecuada indización automática. Por estas razones se continuó trabajando en el procesamiento del lenguaje para aplicarlo a la indización. Este será el asunto que trataremos en el siguiente epígrafe.

2.2. Métodos lingüísticos

A partir de los cincuenta se comenzó a trabajar 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 ya mencionada traducción automática, y por otro lado, trabajos teóricos dirigidos por Chomsky sobre formalización del lenguaje. Paralelamente a estas dos direcciones, se investigaba en Inteligencia Artificial que también incluía aspectos del PLN. 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.

Fue 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 estadísticos, hasta entonces utilizados casi de forma exclusiva.

Antes de introducirnos en este tipo de métodos indicar, que cuando hablamos de criterios lingüísticos nos estamos refiriendo entre otros aspectos a un análisis morfológico, sintáctico o semántico, por lo que a continuación se señala brevemente en que consiste cada uno de éstos. En el análisis morfológico se trata de realizar una segmentación de la palabra ortográfica para obtener la gramatical, y determinar su estructura y propiedades, esto es, en cada palabra se determina su raíz, considerando para ello posibles composiciones tales como prefijos y/o sufijos. Un algoritmo reciente que trata de agrupar vocablos franceses derivados de una raíz común bajo una sola, lo ha elaborado J. Savoy (5)

Hay algunos autores, que dentro de este módulo, incluyen un análisis morfosintáctico que lleva a cabo una revisión del resultado obtenido en el 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. (6)

El análisis sintáctico 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. Por último, el análisis semántico 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.

Después de acercarnos brevemente a algunos conceptos lingüísticos que irán apareciendo, pasamos a aproximarnos a la evolución que se ha ido experimentando en el campo de la indización automática, con la incorporación de estos métodos, a través del estudio de algunos sistemas.

El proyecto SMART constituyó, en los años sesenta y setenta, uno de los sistemas más avanzados en el análisis de texto de forma automática. Este sistema añadió a las herramientas utilizadas en los cálculos estadísticos, otras tales como: un método para extraer las raíces de las palabras inglesas, 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. 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. También 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.

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 "ecom-", 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" (7)

Paralelamente se fue trabajando en éstas y otras direcciones, buscando en cualquier caso un sistema automático que fuera capaz de sustituir al indizador humano, y debido a la proyección que estaban tomando los estudios en este campo, los investigadores se fueron posicionando en los que reconocían los beneficios potenciales del análisis tanto sintáctico como semántico, y en aquellos que planteaban sus reservas en cuanto al poder de resolución que podían aportar estos análisis a la indización automática, pero muchos de ellos reconocían, que en un futuro proporcionarían mayores resultados.

De este modo se llegó a finales de los ochenta. Entonces gran parte de los sistemas implementados 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 suministrar términos de indización adecuados sin apoyarse en el contexto y otras consideraciones semánticas. Por estos motivos, investigadores dedicados al estudio de estas técnicas seguían pensando que mientras la utilización de metodologías sintácticas fuera tan complicada computacionalmente, requiriera tanto espacio de almacenamiento y la disponibilidad de aplicaciones fuese menor que en la metodología estadística, seguirán recomendando ésta última, puesto que ante resultados parecidos se debe elegir la más simple (8) Efectivamente, hay informáticos que reconocen que a pesar de haber diseñado programas que tratan la sintaxis y la semántica de frases completas, sólo servían para contextos limitados, pero incluso en estas situaciones restringidas, los programas son muy complejos.

Algunos autores mantienen 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, es debido a que 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 desambigüar el significado de éstos.

Y como último factor, esta situación es debida a que generalmente, en los sistemas de indización automática cada tipo de palabras es tratado como indepen-

diente 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 (9)

El sistema ARIOSTO —diseñado e implementado a principio de los noventa— no se desarrolló 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 sistema se creó en el núcleo de un grupo de investigadores de Universidades italianas y está dedicado principalmente a la adquisición automática de conocimiento semántico desde corpus. El conocimiento que se adquiere es un grupo de asociaciones de palabras incrementado con marcadores sintácticos y semánticos. Estos datos se extraen a través de cuatro etapas usando técnicas de PLN. En este sistema no se realiza un análisis sintáctico profundo porque lo que importa, según sus autores, 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 por medio de un analizador simple basado en gramáticas discontinuas (que es capaz de detectar las ya mencionadas relaciones sintácticas binarias y ternarias). El sistema realiza el procesamiento de los textos en estas fases: un análisis morfológico, una segmentación del texto, un análisis sintáctico poco profundo, y un etiquetado semántico. Posteriormente el resultado de estos análisis se podrá utilizar en las distintas aplicaciones señaladas anteriormente (10)

