Recuperación de Información en el contexto de la Ciencia de la Computación

TC
Volumen 6, Número 4

Edgar Casasola Murillo
Universidad de Costa Rica, Escuela de Ciencias de la Computación e Informática
casasola@ecci.ucr.ac.cr

Resumen: *Este artículo presenta un vistazo integral y actualizado del área de Recuperación de Información desde la perspectiva de la Ciencia de la Computación. El principal aporte de esta investigación es describir en un marco sistematizado de referencia, los fundamentos de la disciplina, relacionando entre sí temas como: modelos matemáticos, algoritmos, técnicas, y aplicaciones recientes. Se presenta un análisis de los logros alcanzados en el campo, los nuevos retos y las tendencias actuales de investigación en el campo, principalmente los relacionados con el desarrollo del WWW. Este artículo va dirigido tanto a estudiantes, docentes e investigadores en computación, como a profesionales interesados en el estudio de modelos de estructuración, almacenamiento y recuperación automática de información.*

Palabras Clave: *Recuperación de Información, Modelos, Técnicas, Algoritmos, WWW*

1- Introducción

La Recuperación de Información o **RI** tiene como objetivo que el usuario logre satisfacer su necesidad de información, permitiéndole encontrar aquellos documentos que considera relevantes entre una colección de los mismos. Aparte de la recuperación misma, también se centra en la estructuración y en el almacenamiento de la información.

Su origen se remonta a técnicas como las utilizadas antes de Cristo por los Egipcios, Griegos y Romanos quienes organizaban sus documentos en forma de papiros mediante el uso de índices, clasificación alfabética y grupos temáticos. El término "índice" se refería al nombre dado a las tiras de tela que colgaban del extremo de los papiros. Cada tira contenía una descripción del documento y en algunos casos de su autor [16]. Desde el punto de vista de la Ciencia de la Computación podemos decir que los orígenes de la **RI** vienen del trabajo publicado por Vannevar Bush [3] quien basado en un ensayo previo de 1938, plantea, entre otras cosas, la necesidad de automatizar el proceso utilizado para administrar y accesar información. Bush menciona además el uso de hipertexto y la importancia de la compresión, y anuncia que algún día se podrían almacenar colecciones de documentos como la Enciclopedia Británica en dispositivos no más grandes que una caja de fósforos. En los años posteriores a la Segunda Guerra Mundial surge como campo de estudio formal, tomando fuerza cuando autores como Gerarld Salton [13] y Cleverdon [5] publican artículos científicos relacionados a la recuperación automática de documentos indexados, y proponen nuevos

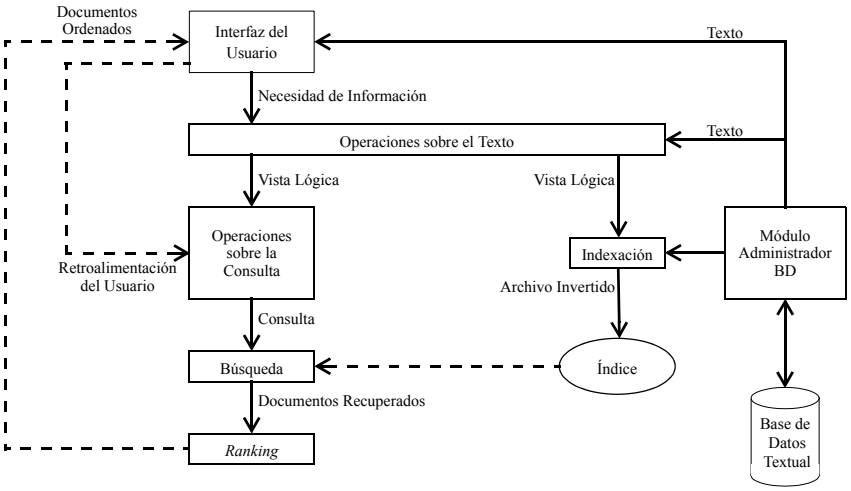

Figura 1. El proceso de recuperación de información. Tanto la consulta del usuario como los documentos son procesados para crear de forma automática vistas lógicas.

modelos, algoritmos y métricas para evaluación de efectividad y rendimiento. Las principales aplicaciones se centraron en la creación de colecciones de documentos con información bibliográfica, pero las aplicaciones comerciales no hicieron uso extensivo de los avances en el campo. Parecía ser que, en ese momento, era más importante el tamaño de las colecciones de documentos, que proveer sistemas eficientes para recuperar la información.

