Falta p. 7

REVISION DEL ESTADO ACTUAL DE LA AUTOMATIZACION DE LOS PROCEDIMIENTOS DE ALMACENAMIENTO Y RECUPERACION DE INFORMACION DOCUMENTAL.

Por: Alejandro Martínez-Márquez*

Introducción

1). La época actual parece invadida del fenómeno de explosión en diferentes áreas de la actividad humana incluyendo el renglón demográfico. Sin embargo la producción literaria ha experimentado un incremento sucesivo cada vez mayor de modo que su crecimiento es exponencial y se estima que para el año 2000 el número de publicaciones periódicas será del orden de un millón de publicaciones científicas. La fig.1 ilustra el proceso de crecimiento del número de revistas a partir de la segunda mitad del siglo XVII y en ella se ha empleado una definición adecuadamente amplia de "publicación o Revista científica".

Se ocurriría a primera vista que hasta muy recientemente nos hemos empezado a preocupar por la cantidad de publicaciones periódicas en el mundo. Sin embargo, en 1945, V. Bush(1) escribía "Profesionalmente nuestros métodos de transmitir y revisar los resultados de la investigación corresponden a los de varias generaciones atrás y esos métodos son definitivamente inadecuados para nuestra época". Ese comentario es hoy en día aplicable a

---

* Doctor en Ingeniería, Sub-Director de Cursos de Graduados del Instituto Politécnico Nacional de México.

nuestro país en forma natural para la gran mayoría de las actividades de investigación. Y el volumen de información generada en las revistas ha crecido desde 1945 a una tasa mucho mayor - que la del crecimiento demográfico, según lo refleja la Fig. 1 en donde se nota que cada 50 años el número de publicaciones - se multiplica por 10.

El crecimiento tan acelerado que se ha producido del número de publicaciones, ha confrontado, al usuario de la información, con una tarea magna cuando de emprender una nueva búsqueda se trata. No es del todo igual para quienes conservan una misma - línea de investigación y se mantienen informados en ocasiones mediante comunicaciones personales o por conducto de un número muy reducido de revistas. Sin embargo, aún en ese caso, el almacenamiento y recuperación de información se convierte, eventualmente, en un problema.

Tratando de corresponder al crecimiento del volumen de información, se ha desarrollado un buen número de procedimientos mecánicos unos y otros de tipo visual que acuden en auxilio del usuario de la información. En la última década se ha promovido de modo muy extenso el uso de la computadora digital como el auxiliar adecuado para otorgar al usuario de la información - aquellas facilidades que le permitan encontrar en forma rápida las referencias y documentos que contienen información relevante a un tema en el que el usuario tiene interés. En conexión - con este objetivo encontramos el desarrollo de un conjunto de áreas de investigación a las que nos referiremos con mayor -

detalle en las Secs. 5 y 6 de este trabajo.

2) Para aclarar nuestras ideas expuestas en el punto anterior, es necesario señalar que la actividad de búsqueda de información científica no es de ninguna manera investigación básica. Pues - si bien es cierto que la producción científica implica la búsqueda de información, esta actividad por sí misma no reditúa en resultados de investigación; la diferencia es el procesamiento que se hace de la información y que es el que eventualmente puede aportar conocimientos a la ciencia. Este procesamiento en ambos casos implica análisis y síntesis de la información.
En forma más o menos convencional se suele diferenciar entre investigación científica e investigación tecnológica(2). Básicamente el producto de la investigación científica son nuevos conocimientos y el producto de la investigación tecnológica es algo que tiene valor económico y en consecuencia puede ser vendido o comprado; tal es el caso de las patentes, un producto químico, etc. Este último comentario nos lleva a considerar los usos de la información clasificándolos de la manera siguiente:

i) información destinada al usuario que trabaja en la investigación científica de tipo básico o de tipo experimental.
ii) información destinada a ingenieros y tecnologistas cuya actividad fundamental es la de diseño y/o la de operación en actividades productivas.
iii) información dirigida a actividades de tipo económico (a nivel nacional y/o industrial) y de manejo empresarial.

La separación entre estos usos de la información no es de ninguna manera perfecta ya que en alguna forma la información destinada a la investigación científica eventualmente se convierte en utilizable en los otros dos usos y viceversa. Sin embargo es conveniente conservar en mente esa división.

Un sistema de información puede ser suficientemente amplio como para auxiliar en cada uno de los usos señalados antes. Desde el punto de vista de procedimientos para almacenamiento y recuperación de información, tal clasificación de los usos de la información es innecesaria aunque desde el punto de vista de diseño e implantación de un sistema sea necesario una discriminación de ese tipo.

3). Un sistema de información documental es propiamente un enlace entre el usuario de la información y la información misma. A través de ese sistema el usuario recibe orientación oportuna sobre aquellos documentos o artículos de revista que contienen información utilizable en el desarrollo de su trabajo de búsqueda, análisis y síntesis. Para que un sistema de información del tipo automatizado logre su propósito es necesario al menos que incluya las siguientes características:

i). Capacidad para que el usuario pueda formular sus preguntas al sistema en el lenguaje del usuario.

ii). Capacidad para seleccionar dentro del archivo de referencias bibliográficas aquellas que responden adecuadamente a la pregunta del usuario.

iii) Capacidad para la actualización del archivo de documentos en forma periódica.

iv) Respuesta eficiente y en el menor plazo posible, al - usuario.

v) Confiabilidad de los resultados.

vi) Costo de la respuesta pequeño en comparación con el costo que implicaría la búsqueda mediante otro procedimiento.

El diseño adecuado de un sistema de este tipo implica la necesaria participación de un equipo de especialistas en diferentes áreas incluyendo bibliotecarios, especialistas en computación y en ciencias de la información. En la Sec. 6 de este documento discutiremos con amplitud los diferentes aspectos involucrados en la adecuada planificación de un Sistema de Almacenamiento y Recuperación de Información Documental (SARID).

Dependencia de la investigación con respecto a un sistema de información.