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 (11)

En SIMPR (Structured Information Management: Processing and Retrieval) se 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 se lleva a cabo un análisis de carácter morfosintáctico que se descompone en subanálisis, con el fin de simplificar y clarificar computacionalmente los problemas, constituyéndose cada uno de los mismos en elementos individualizados del procedimiento general. Y 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 por último, estas secuencias de palabras significativas sufren entre otros un proceso de normalización (12)

Por último vamos a presentar varios programas aparecidos recientemente en el mercado con los que se pueda obtener una indización automática. Uno es el diseñado por la compañía Corporación Iconovex y llamado INDEXICON. Se trata de un software disponible en WordPerfect para Windows, Microsoft Word for Windows, y Microsoft Word para Macintosh.

Esta herramienta lee los documentos, localiza los términos y frases significativas y genera un índice. Para ello sigue un proceso que comienza con la lectura del documento y la marca de los términos indizables, y entonces realiza un análisis semántico y sintáctico con la finalidad de desambigüar el lenguaje empleado, 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. Utiliza también un diccionario compuesto por cincuenta y cinco mil palabras, y un conjunto de normas para determinar las partes de las palabras y los conceptos claves, y de este modo, analizar cada oración.

Asimismo ofrece a los usuarios la opción de elegir la profundidad del análisis, y por tanto el número de términos seleccionados, con hasta seis niveles diferentes. El nivel seis es el más bajo, proporcionando sólo los términos indizables más significativos, y por el contrario, el nivel uno es el más completo, puesto que incluye todos los términos indizables.

INDEXICON excluye automáticamente del análisis el índice de contenido así como las menciones de responsabilidad. Por otro lado, da la opción de descartar otras partes del texto como tablas o gráficos. También se puede intervenir para modificar o eliminar las marcas realizadas automáticamente a los documentos, así como dar énfasis a palabras y frases que el usuario considere importantes para él, de este modo, el programa genera términos de indización propuestos por

el usuario y los seleccionados de forma totalmente automática. Por último, se pueden añadir términos de indización a los obtenidos por el programa.

Este software es capaz de trabajar con información en distintos formatos como manuales, correo electrónico, documentos legales, médicos o técnicos. Y de la velocidad de trabajo cabe señalar que sobre un ordenador de 33 MHz con 16 MB de RAM, genera un índice para cincuenta páginas, de nivel tres en seis minutos y treinta segundos. Este índice, como señala David Haskins en la revista PC Magazine (septiembre, 1994), es bastante exacto y completo, pero reconoce que requiere ser pulido pero que incluso, con la participación del indizador humano para perfeccionarlo, sigue empleando sólo una parte muy pequeña del tiempo y coste que le llevaría a un profesional.

Para instalar el INDEXICON 1.0 para WordPerfect y Microsoft Word es necesario que el usuario cuente con una versión 6.0 o superior de estos programas, un 386 como mínimo, aunque recomiendan un 486 u otro más potente, la Windows 3.1, y ocho megas de RAM. Y el precio en agosto de 1995 de este software no llegaba a los 250 dólares.

La mayor parte de los datos que se han aportado de este sistema provienen de la misma empresa que lo ha diseñado, y no disponemos de otras fuentes donde se expongan posibles deficiencias o corroborar sus resultados, aunque intentaremos comprobarlos nosotros mismos ya que, estamos en fase de obtener una nueva versión INDEXICON que aún no ha salido al mercado, a través de las llamadas BETAS (13)

El segundo software que presentamos es el de la corporación *Excalibur Technologies*, S.A. Y según el director del producto en España, se está aplicando al análisis y recuperación de información una serie de tecnologías tomando como base los mismos principios que la inteligencia artificial y los sistemas expertos para producir otra que la han llamado Proceso de Reconocimiento Adaptativo de Patrones (PRAP), basada en redes neuronales. Esta nueva tecnología es capaz de aprender y decidir sus propias reglas en función de los datos, es decir, al igual que los sistemas biológicos pueden auto-organizarse a sí mismos de forma eficaz para maximizar los recursos disponibles, también lo puede realizar un programa de ordenador basándose en la metodología PRAP.