El advenimiento de la WWW volvió a poner a la **RI** en la mira de la comunidad científica dedicada a la computación. La exponencialmente creciente cantidad de información disponible tiene características similares a las colecciones de documentos con los que se trabajaba en el campo de la **RI**. Sin embargo, las diferencias que presenta la Web en cuanto a volumen y volatilidad, motivaron la reactivación de la investigación en el campo. Estas características fueron el producto de la democratización del acceso y publicación de información en este nuevo medio.

Este artículo retoma los fundamentos teóricos, algoritmos y técnicas clásicas que aún después de los años continúan vigentes y son el pilar fundamental de la disciplina. Aquí se mencionan los principales logros y nuevos retos que han motivado el renacimiento de la **RI** como campo de investigación. Inicialmente, se plantean los aspectos que giran alrededor de la llamada necesidad de información del usuario, posteriormente se mencionan los componentes del proceso de recuperación de información. Los siguientes temas incluyen: modelos formales, el modelo vectorial, filtrado, recuperación de información en el WWW, incluyendo la importancia actual de los motores de búsqueda y el uso de técnicas como el *Page Ranking*. Finalmente, se presentan algunas conclusiones importantes relacionadas con las líneas actuales de investigación en el campo.

2- La necesidad de información del usuario

Satisfacer la necesidad de información del usuario es el objetivo primordial de **RI**. Esto va más allá de la búsqueda determinística de datos típica de los sistemas de bases de datos. Hay que tener presente que la relevancia que cada usuario otorga a un resultado depende de su propia necesidad de informa-

ción, por lo que no puede existir un conjunto óptimo de resultados para cada necesidad. El ejemplo típico es el siguiente: le pedimos a varias personas que clasifiquen el mismo bloque de documentos por relevancia con respecto a un tema. Es poco probable que los ordenamientos coincidan entre sí, ya que no existe una clasificación ideal. Así, un sistema de **RI** se limita a ofrecer, basado en algún modelo, una serie de resultados considerados relevantes, y que son seleccionados acorde a consultas. En algunos casos los resultados son ordenados de mayor a menor en forma de *ranking*.

La representación o *vista lógica* de una necesidad de información se manifiesta como una consulta compuesta de términos, lo que implica que un mal planteamiento de la consulta infiera en una mala calidad de los resultados obtenidos. De acuerdo a la Figura **1**, el mismo procesamiento que se lleva a cabo con las consultas para convertirlos en una representación o vista lógica se aplica a los documentos. Esta transformación incluye la aplicación de algoritmos típicos de eliminación de "*stop words*" (términos o palabras que por ser tan frecuentes son consideradas de poca importancia) en una colección de documentos. Este proceso afecta el volumen de la colección ya que el tamaño de los índices se puede reducir a la mitad del volumen original.

Otros algoritmos comunes son los utilizados para llevar a cabo lematización o *stemming* que pretende reducir las palabras a su raíz. La lematización permite unificar términos sin importar su género, conjugación, o si es singular o plural. Los algoritmos más comunes de lematización incluyen:

- Eliminación de sufijos.
- Basados en reglas como el algoritmo de lematización de Porter
- Agrupación de términos utilizando similitud sintáctica, como el coeficiente de Dice aplicado a los bigramas de las palabras
- Basados en diccionarios.

