LA INFORMACION Y SU BUSQUEDA

Por JOSE EDUARDO ARRECHEA BELZUNCE
Ingeniero de Caminos, Canales y Puertos

La información es uno de los elementos vitales de todo sistema social; el hombre ha vivido siempre rodeado de información lo mismo que de electricidad atmósferica y campos electromagnéticos, pero así como no supo captar y dominar la energía eléctrica hasta mediado el siglo XIX, tampoco ha sabido manejar masas considerables de información hasta mediado el siglo XX, valiéndose de los computadores electrónicos.

Si el control de la energía ha tenido las trascendentales consecuencias que todo el mundo reconoce, no va a ser menor el impacto que producirá la elaboración, depuración y control de la información a cualquier nivel y a cualquier distancia.

Existe un profundo paralelismo entre la energía y la información. Ambas con elementos del mundo moderno que sin ellas dejaría de ser lo que es hoy; ambas tienen un doble aspecto de continuidad y discontinuidad; ambas son medibles, es decir, expresables numéricamente: la energía en cuantos, en kgm, en kwh, etc., y la información, en bits; ambas se presentan en formas variadísimas. Norberto Wiener, uno de los padres de la moderna tecnología, dijo en cierta ocasión estas palabras: "La información no es materia, ni energía; la información es sólo eso: información".

Si esta definición no nos satisface podemos decir: información es algo contenido en un mensaje formado por dígitos y letras, o bien por una señal variable, continua como el sonido de una sirena o el nivel de una columna termométrica.

Además, estas señales contienen verdadera información si el cambio de signos se realiza de un modo que el receptor no puede predecir exactamente.

Si el mensaje se realiza utilizando dígitos binarios, cada uno de estos dígitos recibe el nombre de bit; el bit es una señal bivalente, la extensión del mensaje puede representarse numéricamente por el número de bits que contiene.

Sin embargo, dos mensajes sobre un mismo asunto, conteniendo el mismo número de bits, pueden poseer valor informativo muy diferente y eso lo saben las agencias periodísticas.

La teoría de la información trata, desde luego, de establecer el valor práctico de un mensaje según los efectos que produce su difusión.

Estos efectos se refieren a las modificaciones de opinión, de criterio o de conducta que el mensaje produce en el destinatario.

Por ejemplo, un mensaje cuyo texto dice: "hoy es lunes y mañana será martes", aunque su medida en bits sea, por ejemplo, 230, su contenido de información es prácticamente nulo.

Si yo no juego o no me interesa la lotería nacional, la lista completa de números premiados tiene para mí escaso o nulo valor informativo, y digo escaso porque tal vez signifique la fortuna de algún familiar o amigo mío.

Análogamente, si yo jugase un décimo de cada uno de los billetes que entran en el sorteo, tampoco me aportaría información alguna la lista de los números premiados.

Esto indica que, una noticia, un mensaje, un informe recibido por un determinado destinatario en una determinada situación, carece de valor si el destinatario está absolutamente seguro de su contenido, es decir, si la probabilidad del hecho o suceso en cuestión es 1.

Esta es la idea fundamental sobre la que se apoya la teoría de la información en su aspecto valorativo, lo que podemos llamar valor práctico de un informe.

Sea x_i la noticia o el informe; desde luego su valor será una cierta función que representaremos por $I(x_i)$.

¿Cómo podemos obtener esta función? Para ello Shannon hace las cuatro observaciones siguientes:

1.ª Si x_i e y_i son dos noticias completamente independientes, la información suministrada

por la noticia compuesta $(x_i \mid y_i)$ debe ser suma de las informaciones parciales de x_i e y_i, o sea:

$$I(x_i \mid y_i) = I(x_i) + I(y_i).$$

2.ª El valor informativo de la noticia x_i debe ser función exclusivamente de la probabilidad de realización del suceso a que se refiere la noticia, o sea:

$$I(x_i) = f[p(x_i)].$$

3.ª A un incremento infinitésimo de la probabilidad debe corresponder un incremento también infinitésimo del valor de la información, o sea que las funciones f e I deben ser continuas.

4.ª Cuando la noticia o informe conste de un solo bit, la probabilidad asignada a esta noticia debe ser $\frac{1}{2}$.

Esta condición es lógica, porque si yo espero que se produzca un 0 o un 1, con esperanzas iguales, la probabilidad *a priori* de cualquiera de los dos resultados es $\frac{1}{2}$.