La Fig. 2 ilustra lo que consideramos es un proceso regular de revisión y análisis de información para cualquiera de los usos antes señalados. En esa figura se ha pretendido enfatizar la dependencia de la búsqueda de información con respecto a la operación de un sistema de información. En el diagrama de la Fig. 2 se han señalado con uno y dos asteriscos las actividades dentro del proceso en las que un sistema de información adecuado suministra ayuda determinante. Durante la etapa de familiarización con lo existente la información bibliográfica es importante ya que permite una ubicación rápida del usuario en relación con su problema. La siguiente etapa es igualmente dependiente de la información documental y corresponde al análisis de la importancia de cada uno de los documentos disponibles. En ocasiones el usuario puede preparar en esta etapa su fichero bibliográfico que le puede servir para la elaboración de un reporte en el caso de que la investigación amerite una publicación.

Realizada la etapa de análisis selectivo es necesaria una evaluación crítica de las publicaciones proporcionadas por el S.I. para establecer ulteriormente la línea de trabajo que más convenga a la resolución del problema planteado. En esta etapa es necesario suprimir aquellos documentos que resultan no-relevantes o que incluyen información redundante. El auxilio que puede prestar en esta etapa el S.I. es relativamente poco ya que se requiere fundamentalmente de los conocimientos del usuario. En

S.I. eficiente a través de una amplia interacción con el usuario podría eventualmente cubrir ampliamente esta etapa. Buena parte de las tendencias actuales en el área de la automatización se dirigen hacia este aspecto y existe hoy en día un buen número de publicaciones e investigaciones en proceso que intentan lograr, básicamente, hacer programable la acción del usuario frente a la discriminación de documentos no-relevantes(3).

Realizado el análisis y evaluación exhaustiva de la información obtenida del S.I. el usuario realizará el procesamiento de la misma formulando sus hipótesis de trabajo y recurriendo a la - prueba de las mismas con respecto al problema que intena resolver. Este procesamiento es de tipo iterativo o de suposición y error que concluye una vez que el problema queda resuelto a satisfacción del usuario. El usuario formula sus conclusiones y presenta sus resultados y recomendaciones. El S.I. le vuelve a auxiliar cuando con visión retrospectiva desea incorporar a una publicación lo que encontró en la información disponible y como influyó ella en el desarrollo de su trabajo.

La presentación anterior se consideró necesaria puesto que ubica muy claramente la importancia de las funciones de un S.I. y su interacción con el usuario. Ciertamente el modelo descrito antes podría no ser exhaustivo pero el propósito de su introducción - corresponde a la conveniencia de establecer la base de interacción entre el usuario y el S.I.

Un sistema conceptual de almacenamiento y recuperación de Información documental.

Puesto que en la sección anterior discutimos con amplitud la interacción entre la actividad del usuario de la información y un sistema de información, es pertinente definir aunque solo sea de manera conceptual una versión idealizada de un S.I. La Fig. 3 - pretende auxiliarnos en el propósito de esta sección. Esa figura destaca nuevamente la interacción entre el S.I. y el usuario pero ahora vamos a describir la organización ideal de un S.I.

Podríamos empezar por señalar que lo que llega al S.I.son documentos o referencias bibliográficas y las solicitudes de información del usuario. Lo que sale del S.I. son las reproducciones de aquellos documentos que se consideran relevantes a través de la interacción entre la parte automatizada del S.I. y el usuario.

Según se indica en la Fig.3 los documentos que ingresan al S.I. experimentan dos procesos distintos uno es el de clasificación y catalogación y el otro es el de almacenamiento de abstracts en el archivo de una computadora digital que pertenece a la parte automatizada del S.I. y que se ha encuadrado con línea interrumpida en la misma figura. En ocasiones la tarea de clasificación y catalogación de libros puede realizarse mediante el concurso de la computadora. Sin embargo las experiencias tenidas en este sentido no son completamente halagadoras según lo ha señalado D. Melcher en la Ref. 4.

Al ingresar los documentos a la parte automatizada del S.I. se

logra la actualización de los archivos y es posible solicitar - por conducto de la computadora y un sistema de comunicación entre el usuario y la máquina aquellas referencias que responden de modo relevante a la pregunta del usuario. Cuando el usuario interactúa con la computadora tiene opción de elegir aquellos documentos que se supone reunen las mejores condiciones de relevancia y solicitan les sean reproducidos. El sitema entrega - finalmente al usuario las reproduciones solicitadas las que se toman del Almacén físico y se devuelven después de su copiado.

Como puede apreciarse de lafig 3 la parte automatizada es solamente una porción reducida de todo el sistema. En concepto, el almacén físico de documentos no necesita estar en un lugar determinado, sino que puede encontrarse distribuído en una región determinada. En el caso de nuestro país, el almacén físico de documentos pueden encontrarse acumulado en las diferentes bibliotecas de las Universidades de los estados y/o bibliotecas de cad estado. Para llevar a cabo la reproducción de documentos es conveniente la instalación de los recursos pertinentes en cada lugar o recurir al préstamo interbibliotecario .

Acerca de la creación de un S.I., como el que antes se describe, en México podría darse seguramente un importante debate y esa es una de las razones por las que hemos considerado pertinente incluir el tema en este trabajo. Una discusión del mismo asunto a nivel internacional ha sido publicada por la UNESCO en la ref. 5.

Evaluación de un sistema de almacenamiento y recuperación de información.

Sin lugar a dudas la introducción de la parte automatizada en el S.I. introduce una facilidad apreciable para el usuario y viene a coadyuvar significativamente a la tarea cotidiana del bibliotecario quien solo de modo muy general puede suministrar orientación al usuario sobre el lugar en donde es posible encontrar la información que él busca. Como ya se dejo establecido la parte automatizada del S.I. deberá realizar de modo programado la selección de aquella información que es pertinente en relación con la pregunta del usuario. Esta labor ha dado lugar al desarrollo de técnicas que pretenden simular el proceso mental que se desarrolla en el usuario cuando realiza él personalmente la tarea de selección de material relevante. Lamentablemente, y como es natural, todavía no es factible hoy en día la ejecución satisfactoria de esta tarea por parte de la máquina. En tales condiciones el diagrama de la fig. 4 pretende caracterizar el fenómeno de busqueda de documentos con auxilio de la computadora con el propósito de evaluar el funcionamiento de la parte automatizada del S.I.