Dichos algoritmos son comunes en los libros actuales de **RI** tales como el de Baeza [2] y forman parte del preprocesamiento de texto. Lo importante es notar que la mayoría de estos algoritmos implícitamente están diseñados para aplicarse dentro del contexto del modelo más común actualmente utilizado, el modelo vectorial.

En algunos casos, tal y como se muestra en la Figura **1**, se puede utilizar la retroalimentación provista por el usuario para refinar las consultas originales con el fin de obtener mejores resultados. La retroalimentación por relevancia de Rochio [11] es un ejemplo de este tipo de técnicas. Al igual que en los casos anteriores, estas técnicas presuponen que la vista lógica de las consultas se basan en representaciones vectoriales de pesos.

La consulta es procesada por un motor de búsqueda que con la ayuda de un índice invertido obtiene y clasifica los documentos donde aparecen los términos presentes en la consulta, estableciendo como resultado un orden o *ranking* entre ellos.

3- Recuperación de Información en el WWW

La exploración de la WWW mediante el seguimiento de enlaces llevó a nuevos mecanismos de recolección de recursos disponibles para indexación. El uso de arañas o agentes automatizados de recolección de documentos es común en este medio. La exploración de la WWW y el mejoramiento de las técnicas de exploración han sido ampliamente estudiados y han evolucionado desde modelos de exploración en ancho primero hasta modelos dirigidos tales como los presentados en [8].

Cabe aclarar que a pesar de contar con motores de búsqueda, el usuario continúa invir-

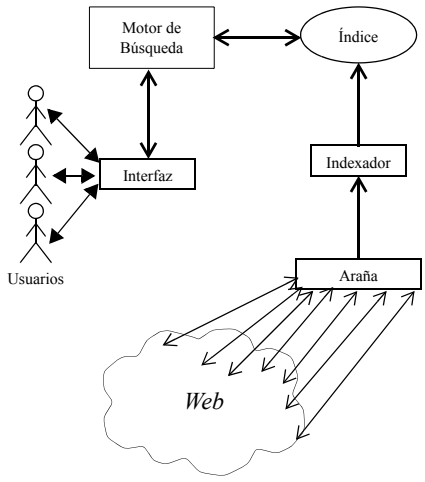

Figura 2. Estructura típica de un indexador automáticos para recolección de documentos en la WWW. Figura basada en Diagrama 13.3 de [2]

tiendo una cantidad importante de tiempo en labores de recolección de información, iterando entre búsqueda y navegación. El resultado final de esta combinación es una serie de documentos en formatos específicos que fueron recolectados al ir navegando por el WWW. Actualmente, se pretende evaluar el impacto de *agentes inteligentes* para llevar a cabo estas tareas en función de un usuario humano.

4- Caracterización formal de los modelos de Recuperación de Información

Los modelos de **RI** pueden ser caracterizados formalmente, según [2], como una tupla de cuatro elementos <D, Q, F, $R[q_i,d_j]$> donde:

- D es el conjunto de representaciones o vistas lógicas de los documentos en una colección.
- Q es el conjunto de representaciones o vistas lógicas de las consultas.
- F es un marco para modelar representaciones de los documentos, consultas y sus relaciones
- $R[q_i,d_j]$ es una función de clasificación por *ranking* que asocia un valor real al par compuesto por una consulta i y un documento j.

El modelo de espacio vectorial

En Salton [12] se introduce un modelo que representa a cada documento como un vector en un espacio n dimensional, donde n es la cardinalidad del conjunto de términos posibles. Cada consulta posible se representará también como un vector de n dimensiones, y se presume que existe un grupo de documentos *cercanos* que podrían ser relevantes a la consulta. En este modelo D y Q son espacios vectoriales, F es el álgebra vectorial y $R[q_i,d_j]$ podría ser la función coseno entre vectores.

Por tanto, la vista lógica d_j del documento j es de la forma $[w_{0j}, w_{1j}, w_{2j}, ..., w_{nj}]$. Cada entrada w_{ij} es un peso calculado como $tf_{ij}*idf_i$, donde tf_{ij} es alguna variación de la frecuencia del término i en el documento j, mientras que idf_i es la *rareza* del término i dentro de la colección que se indexa, calculada como:

$$idf_i = \log_2 \frac{N}{n_i} + 1$$

donde N es el total de documentos en la colección y n_i la cantidad de documentos en la colección en el que el término i aparece.