Sin grandes dificultades puede verse que la única función que cumplen estas cuatro condiciones es la siguiente:

$$I(x_i) = l\frac{1}{p(x_i)}.$$

En la práctica suelen tomarse logaritmos decimales o logaritmos binarios; en el primer caso la unidad se llama decit, y en el segundo, se llama bit, aunque esta denominación puede causar confusión entre los bits cuantitativos anteriormente explicados y los bits de valoración a que ahora nos referimos.

En definitiva, acabamos de ver que el valor de una información es el logaritmo de la inversa de la probabilidad asociada a esa información.

El propio Shannon quedó maravillado al encontrarse con este resultado, pues su fórmula era idéntica, salvo el signo, a la que nos da la entropía de un sistema según la mecánica estadística.

Recordemos que la entropía es el logaritmo de una probabilidad; recordemos también que el segundo principio de la termodinámica, una de las leyes supremas de la física, dice que la entropía del universo siempre va en aumento.

De todo ello se deduce que la información adquiere un rango capital, tanto en la teoría como en la práctica, por ser, en fin de cuentas, una entropía negativa.

Sabemos que el aumento de entropía significa tendencia hacia la desorganización y el desorden; la ley de la entropía nos dice que el Universo abandonado a sí mismo, como un puro mecanismo sin conductor, tiende a la desorganización y al desorden; ahora bien, la información por su carácter de entropía negativa puede restablecer la organización y el orden en aquel sistema.

A la informática, ciencia de la información, actualmente en desarrollo, le espera un futuro brillantísimo, dentro de la ingeniería, gracias a la ayuda de los computadores electrónicos.

En realidad, una computadora electrónica no es más que un generador de bits, en la cantidad que sea necesaria y a ritmo tan rápido que va más allá de todo lo imaginable; el corazón de las computadoras puede dar más de mil millones de latidos por segundo; en cualquier terminal del circuito podemos disponer de mil millones de señales por segundo.

Decimos que corriente eléctrica es aquello que circula dentro de un cable conductor; hoy sabemos que eso que circula son los electrones; también podemos decir que información es aquello que circula dentro de una computadora electrónica, y eso que circula son los bits.

Pero la información no se crea en las entrañas de la computadora; la información se engendra fuera de la máquina y su destino también está fuera de ella.

Tanto la información como la energía poseen, además de su valor intrínseco propio, otro valor que depende de la circunstancia.

De ahí la importancia de los sistemas de transporte de la energía y de transporte de la información.

El esquema general de un sistema de transmisión de información, aparece en la figura 1.ª.

La transmisión de datos y su procesamiento a distancia, recuperando en origen la información elaborada, se esquematiza en la figura 2.ª.

Por otra parte, la vida moderna exige crear embalses de información en archivos, bibliotecas, centros bibliográficos, etc.

Esto plantea el problema de la información retrieval (I.R.), que trata de localizar y repro-

ducir automáticamente un informe, una nota o un dato que nos interesa, contenido en un archivo.

La intensificación de las actividades científica y administrativa a cualquier nivel hace inservibles los archivos tradicionales.

El problema presenta dos aspectos bien conocidos de antiguo; la creación de los ficheros y su utilización práctica en la oficina.

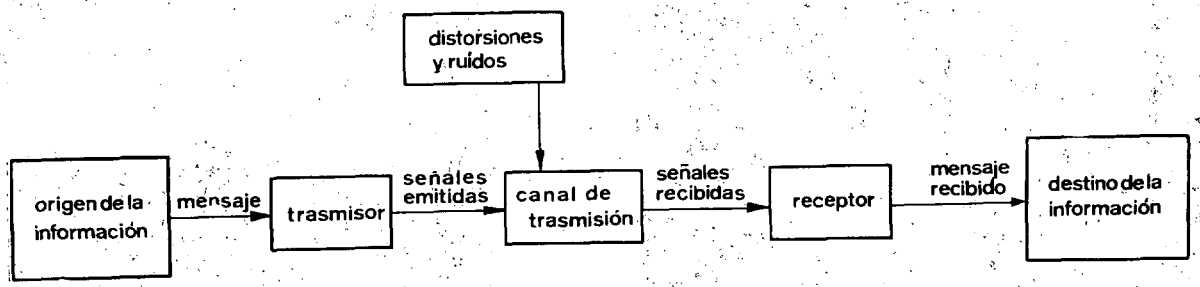


Figura 1.ª

La creación del fichero supone analizar los textos, establecer rúbricas, agrupar y clasificar las materias. Una vez sentadas las bases más convenientes, el fichero crece al ritmo de la información que diariamente entra en la oficina, pero en el momento de utilizar ese fichero nos encontramos con las dificultades, muchas veces insuperables, de todo problema de grandes masas, que exigen tiempo y molestias considerables.