Si el usuario llevará a cabo la tarea de selección, tendríamos que él optendría como total de documentos relevantes la cantidad (a+c) que es solo una parte de todos los documentos incluidos en el archivo bibliográfico y que se representa en la fig. 4 por (a+b+c+d). Los documentos considerados no-relevantes por la computadora hacen un total de (c+d), y de entre todos aquellos que la computadora recupera mediante su pertinente programación, la

cantidad b resulta ser no relevante para el usuario. La situación reflejada en la fig. 4 es semejante a la que ocurre en estadística con la prueba de hipótesis y corresponde a los errores I y II.

En la misma figura 4 se aprecian las diferentes medidas que llevan a la evaluación de la parte automatizada de un S.I. La posibilidad de una interacción efectiva entre el usuario y la computadora reduce significativamente la cantidad de los documentos que no siendo relevantes ocurre que son recuperados, e.d. la cantidad de la fig. 4 . Al mismo tiempo la interacción permite que la cantidad c disminuya de valor.

A los propósitos de la eficiencia de un S.I. es muy importante la técnica de recuperación que se utilice. El modelo matemático más usual para este propósito es el suponer que con el thesaurus se puede construir un espacio de dimensión N, siendo este número el de elementos que contiene el thesaurus. De acuerdo con este modelo a cada documento que ingresa al archivo de la computadora le corresponde un punto o vector de posición que queda definido por el peso relativo de las palabras que aparecen en el abstracto y en título del documento. De esta manera se originan amontonamientos o "Clusters" de documentos en el espacio a que hacíamos mención antes. Al ingresar una solicitud de un usuario es posible evaluar el vector de posición de la pregunta pesando adecuadamente los elementos del thesaurus que contiene. Al hacerlo así se puede entregar como respuesta al usuario aquellos documentos que mejor responden a su pregunta porque se encuentran muy cerca del punto que le corresponde a ella.

La fig. 5 sugiere un modelo relativamente simple de recuperación de información basado en una matriz término-documento en donde aparecen en cada renglón los términos que caracterizan cada documento de los almacenados. El vector $\tilde{q}$ es el vector que contiene las palabras que aparecen en la pregunta que formuló el usuario. El producto de la matriz término-documento con el vector $\tilde{q}$ señala aquellos documentos que se ajustan a la pregunta. Ahí vemos en el caso particular de la fig. 5 que el documento que mejor responde a la pregunta es el 3 puesto que contiene 3 de las palabras señaladas en la pregunta.

Un aspecto que es esencial a la evaluación del S.I. es la planificación del mismo. Este tema amerita una revisión exhaustiva y en consecuencia se invita al lector a consultar las referencias citadas al final. Las figs. 6 y 7 incluyen un proyecto de un S.A.R.I.D. Las figs 8 9 y 10 incluyen un ejemplo experimental típico del comportamiento del sistema denominado SMART y que se detalla ampliamente en la Ref. 13 .

CONCLUSIONES

El objetivo de este artículo es el de informar a las personas interesadas en los Sistemas de Información de una variedad de aspectos que ameritan consideración en el desarrollo de un proyecto de creación de un S.I. al nivel nacional. Se espera que este documento pueda servir de motivación para una revisión completa de

las implicaciones, a nivel nacional, de una política de desarrollo de la Ciencia y la Tecnología que solo puede implementarse cuando se dispone de un sistema de información capáz de auxiliar a quienes requieren de información actualizada para realizar su actividad investigadora a través de la búsqueda de información.

REFERENCIAS.

1. Bush, V. "As we may think", reproducido del original publicado en 1945 en "Readings in Information retrieval", 1964, p. 20

2. Price, D. "The difference between Science and Technology". Edison birthday lecture. Dallas, February 10, 1968 p.11

3. Computer and Information Science Center, "Abstracts of the report to the NSF Office of Science Information Service", The Ohio State University, 1970.

4. Melcher, D. "Cataloging Processing, and Automation", American Libraries, July-August, 1971. pp. 701-713

5. "Information Systems Concept", Cap. 6 de un volumen publicado por UNESCO en 1969, pp. 219-276

6. Rees, A M. "The evaluation of retrieval Systems",

7. Murdock, J.W. and Liston, D.M.Jr. "A general model of Information Transfer: theme paper 1968 Annual Convention", American Documentation, October 1967, pp. 197-208

8. Weisberg, A.M. "Scientific Communication", Science and - Technology, 1963, pp.

9. Luhn, H.P. "The automatic creation of literature abstracts", IBM Journal of Research and development, April 1958, pp. - 159-165

10. Overhage, C.F. and Harman, R.J., "The On-line Intellectual Community and the Information Transfer System at M.I.T. in 1975", en "The Growth of Knowledge" by Manfred Kochen, J. Wiley, 1967, pp. 77-95

11. Salton, G., Keen, E.M and Lisk, M. "Design Experiments in Automatic Information Retrieval", en el mismo libro que la Ref. 10, pp. 336-351

12. Davis, W., "The Universal Brain: Is Centralized Storage of All Knowledge Possible, Feasible or Desirable ?, en el mismo libro que Refs. 10 y 11, pp. 60-65