Similarmente, la consulta q es de la forma $[w_{0q}, w_{1q}, w_{2q}, ..., w_{nq}]$, con w_{iq} definido como:

$$w_{iq} = \left(0.5 + \frac{0.5 \times freq_{iq}}{max_l \times freq_{lq}}\right) \times idf_i$$

donde $freq_{iq}$ es la frecuencia del término i en la consulta q, y $freq_{lq}$ es la frecuencia máxima de algún término l en la consulta q. Las fórmulas anteriores son de las más utili-

zadas y fueron definidas por Karen Spark Jones y Salton respectivamente [14].

Dado que d_j y la consulta q son vectores de n dimensiones, podemos calcular el coseno del ángulo entre ambos vectores, lo que hace que si d_j y q son similares el valor se aproxima a 1 y en caso contrario, si tienen pocos términos en común, se aproxima a 0. Así la similitud entre d_j y q es:

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{\left\|\vec{d_j}\right\| \times \left\|\vec{q}\right\|}$$

o sea:

$$sim(d_j, q) = \frac{\sum_{i=0}^{n} (w_{ij} \times w_{iq})}{\sqrt{\sum_{i=0}^{n} (w_{ij})^2} \times \sqrt{\sum_{i=0}^{n} (w_{iq})^2}}$$

Note que sólo es necesario tomar en cuenta los pesos diferentes a cero para el cálculo, lo cual tiene una implicación importante para la implementación del modelo.

El **filtrado** puede ser visto como una adaptación del modelo formal de **RI**. Aquí la necesidad de información del usuario se mantiene mientras que los documentos son los que entran en el sistema de manera dinámica para ser clasificados como relevantes o no relevantes. En la Figura **3** se puede observar en el punto 1 como la colección de documentos se compone de dos macro documentos, el perfil de documentos relevantes y el perfil de documentos no relevantes, estos perfiles son los vectores resultantes del producto cruz entre los documentos relevantes y no relevantes respectivamente. El documento a ser juzgado toma el lugar de la consulta cuya similitud será evaluada contra los dos perfiles tal como se muestra en el punto 3. Si la similitud con el perfil relevante está por encima de la similitud con respecto al no relevante el documento evaluado será considerado relevante. Si la diferencia entre ambos documentos es menor a un cierto umbral el filtro se abstiene de juzgarlo. De esta forma los modelos clásicos de **RI** se han generalizado para aplicarse a problemas de ordenamiento, filtrado, agrupamiento y ruteo de información [4].

A pesar de que se han desarrollado otros modelos, el vectorial es el más popular por su fácil implementación y resultados relativamente aceptables en términos comparativos. Otros modelos clásicos existentes son el booleano y el probabilístico. Los más recientes involucran el uso de redes de inferencia [15], conjuntos difusos, vectores de vectores, sin embargo el rendimiento obtenido versus complejidad de los mismos sigue inclinando la balanza a favor del modelo vectorial, por lo que los nuevos desarrollos, casi sin omisión, se orientan a este modelo. Por ejemplo, la retroalimentación por relevancia estándar de Rochio [11] genera una nueva consulta al recalcular los pesos del vector de consulta original aumentándolos o disminuyéndolos de acuerdo a las frecuencias de los términos en los documentos considerados relevantes y no relevantes que fueron obtenidos.