Sin embargo, hoy viene observándose que es en la creación de los modernos ficheros donde se presentan los mayores obstáculos para el buen encauzamiento y embalse de la información, que crece a un ritmo imposible de seguir con los medios convencionales.

De ahí que los métodos a emplear en la documentación automática deban obedecer a directrices nuevas.

Parece que las palabras clave contenidas en el propio texto son el mejor fundamento del análisis que debe preceder a la creación del fichero.

Este análisis lo efectuará la propia máquina, registrando esas palabras clave que nosotros hemos subrayado en el texto original; también la máquina realizará los trabajos para clasificar, intercalar y localizar estas palabras clave, mediante las cuales se podrá encontrar la información deseada en el momento preciso.

El sistema KWIC (Key Word-in-Context) es un paso hacia ese objetivo, que puede verse, por ejemplo, en el Catálogo de Programas IBM 1.620/1.710-20 del primer semestre de 1966.

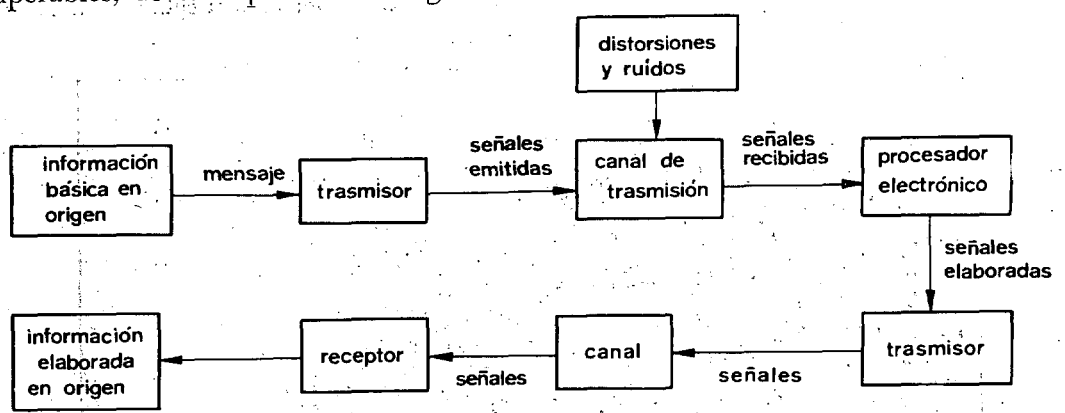


Figura 2.ª

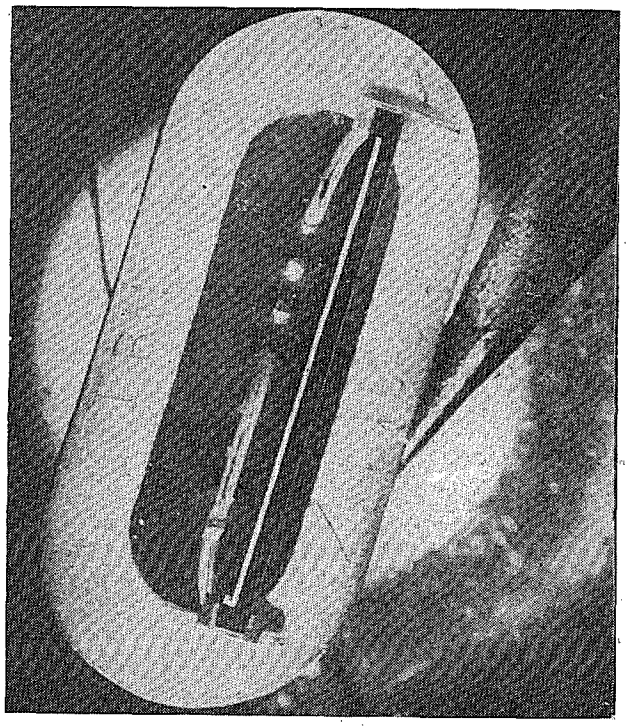
En resumen, el problema de la selección automática de la documentación tiene tres aspectos:

1.° Lectura automática de los textos originales.