13. Salton, G. "Automatic Information Organization and Retrieval", Mc Graw Hill Co., 1968, pp. 12-16

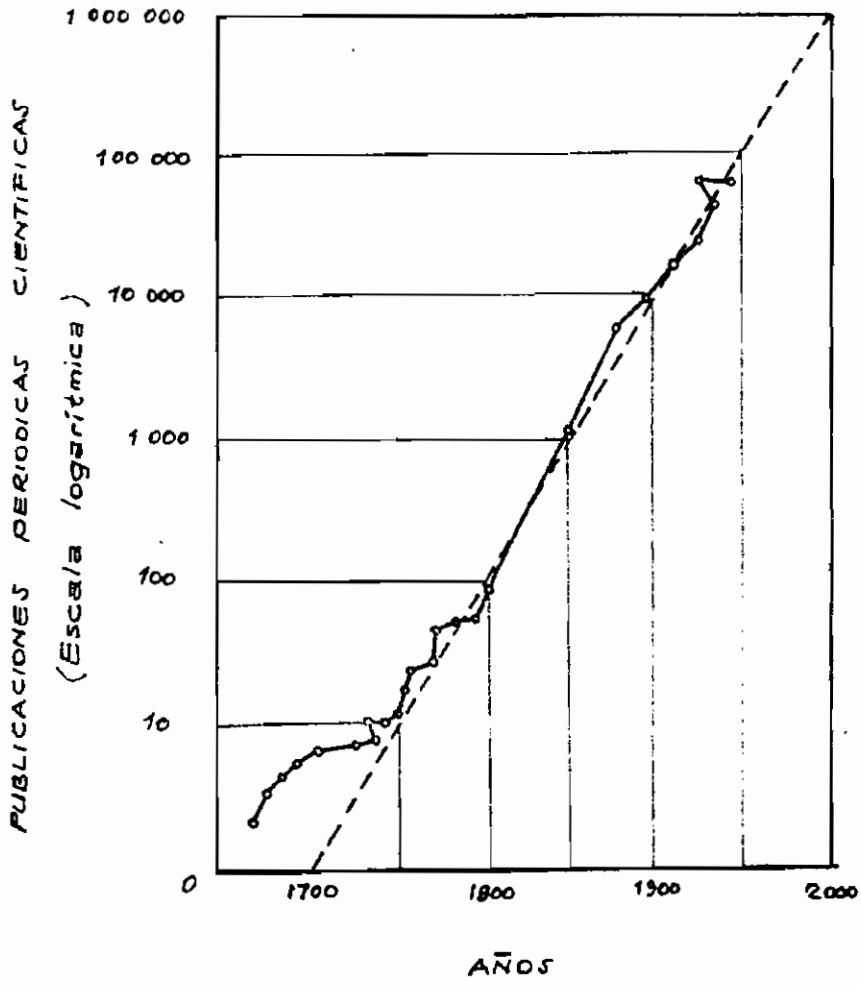


Fig. 1 Crecimiento de las publicaciones periódicas científicas en el mundo.

Tomado de: Price D. J. de S. "Science since Babylon". New Haven, 1961, p. 97.

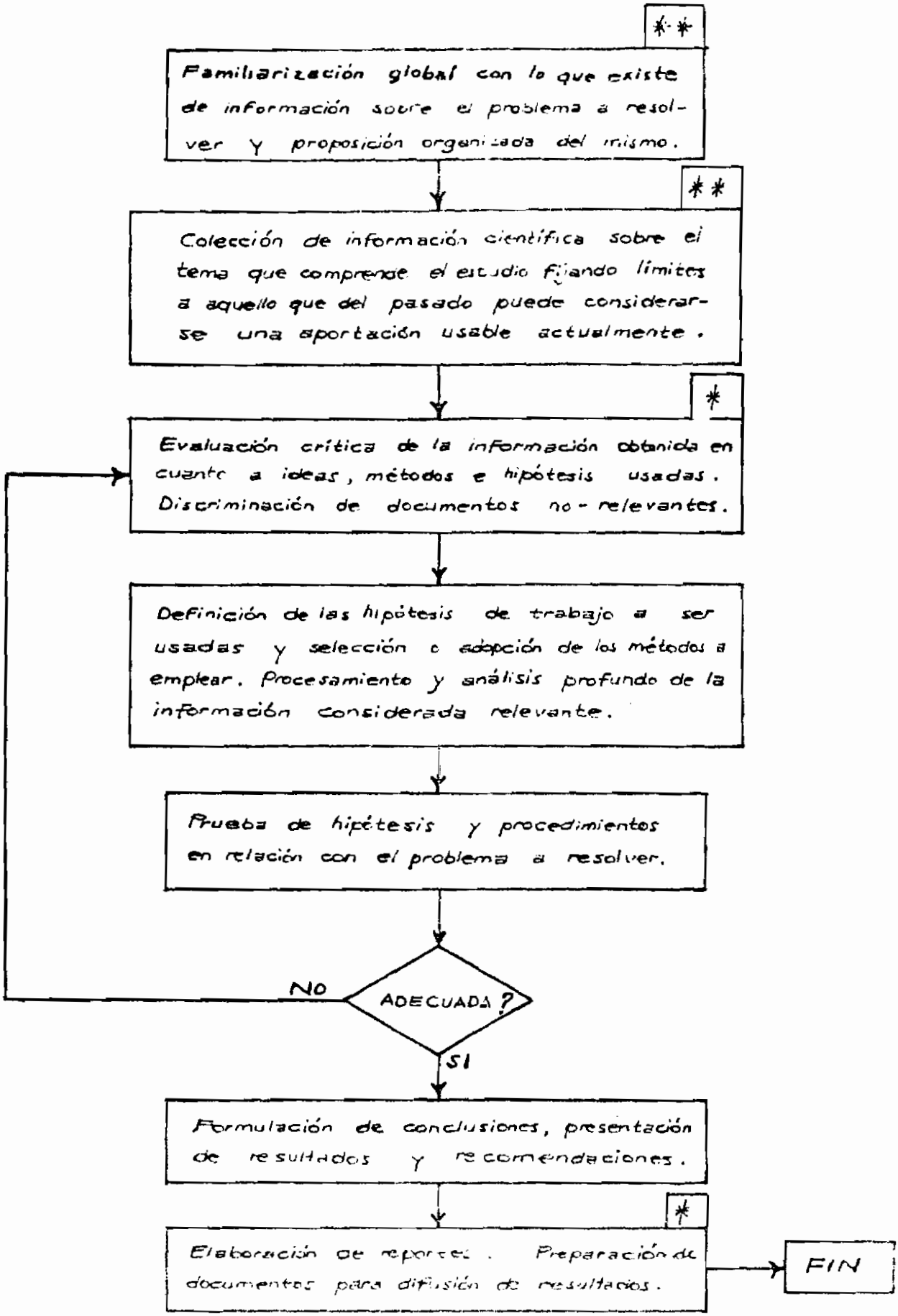


Fig. 2 La dependencia de la busqueda científica de información con relación a un sistema de información documental.