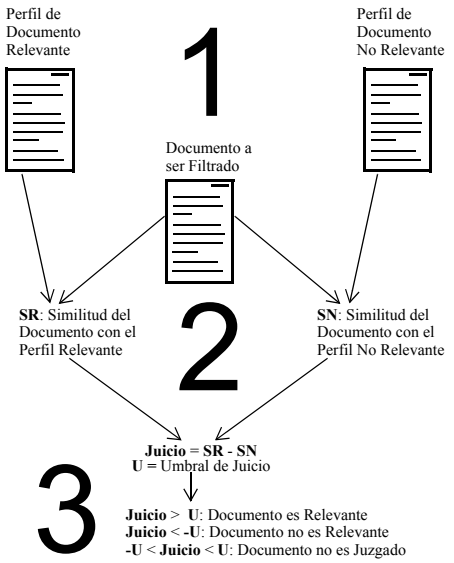

Figura 3. Filtrado visto como una generalización de un modelo formal de recuperación de información.

Tabla 1. *Porcentajes de Utilización de Motores de Búsqueda en el WWW a Julio del 2006.*

Motor de Búsqueda	%
Google	**49.20**
Yahoo Search	**23.80**
MSN Search	**9.60**
AOL Search	**6.30**
Ask.com	**2.60**
My Way	**2.30**
Earth Link	**0.60**
iWon	**0.60**
Netscape Search	**0.50**
Dogpile	**0.40**

Implícitamente se utilizan modelos algebraicos vectoriales para llevar a cabo cálculos utilizados en algoritmos de agrupamiento, clasificación, y sumarización. Todos derivados del cálculo de similaridad entre documentos propio del modelo vectorial. Incluso el algoritmo de *ranking* utilizado por los motores de búsqueda en Internet más modernos reutilizan un modelo que cuenta ya con casi cuarenta años de existencia.

En cuanto a la implementación del modelo, la prevalecencia del índice invertido como estructura preferida para creación de bases de datos de documentos textuales sigue predominando. El uso de sistemas de bases de datos para la implementación de índices en **RI** no ofrece resultados aceptables, si se piensa en índices que indexan tablas compuestas por un solo término esto resulta obvio. La existencia de la tabla en sí no tiene sentido en este caso. Además la estructura de los índices invertidos se presta para aplicar algoritmos de compresión específicos dentro de los cuales predominan el nuevo algoritmo de Huffman orientado a bytes cuya tasa de compresión cae cerca de la entropía, y algunos otros que permiten llevar a cabo búsqueda sobre texto comprimido, o aplicar variaciones de los algoritmos de búsqueda de texto en texto como los de Boyer More Hospol y el Shift-Or.

Luego de la creación de Google, la popularización de sistemas de **RI** se canaliza a través del uso de motores de búsqueda. En Estados Unidos, casi la mitad de las consultas con motores de búsqueda fueron hechas con Google, tal y como se muestra en la Tabla **1** [9]. El éxito de este motor viene de la indexación de lo popular, de su interfaz minimalista poco cargada de imágenes o animaciones, y por supuesto de su tiempo de respuesta. El algoritmo de *Page Ranking* [10] se basa en el establecimiento de un ranking absoluto entre los documentos basados en la cantidad de enlaces entrantes hacia los mismos. Esta medida de popularidad no es ideal, pero ha ofrecido buenos resultados.

En todo caso, de existir alguna información mejor que la que encuentra un usuario, de la que nadie sabe, que nadie puede encontrar, y es poco o nada referenciada, la misma no va a ser echada de menos por la mayoría de los usuarios. Lo importante es que el viejo modelo vectorial sigue siendo utilizado, y las opciones siguen utilizando los modelos más tradicionales: por ejemplo, la funcionalidad "*más documentos como este*" es una aplicación típica de retroalimentación por relevancia pero utilizando solo retroalimentación positiva proveniente de un documento.