Disponemos de pocos datos de cómo el sistema lleva a cabo la indización automática de los documentos, sólo sabemos que es cada red neuronal la que se convierte en una memoria basada en el contenido, que se optimiza para la información en cuestión que gestione; del mismo modo el sistema también se auto-optimiza para el patrón del lenguaje, y para el patrón del tema tratado (14)

En el documento que la compañía tiene consultable en Internet a través del Web, esta tecnología compuesta por la ya mencionada metodología PRAP y las

redes semánticas (que incorpora aspectos del PLN entre otros), es capaz de indizar información multimedia compuesta por texto, imagen, gráficos, vídeo o sonido (15). No disponemos de información de cómo realiza el proceso de análisis pero, en cualquier caso también es un producto a tener en cuenta de cara a posibles resultados que pueda ofrecer, si no ahora en un futuro muy próximo.

Por último, para concluir con este apartado donde se ha estudiado la contribución del PLN a la indización automática, ofrecemos un organigrama donde aparecen esquemáticamente algunos de los distintos módulos que podrían componer un Sistema de Indización Automática. (Fig.1)

3. Otras investigaciones en indización automática

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, gráficos y sonido, como fruto del desarrollo que ha experimentado, entre otros, el procesamiento digital de imágenes (PDI) por la incorporación de soluciones desde áreas tan diversas como la inteligencia artifi-

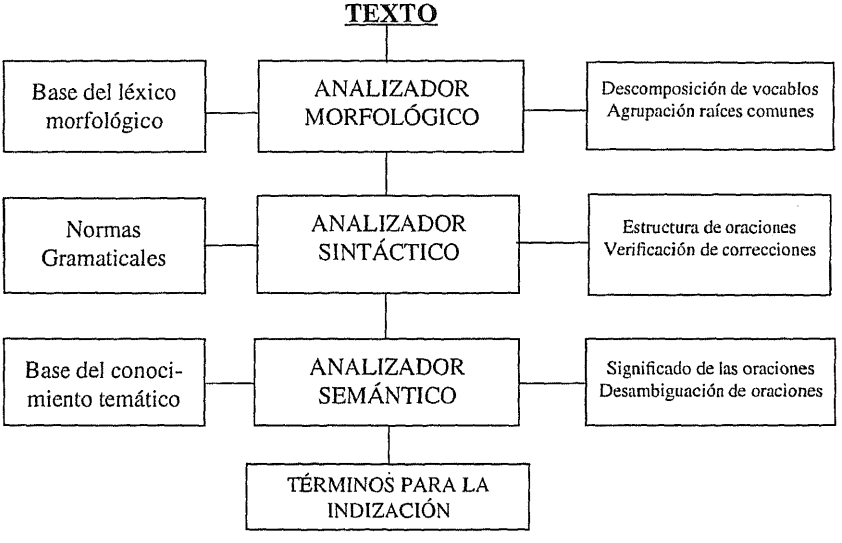

Fig.1. Módulos de un Sistema de Indización Automática

cial, la teoría de ondas o la morfología matemática. Hoy en día el PDI ocupa un espacio fundamental en campos como las telecomunicaciones o la robótica (16)

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 (17)

F. Rabitti y P. Savino (18) tratando de facilitar el acceso a las bases de datos de imágenes han ideado un sistema capaz de indizarlas de forma automática. Ellos 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.

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 se 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. En la Britihs Library se viene trabajando para la indización automática de imágenes con Excalibur EFS, software anteriormente mencionado (19)

Otro campo al que se está dirigiendo la indización automática es al del sonido. Con la ya señalada expansión de la multimedia se está incrementando el número de bases de datos que contiene este tipo información. Y un ejemplo para facilitar tanto el acceso como el tiempo y esfuerzo para seleccionar sonidos es un ensayo que se ha realizado utilizando redes neuronales (20)

Finalmente, señalar que investigadores de la Universidad de California (Berkeley) han tratado de indizar recientemente un tipo de información textual 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 programa se extraen palabras y frases que con-

tienen nombres de lugares geográficos o características de éstos que serán empleadas como posibles términos de indización. El objetivo 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 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 (21).

4. Conclusiones

Como se ha podido comprobar al comienzo de este trabajo todavía existen aspectos que solucionar en la indización humana. Los problemas surgen porque es una tarea que entraña cierta dificultad, y quizás principalmente porque es complicado conseguir una plena normalización en las operaciones que comprende. Otro de los problemas es el debate de si indización humana o automática, indudablemente se lleva mucho tiempo trabajando para conseguir un algoritmo que sustituya la labor intelectual de una persona, y hemos podido comprobar que procede de la aplicación de medios lingüísticos y no estadísticos.