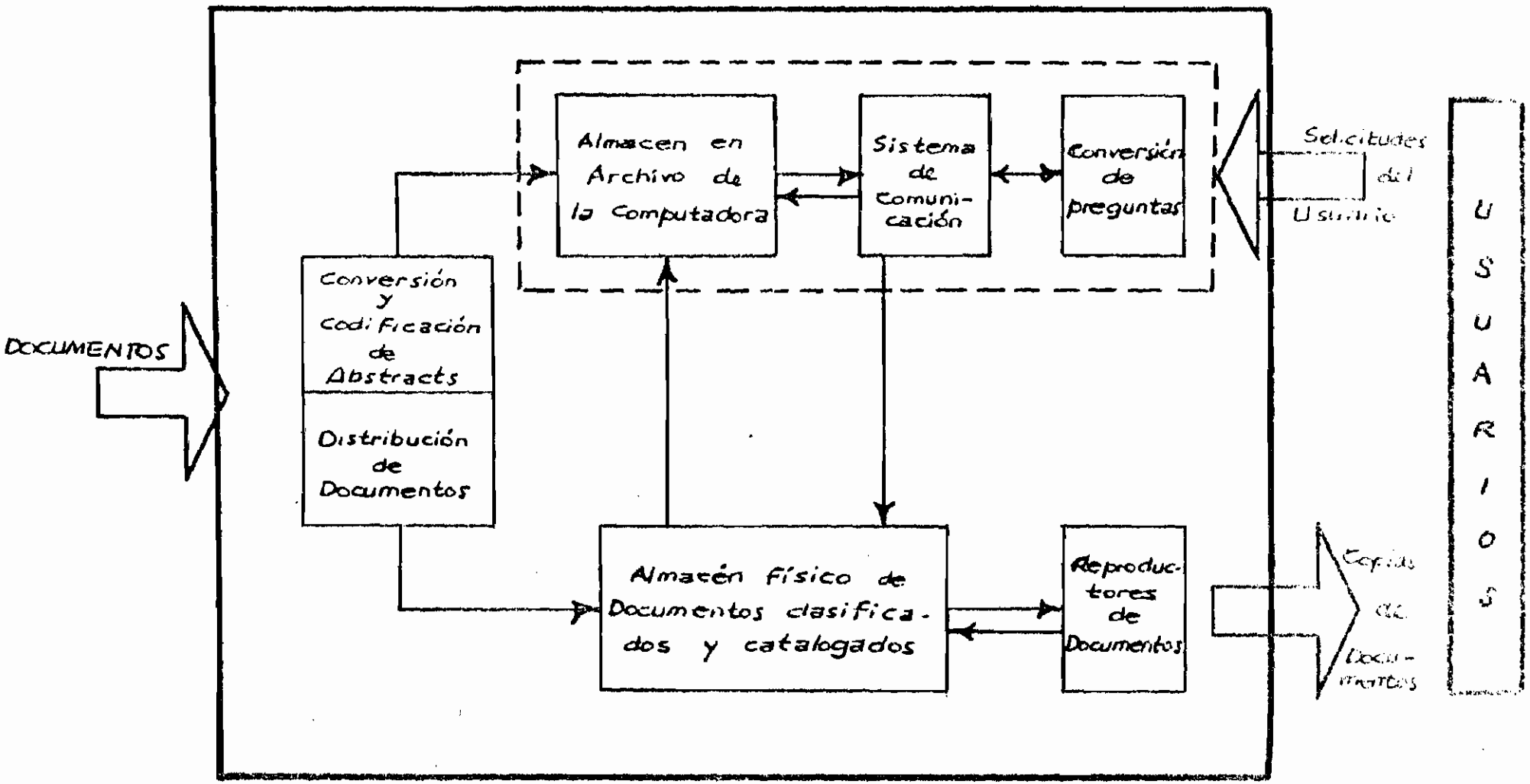
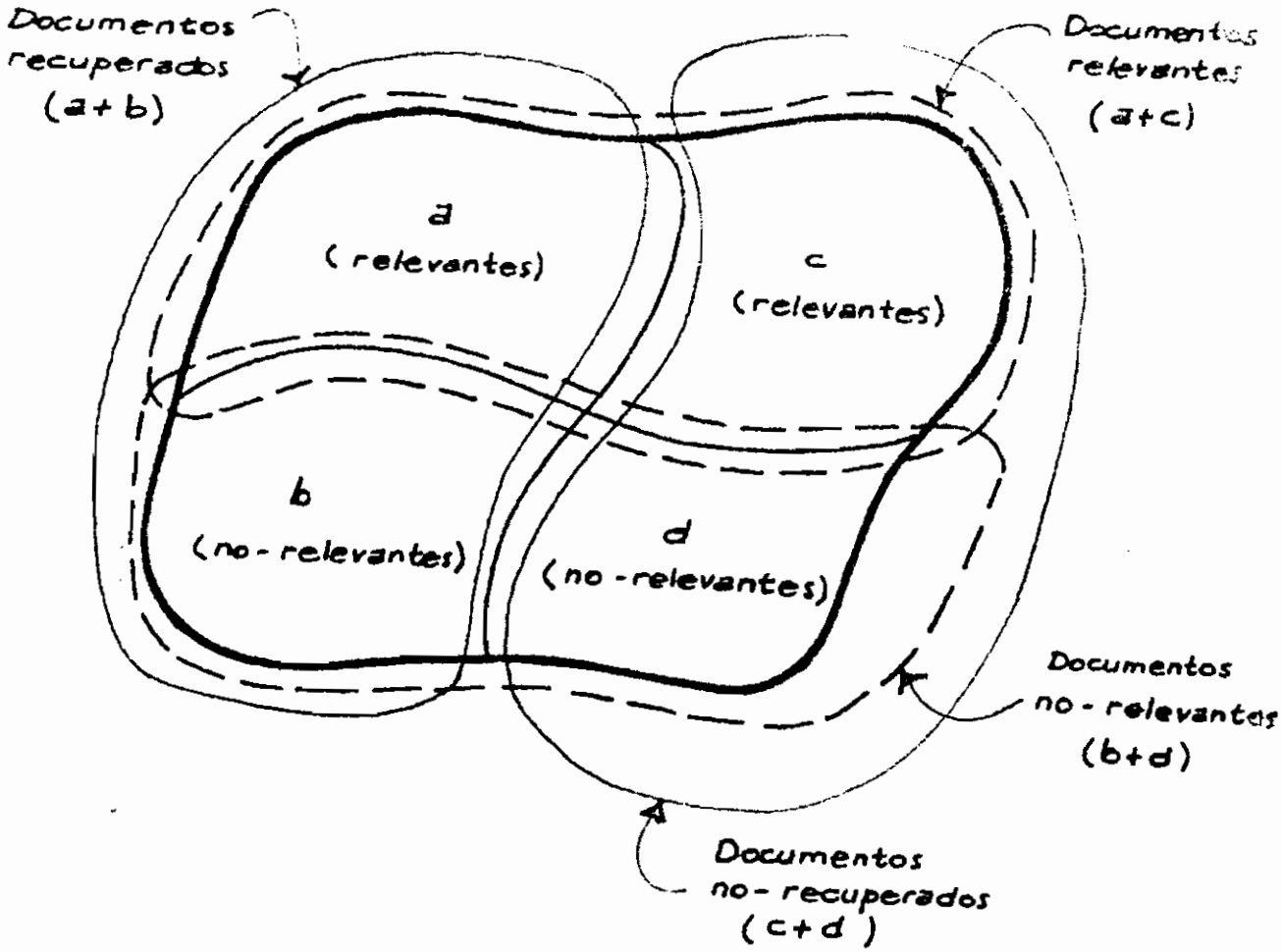