2.° Análisis resumido e indexación automática.

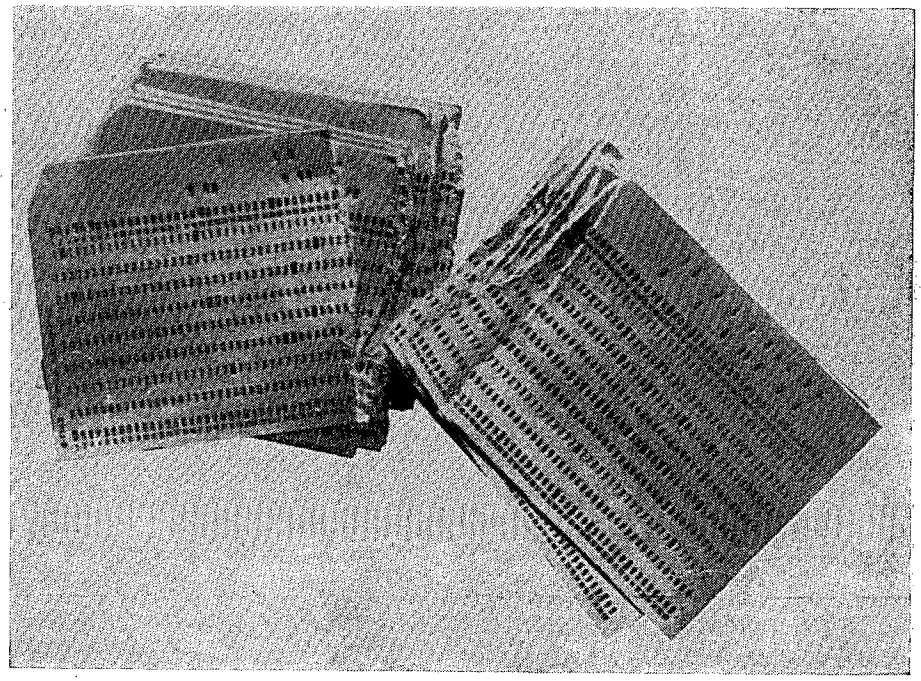
3.° Selección y localización de los documentos, referencias o notas que interese conocer.

El problema está todavía lejos de alcanzar



El presente: Explorador óptico de estado sólido capaz de convertir directamente imágenes en señales eléctricas. Su tamaño aparece contrastado con la punta de un lápiz.

El futuro: Lectora óptica de documentos.



¿El futuro?

una solución satisfactoria y definitiva; sin embargo se han conseguido resultados muy aceptables.

Uno de ellos, entre los primeros, fue realizado en Francia, por la Compañía de Saint Gobain, con la asistencia técnica de IBM y de la cual se dio cuenta en el II Congreso de AFIRO (año 1961).

Se trataba de indexar multitud de textos químicos mediante un sistema de palabras-clave sin codificación previa.

De este modo el trabajo del analista se simplificaba al máximo, pues se limitaba a una lectura atenta del original para formar una lista de palabras clave características, a su juicio, del documento.

Estas palabras conservaban su lexicología natural y eran ordenadas alfabéticamente.

Con estos preparativos y la indexación, el fichero estaba creado.

Para la explotación del fichero sería suficiente, en cada caso, formar una lista de palabras clave, vinculadas al tema en cuestión de ciencia química, u otra disciplina, pero que también podría ser una disposición oficial, un nombramiento, una concesión, etc. Así podremos saber cuáles son los documentos indexados o rubricados, mediante dichos conceptos, hasta localizar exactamente el documento o nota deseados mediante selecciones progresivas.

Teóricamente el método es sencillo, pero se observaron dificultades en orden a las sinonimias, jerarquías de conceptos (por ejemplo, Chlore y Halogene), sintaxis (por ejemplo, forets pour acier y acier pour forets).

La experiencia se hizo sobre una masa de 10.000 documentos, entre libros, revistas, patentes industriales, todas ellas relacionadas con la Química.

La creación y tratamiento del fichero se hizo con un Ordenador IBM 705, aunque dada la reducida masa de datos hubiera servido una máquina menos potente.