Los nuevos avances van de la mano con el conocimiento que se tiene de la relación existente entre documentos, el vocabulario utilizado, los usuarios, su necesidad de información y su interacción con los sistemas de **RI**. La Minería de Consultas, descendiente de la Minería de Uso de la Web [1], utiliza la información obtenida de las bitácoras de consultas llevadas a cabo en motores de búsqueda para modificar la forma en que estos organizan la información, identificando grupos de consultas semánticamente relaciona-

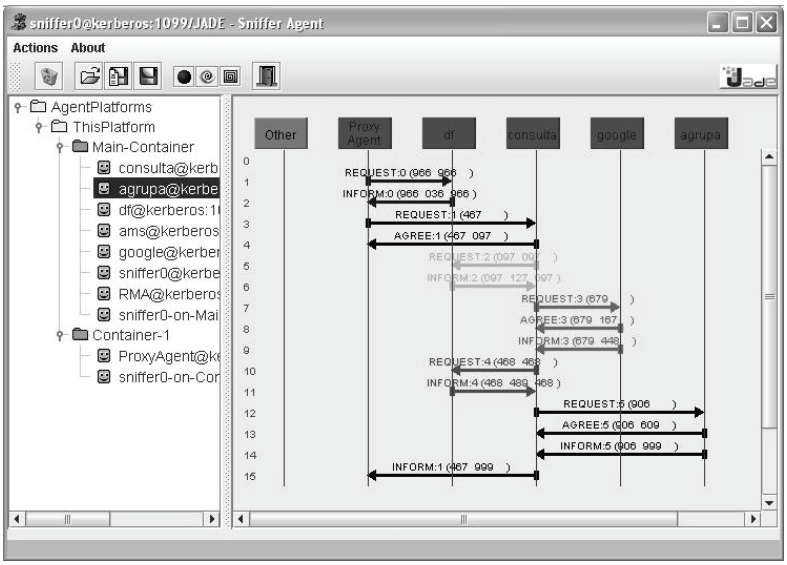

Figura 4. Ejemplo de una aplicación de un Sistema Multi-Agente para Filtrado de Información Proveniente del WWW.

das, para transformar consultas a categorías de búsqueda para que sin retroalimentación explícita y en forma automática y anónima el usuario pueda retroalimentar a través del historial de búsquedas.

Otros apuestan a tecnologías tales como el uso de marcado semántico con lenguajes como OWL [6] o al uso de ontologías [7] para construir el Web Semántico y permitir que Agentes Inteligentes lleven a cabo las tareas de **RI** en nombre del usuario. En este caso lo complejo es la tarea de extracción y formalización de la información no estructura proveniente del Web por lo que otros más conservadores tratan de utilizar Agentes Inteligentes en combinación con los algoritmos clásicos de **RI**.

Actualmente, en la Escuela de Ciencias de la Computación de la Universidad de Costa Rica hay varias investigaciones en proceso que experimentan con el uso de sistemas multi-agente y agentes inteligentes de **RI** y filtrado personalizado de información proveniente del WWW. Se espera que el uso de agentes disminuya el tiempo que invierten los usuarios al recolectar información proveniente de la Web, y que los resultados que se obtengan en áreas como la minería de consultas será de gran utilidad para guiar los esfuerzos de trabajo de los agentes. En el ejemplo mostrado en la Figura **4**, una serie de agentes con labores especializadas utilizan los servicios de otros agentes para sacar adelante una tarea de **RI** más compleja.

5- Conclusiones

La Web llevó al resurgimiendo la investigación en **RI**, y sin dejar de cumplir con las expectativas originales planteadas por Bush se han logrado desarrollar teorías y algoritmos cuyas aplicaciones han producido buenos resultados. Modificaciones a los algoritmos y modelos clásicos han permitido crear herramientas como por ejemplo el motor de búsqueda Google, que con la indexación de lo popular, su interfaz minimalista y su buen tiempo de respuesta ha sido ampliamente aceptado entre los usuarios de la Web. Sin embargo, no se puede decir que motores como Google sean la solución a todos los problemas de **RI**, ya que los usuarios continúan invirtiendo gran cantidad de tiempo y esfuerzo para recuperar informa-

ción, y algunos incluso aún no pueden dar con ella al no poder plantear en forma de consultas su necesidad de información.