Fig. 3 La operación de un sistema de información como auxiliar del usuario



Medidas de eficiencia de un S.I.D.

$$\text{Recordación} = \frac{a}{a+c} = R \quad \text{(También sensitividad)}$$

$$\text{Presición} = \frac{a}{a+b} = P$$

$$\text{Eficiencia} = \frac{a+c}{a+b+c+d} = E_f$$

$$\text{Especificidad} = \frac{d}{b+d} = E_s$$

$$\text{Efectividad} = R + E_s = E_c$$

Fig. 4 Evaluación de un sistema de información documental.

Palabras

Documentos

$\tilde{C}$ =

	A	B	C	D	E	F
1	1	0	0	1	0	0
2	1	1	0	1	0	0
3	0	0	1	1	1	1
4	0	1	1	0	0	1
5	1	0	0	1	1	0
6	0	0	1	0	1	0
7	0	1	0	1	0	1
8	1	1	1	0	0	0

a) La matriz término-documento.

$$\tilde{q} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} ; \qquad \tilde{r} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 0 \\ 1 \\ 2 \\ 2 \end{bmatrix}$$

b) $\tilde{r} = \tilde{C}\ \tilde{q}$

Fig. 5 .- La recuperación mediante el empleo de palabras o términos clave.

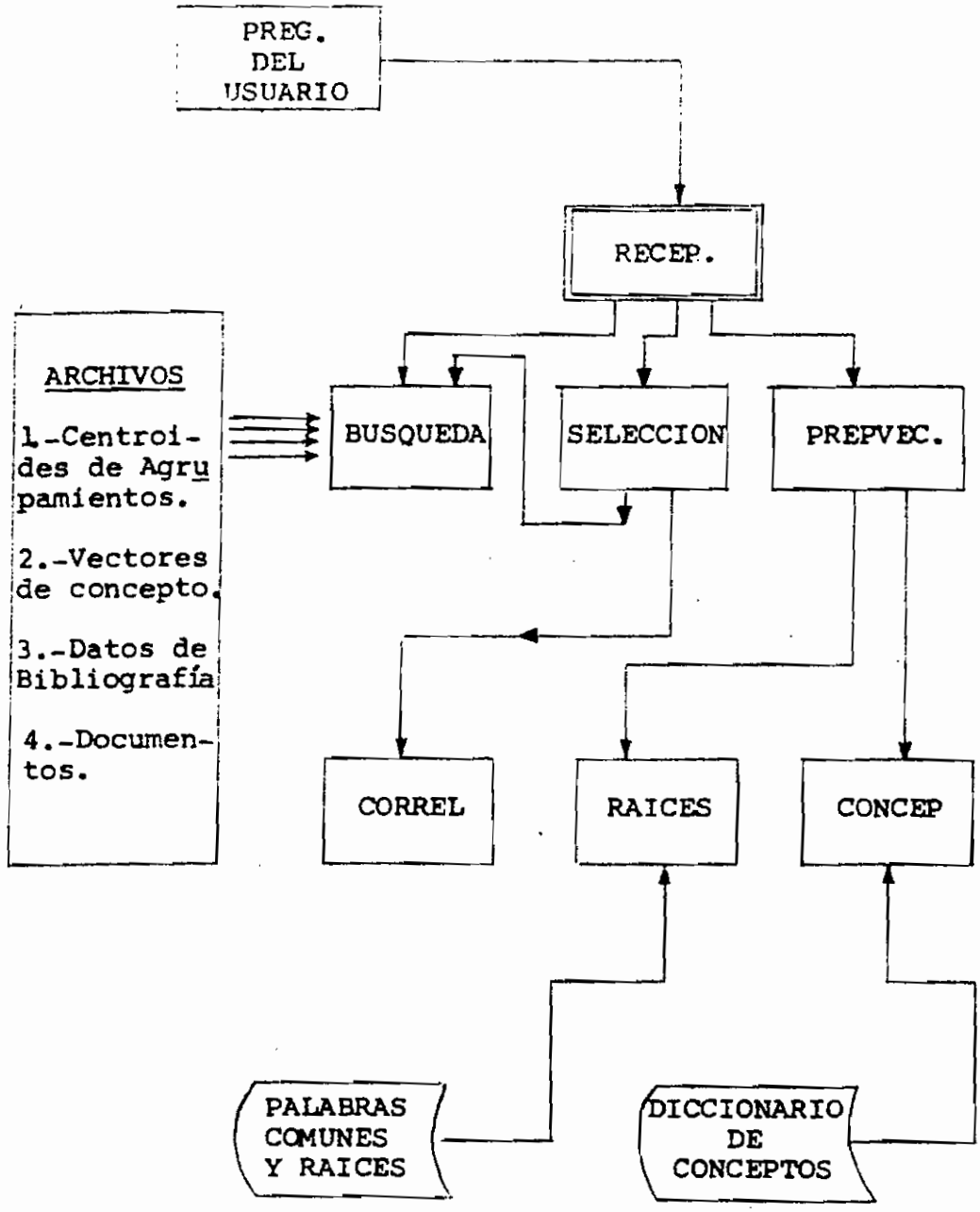


Fig. 6.- ORGANIZACION DE SARID.