En el procesamiento del lenguaje natural la complejidad deriva de la comprensión de los textos, aunque por lo que hemos visto, se ha avanzado enormemente en este terreno. Hemos constatado también que no hay una única corriente de cómo diseñarse un sistema de indización automática, ya que por un lado, hay quienes proponen sistemas que no ejecutan un análisis sintáctico completo, mientras otros prefieren no realizar el semántico.

Sin embargo, mientras todavía se aportan distintas opciones para conseguir un procesamiento del lenguaje natural apropiado y poder aplicarlo en el análisis del contenido de los documentos, se lleva unos años investigando en cómo lograr indizar no ya la tradicional información textual sino imágenes, sonidos, vídeos o gráficos. Y creemos que éste es el camino por el cual deben girar las investigaciones actuales, porque la información multimedia tiende a estar cada vez más presente.

Por último, en cuanto a los dos software disponibles en el mercado que se han presentado, hay que esperar a su evaluación para poder juzgar su verdadera capacidad tanto en la indización como en la validez de los términos propuestos para ejecutar una recuperación efectiva de los documentos.

5. Notas

(1) Grishman, R. (1991). Introducción a la lingüística computacional. Madrid : Visor Distribuciones, 1991

(2) Stevens, M.E. (1965). Automatic indexing: a state of the art report (Monograph 91).

National Bureau of Standars : Washington, D.C., 1965.

(3) Salton, G.; H. Wu; CT. YU (1981). The measurement of term importance in automatic indexing. // Journal Of the American Society for Information Science. May (1981) 175-186.

(4) Artandi, S. (1976). Machine indexing: linguistic and semiotic implications. //Journal of the American Society for Information Science. (1976) 235-239.

(5) Savoy, J. (1993). Stemming of french words based on grammatical categories. // Journal of the American Society for Information Science. 44 : 1 (1993) 1-9

(6) Antonacci, F.; et al.(1988). Representation and control strategies for large knoledge domains: A application to NLP. // Applied Artificial Intelligence. 2 (1988) 213-249.

(7) Salton, G. (1983). Introduction to modern information retrieval. McGraw Hill : New York, 1983.

(8) Salton, G. ; et al. (1990). On the application of syntactic methodologies in automatic text analisys. // Information Processing & Management. 26 : 1 (1990) 73-92.

(9) Deerwester, S.; et al. (1990). Indexing by latent semantic analysis. // Journal of the American Society for Information Science. 41 : 6 (1990) 391-407.

(10)Pazienza, M.T. (1994). Extraction of semantic knowledge from text: a goal or a starting point? // Actas X Congreso SEPLN. (1994) 1-23.

(11)Velardi, P. (1991). How to encode semantic knowledge: a method for meaning representation and computer-aide acquisition. // Association for Computational Linguistics. 17 : 2 (1991) 153-170.

(12)Karetnyk, D. (1991). Knolewledge-based indexing of morpho-syntactically analysed language. // Expert Systems for Information Management. 4 : 1 (1991) 1-29.

(13)Información obtenida en INTERNET en noviembre de 1995, en URL : <http://www.iconvex.com>.

(14)Maseda, F. La nueva generación de sistemas de gestión documental: tecnología de redes neuronales aplicada a la recuperación textual. // 5es. Jornades Catalanes de Documentació. 1995, p. 267-276.

(15)Información obtenida en INTERNET en noviembre de 1995. URL: <http://www.xrs.com>

(16)Ramirez, J. Información aparecida en un grupo de noticias de INTERNET, junio 1995. <jrp@esga.es>

(17)Bordogna, G.; et al.(1990). Pictorial indexing for an integrated pictorial and textual IR environment. // Journal of Information Science. 16 (1990) 165-173

(18)Rabitti, F., P. Savino (1992). Automatic image indexation to support content-based retrieval. // Information processing & management. 28 : 5 (1992) 547-565

(19)Alexander, M. (1995). Automatic indexing of document images using Excalibur EFS. .// Library Technology News. 16 (1995) 4-8.

(20)Feiten, B., S. Gunzel. Automatic indexing of a sound database using self-organizing neural nets. // Computer music journal. 1994, 18 (3) 53-65.

(21)Gyle Woodruff, A.; Plaunt, C. (1994). Gipsy Automated Geographic Indexing of Text Documents. // Journal of the American Society for Information Science. 45 : 9 (1994) 645-655.