Estamos en un momento donde los algoritmos clásicos y técnicas desarrollados desde hace más de cuarenta años continúan siendo los fundamentos primordiales del trabajo realizado en el presente. Los nuevos esfuerzos apuestan por tecnologías para automatizar aún más el proceso mediante el uso de Agentes Inteligentes, el Web Semántico, y la minería de uso del Web.

Agradecimientos

Se agradece a los estudiantes Didier Cerdas, Rodrigo Bartels, Kryscia Ramírez y Mauricio Ulate por su colaboración durante la elaboración de este artículo, y especialmente al Dr. Francisco Torres por su constante motivación y gran paciencia.

Referencias

[1] **Baeza-Yates, Ricardo**. "*Applications of Web Query Mining*". Proc. of the 27th European Conf. on IR Research, ECIR 2005, Vol. 3408. ISBN: 3-540-25295-9. Santiago de Compostela, Spain, March 21-23, 2005.

[2] **Baeza-Yates, R., Ribeiro-Neto, B.**, "*Modern Information Retrieval*". Addison-Wesley, Wokingham, UK, 1999

[3] **Bush, Vannevar**, "*As We May Think*", Atlantic Monthly 176,1, pp. 101-108, 1945.

[4] **Casasola, Edgar**. "*ProFusion Personal Assistant: An Agent for Personalized Information Filtering on the WWW*". Master's Thesis. University of Kansas, Lawrence, Kansas. 1998.

[5] **Cleverdon, C.W., Mills, J. and Keen, E.**, "*Factors Determining the Performance of Indexing Systems*", ASLIB Cranfield Research Project. 1966

[6] **Daconta, M.C., Obrst, L.J. & Smith, K.T.**, "*The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management*". 1era ed., Editorial Wiley. Indianápolis, Indiana, 2003.

[7] **Dean, M. & Schreiber, G., Editores**. "*OWL Web Ontology Language Reference*". W3C Recommendation, 10 Febrero 2004. www.w3.org/TR/2004/REC-owl-ref-20040210/

[8] **Hersovici, M., Jacovi, M., Maarek, Y. S., Pelleg,D., Shtalhaim, M. and Ur, S.** "*The shark-search algorithm --- An application: Tailored Web site mapping*". In Proc. 7th Intl. World-Wide Web Conference, 1998

[9] **Nielsen**, "*NetRatings Inc. Anounces U.S. Search Share Rankings*", July 2006. www.nielsen-netratings.com/pr/pr_060821.pdf

[10] **Page, L., Brin S., Motwani, R., T. Winograd**. "*The PageRank Citation Ranking: Bringing Order to the Web*". Stanford Digital Library Technologies Project, 1998.

[11] **Rochio, J.J.** "*Relevance Feedback in information retrieval*". In G. Salton, editor, The SMART Retrieval System, Experiments in Automatic Document Processing. Prentice Hall Inc. Englewood Cliffs, NJ, 1971

[12] **Salton, G., Wong, A., and Yang, C.S.** "*A vector space model for automatic indexing*". Communications of the ACM, 1975

[13] **Salton, G.**, "*Manipulation of trees in information retrieval*". Commun. ACM 5[2]: 103-114, 1962

[14] **Spark Jones, K.**, "*A Statistical Interpretation of Term Specificity and Its Application in Retrieval*", J. Documentation. 28 (1), 11-20 1972

[15] **Turtle, H., and Croft, W.B.,** "*Inference networks for document retrieval*". In J.L. Vidick [Ed.], Proc. of the 13th Int. Conf. on Research and Development in Information Retrieval, pp. 1-24, ACM, New York, 1990

[16] **Wellisch, Hans.**, "*The Oldest Printed Indexes*", The Indexer 15 [2], pp. 73–82, 1986.

Edgar Casasola Murillo *obtuvo su Bachillerato en Ciencias de la Computación e Informática de la Universidad de Costa Rica, su Maestría en Ciencias de la Computación en la Universidad de Kansas en Lawrence. Actualmente es profesor e investigador en la Escuela de Ciencias de la Computación e Informática de la Universidad de Costa Rica.*