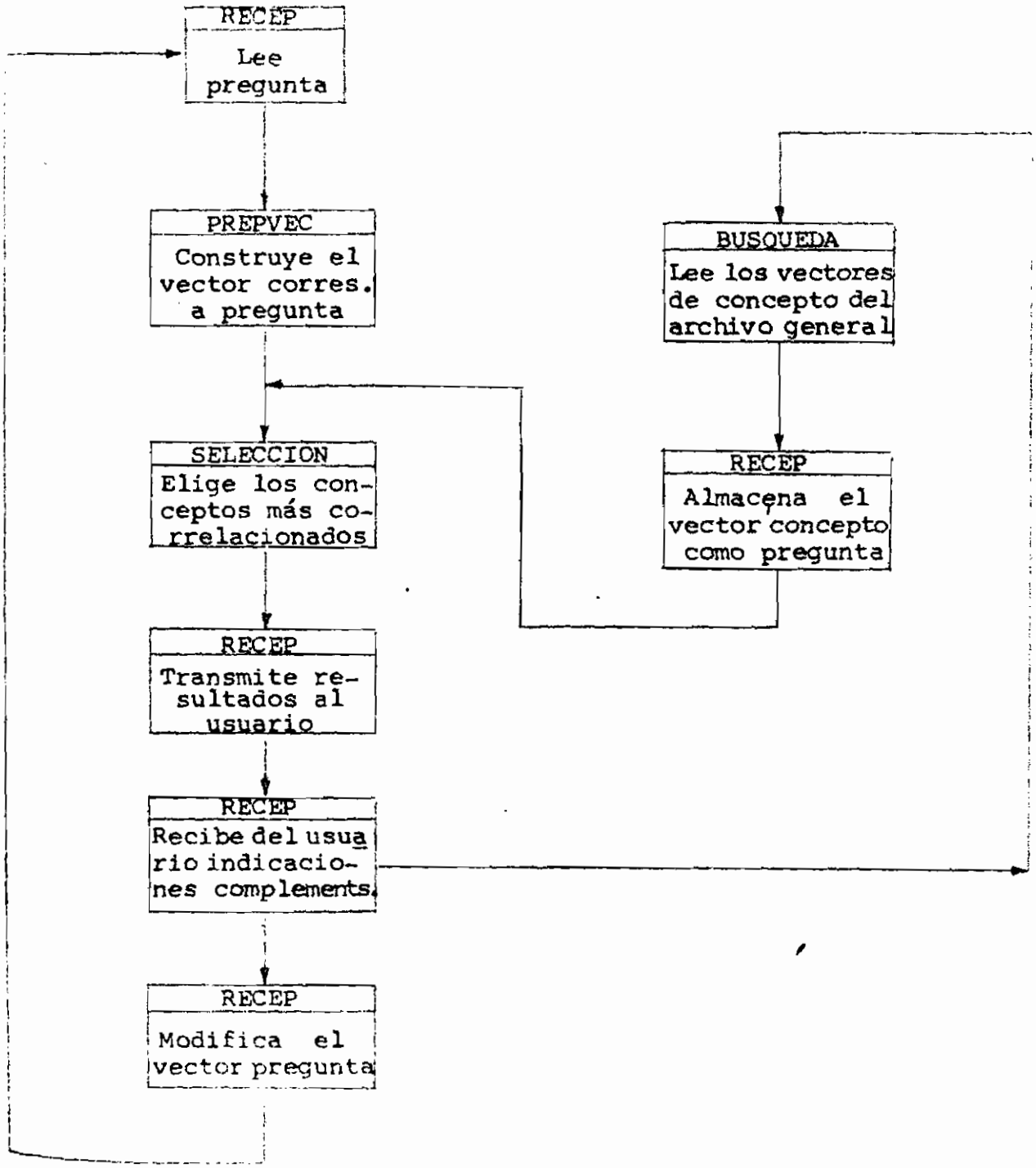


Fig. 7 -- OPERACION DE SARID

GIVE ALGORITHMS USEFUL FOR THE NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS AND PARTIAL -- DIFFERENTIAL EQUATIONS ON DIGITAL COMPUTERS. - - EVALUATE THE VARIOS INTEGRATION PROCEDURES (TRY RUNGE - KUTTA, MILNE-S METHOD) WITH RESPECT TO -- ACCURACY, STABILITY AND SPEED.

Fig. 8 .- Una pregunta Típica.

RANGO	No. DE REF.	CORRELACION
1	80	0.51
2	102	0.44
3	81	0.42
10	82	0.28
11	193	0.27
14	83	0.26
15	87	0.26
20	88	0.23
40	86	0.19
50	109	0.16
69	84	0.13
78	85	0.12

Fig. 9 . - Lista de documentos relevantes.

Recordación-Preción después de la recuperación de X documentos

X	Recordación	Precisión
1	0.0833	1.000
2	0.1667	1.000
3	0.2500	1.000
9	0.2500	0.3333
10	0.3333	0.4000
11	0.4167	0.4545
13	0.4167	0.3845
14	0.5000	0.4286
15	0.5833	0.4667
19	0.6667	0.3684
20	0.6667	0.4000
39	0.7500	0.2051
40	0.7500	0.2250
49	0.8333	0.1837
50	0.8333	0.2000
68	0.9167	0.1470
69	0.9167	0.1594
77	0.9321	0.1428
78	1.000	0.1538

Fig. 10a .- Datos de la medida Recall-Precisión.