El trabajo manual de los analistas se reducía a redactar por cada documento una ficha con los siguientes datos:

1.° Número de identidad del documento.

2.° Título, autor y detalles complementarios.

3.° Lista de palabras clave, en orden alfabético.

En promedio, se invirtieron siete minutos en preparar cada ficha, y el número medio de palabras clave por ficha, fue 8.

Después de formar 1.500 fichas se redactó una relación alfabética de todas las palabras clave utilizadas hasta ese momento, y se estudiaron las posibles sinonimias; se asignaron índices numéricos a las palabras clave y al conjunto de sinónimos.

Al proseguir la experiencia se fueron detectando automáticamente las nuevas palabras clave para incorporarlas a la lista general.

Estas fichas del analista eran perforadas en tarjetas IBM, y grabadas después en tres ficheros distintos sobre bandas magnéticas.

Los 10.000 documentos sirvieron para crear un fichero de léxico con 12.000 palabras.

Las palabras claves, "aglomeración" y "análisis", dieron 30 y 26 sinónimos, respectivamente.

Se obtuvo un fichero llamado de biblioteca y otro de selección.

También pudo comprobarse la influencia del ruido en el proceso de selección, que alcanzó el 25 por 100, es decir que de cada cuatro documentos señalados por la máquina, uno era superfluo o no relacionado con la pregunta formulada. En cambio, el silencio, o sea el conjunto de registros pertinentes al caso y no acusados por el ordenador, fue del 8 por 100.

Ambos defectos del sistema eran consecuencia de un análisis incompleto, o de la incorrecta agrupación de los términos básicos.

Cada pregunta al ordenador requería una preparación de diez minutos, en promedio, para fijar la selección y orden de las palabras clave, perforar las instrucciones lógicas especiales y los bits de indexación.

El ordenador IBM 705 contestaba, en promedio, una pregunta por minuto.

Consciente de estas dificultades, el Engineers Joint Council editó (año 1964) el Thesaurus of Engineering Terms sobre la base de 119.000 tecnicismos, aportados por 18 organismos del máximo prestigio en U. S. A.

Indudablemente este Thesaurus es instrumento de gran valor para controlar el vocabulario, indexar la información, archivarla y recuperarla en un momento dado.

Sería muy interesante la preparación de un Thesaurus de análoga calidad para los aspectos administrativos de la actividad ingenieril no considerados por el EJC.

La V reunión de AFIRO (1966) se ocupó de otros aspectos básicos de la I.R., que el perfeccionamiento del hardware permite enfocar con creciente optimismo.

Se trata no sólo de localizar el documento que interese y obtener mecanografiado un breve extracto del mismo, sino de ver reproducido y poder consultar el original a través de la pantalla de rayos catódicos que poseen los más modernos ordenadores, e incluso se pretende lograr esa información completísima desde cualquier despacho u oficina auxiliar mediante dispositivos de teleproceso.

Desde hace pocos años, las grandes casas constructoras de máquinas electrónicas han ideado dispositivos para guardar millones de documentos, que son variantes de la técnica del microfilm; en una película de 33 mm.^{2} caben un centenar de imágenes o documentos.

El sistema denominado WALNUT, utiliza unas cubetas que contienen 200 celdillas de plástico, cada una de las cuales, a su vez, puede contener 50 películas, o sea un total 1 000 000 de imágenes o documentos, equivalentes a unos 3 000 libros de tamaño corriente.

Las conclusiones actuales sobre el problema, pueden resumirse del siguiente modo:

1.º Un cierto grado de automatización en la I.R. es económicamente realizable con la técnica actual e incluso se está utilizando en algunos países.

2.º Dado el tenaz esfuerzo de los investigadores y la importancia del problema, es indudable que en plazo no muy lejano la I.R. constituirá uno de los más brillantes triunfos de la automación; el plan WARREN, que se está estudiando en Norteamérica, es buena promesa de éxito, así como la existencia de lenguajes especializados como son el Tabsol, el Smart, el Asm, el Satire y otros.

Bibliografía.

1. SHANNON y WEAVER: "The Mathematical Theory of Communication." *University of Illinois Press*, 1949.
2. CULLMAN, DENIS PAPIN y KAUFFMAN: *Elements de Calcul Informationel.* Albin Michel. París, 1960.
3. J. BECKER y R. HAYES: *Information Storage and Retrieval: Tools, Elements, Theory.* J. Wiley Sons, Inc., 1963.