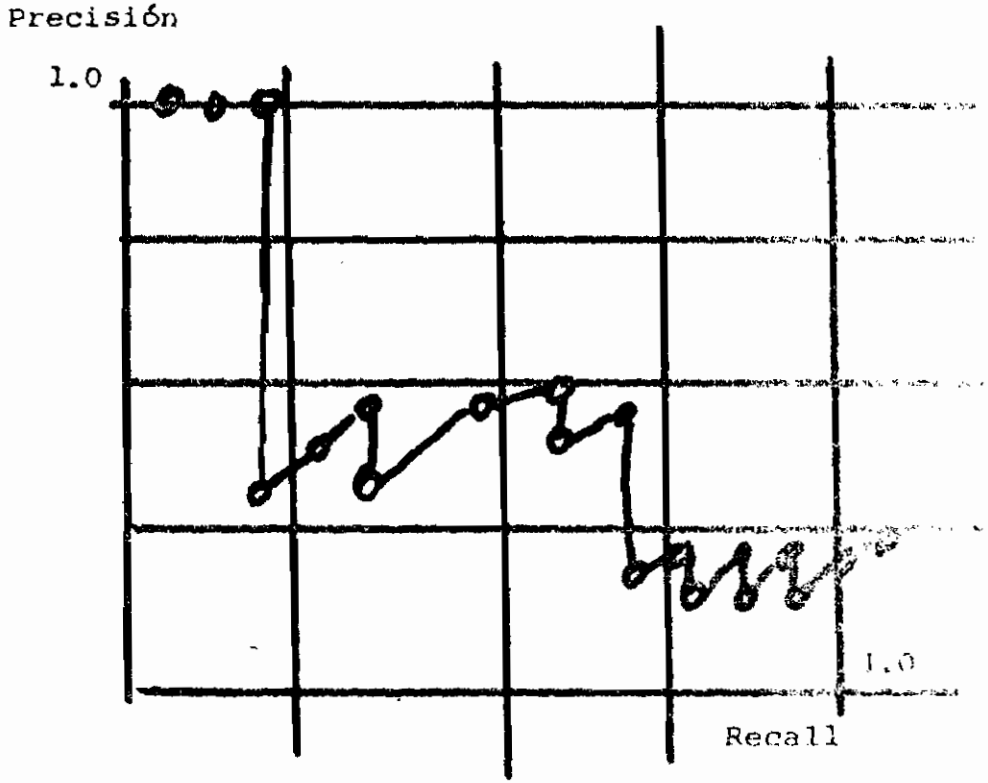


Fig. 10b. - Gráfica.

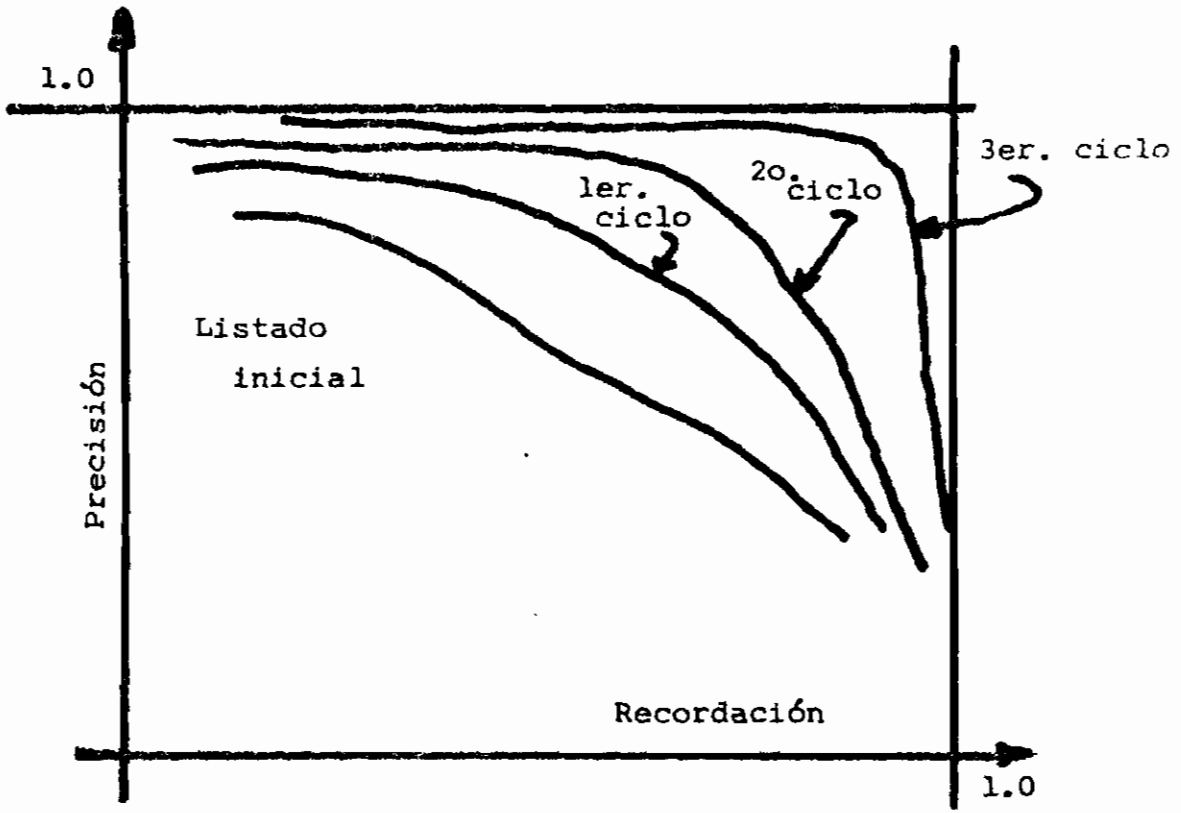


Fig. 11 . - El efecto de la retroalimentación de relevancia.