Sistemas de recuperación de información

Sonia Ordóñez Salinas[1]

Fabio A. González O.[2]

RESUMEN

Este documento presenta una revisión de los principales aportes que se han hecho en el tema de los sistemas de Recuperación de Información (RI). Dado que la eficiencia y el desempeño de dichos sistemas depende de varios subsistemas y que cada uno de ellos ha ido creciendo y sufriendo cambios de manera independiente, esta revisión discrimina a grandes rasgos los sistemas de recuperación de información en 4 grandes temáticas a saber: Representación de documentos y consultas, estructuras de datos, selección de documentos relevantes y eficiencia de los Sistemas de Recuperación. Por último y a partir de los diferentes documentos estudiados se plantea el trabajo futuro basado en la semántica.

Palabras clave: Sistemas de Recuperación de Información, representación de documentos, categorización de documentos.

ABSTRACT

This document presents a survey of the main contributions that has been made in the topic of the Information Recovery systems (IR). Since the efficiency and the performance of this systems depend on several subsystems that have been growing and suffering changes in an independent way, this survey is organized around four broad thematic areas: representation of documents and queries, data structures, selection of relevant documents and efficiency of Recovery Systems. Finally IRS based en semantics are discussed.

Key words: Information Recovery Systems, documents representation, documents categorization.

I. INTRODUCCIÓN

Con la posibilidad de almacenar documentos u otro tipo de objetos diferentes a registros, el problema de la gestión de la información y su recuperación, se ha convertido en un problema y por ende en tema de estudio.

El contar con una gran cantidad de información almacenada y poder recuperarla de manera precisa y rápida se estudia bajo el tópico de Técnicas de Recuperación. Sin embargo, en cuanto a técnicas de recuperación existe una clara distinción entre recuperación de datos y recuperación de información. Entre las diferencias principales debe señalarse que mientras en la primera se utilizan funciones de correspondencia exactas, que verifican si está o no un ítem dentro de un archivo o registro, en la recuperación de información se utilizan funciones que permiten que parcialmente se relacionen los documentos con el requerimiento y de estas se seleccionan las que mejor se ajusten con la consulta. En la recuperación de datos, la inferencia se da por deducción basada en reglas exactas, que permiten recuperar los datos siempre que este exista. La inferencia se hace con simples reglas deductivas de relación, es decir si aRb y bRc entonces aRc. Mientras que en la RI, la recuperación se hace por inducción, es decir que las relaciones se especifican con un grado de incertidumbre. De igual forma, el lenguaje de consulta para la recuperación de datos es artificial y debe ser completo, mientras que en la recuperación de información, el lenguaje de consulta es natural e incompleto [1], esto es, que depende del usuario.

Otro aspecto relevante que marca una diferencia entre la Recuperación de Datos y la Recuperación de Información es que el hecho que un objeto cuente con un atributo lo hace pertenecer a una clase. En el segundo caso, un objeto puede contar con cierto atributo que a simple vista lo haría pertenecer a una clase, y que sin embargo no pertenece. El poder afirmar si un objeto pertenece a una clase o no, depende no solamente de varios atributos, sino de varios factores.

Es así que, la recuperación de datos se implementa generalmente en los motores de datos tradicionales, mientras que la aplicabilidad principal de la recuperación de información se presenta en los llamados buscadores web.

El sistema de recuperación de información, para efectos de esta revisión se propone subdividirlo en seis subsistemas que permitan: a) registrar el documento; b) representar el documento; c) almacenar el documento; e) registrar el documento dentro de una estructura que facilite la búsqueda; f) seleccionar los documentos relevantes; g) representar la consulta y buscar los documentos relevantes. Es claro anotar que cada uno de estos subsistemas a su vez se puede dividir o mezclar, como en Lan Huang [2], donde divide la arquitectura de los sistemas de recuperación de información en un indexador, un rastreador y un servidor de consultas. El rastreador en este cado se encarga de coleccionar las páginas en el Web, el indexador procesa los documentos recuperados y los representa en una estructura de búsqueda eficiente. El servidor de consultas acepta las consultas de el

[1] **Sonia Ordóñez Salinas**
Directora Grupo de Investigación Gesdatos - Facultad de Ingeniería, Universidad Distrital.

[2] **Fabio A. González O.**
Profesor asociado Universidad Nacional de Colombia.

Artículo recibido en Abril de 2004
Aceptado en Junio de 2004

usuario y retorna las páginas de resultado consultando con las estructuras de búsqueda.

En la figura 1, se esquematiza un modelo general para la recuperación de información. Como se puede observar, los subsistemas del registro del documento, del almacenamiento del documento y de digitación o introducción de la consulta en un lenguaje natural, no requieren mayor complejidad, razón por la cual que no se hará referencia a ellos en esta revisión. De igual forma, el subsistema de búsqueda de documentos relevantes hace referencia a la ubicación de los documentos con base en una estructura, por lo que tampoco merece mayor atención.

Una vez que un investigador o en general el autor de un documento, lo registra en un sitio web, el documento se debe representar de tal forma que al compararse con una consulta se pueda, a través de algún indicador, saber si el documento es o no relevante a dicha consulta. Luego, el subsistema de selección de documentos relevantes utiliza las representaciones tanto de documentos como de consultas para realizar el proceso de relacionar sobre una misma unidad de medida.

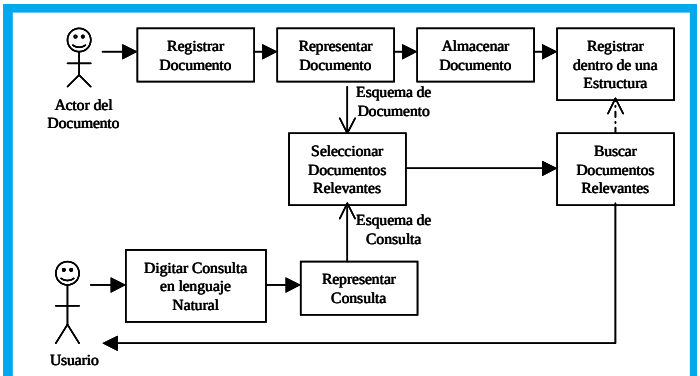

Figura 1

El documento debe ser registrado en alguna estructura que facilite la búsqueda física del mismo. Dicha estructura puede darse en términos clásicos como un índice o complicarse al incluir la representación de los documentos dentro de esta.

Un subsistema que no aparece dentro de la figura 1, y que merece especial atención, hace referencia a la forma como se mide la eficiencia del sistema de recuperación y que esta relacionada con la satisfacción del usuario al obtener los documentos que verdaderamente él espera.

Google por ejemplo, utiliza un rastreador web conformado por varios rastreadores distribuidos. Los servidores de URL envían una lista de URL´s para que los rastreadores la busquen. Las páginas encontradas son enviadas a un servidor de almacenamiento. Dicho servidor comprime y almacena las páginas dentro de un repositorio y les asocian un número de identificación. Un indexador se encarga de descomprimir los documentos de dicho repositorio y analizarlos gramaticalmente. Cada documento lo convierte en un conjunto de palabras de ocurrencias llamadas aciertos, es decir representa cada documento por dichas ocurrencias[2].

En particular esta revisión pone su atención a la representación de los documentos y consultas, las estructuras de datos, la selección de los documentos relevantes y la eficiencia del algoritmo en general.

II. ESTADO DEL ARTE

De acuerdo a lo explicado en la introducción, el desempeño de un sistema de recuperación de Información está marcado por varios aspectos, por lo que se tratará de mostrar en que estado está cada uno de los subtemas más significativos.

Representación de Documentos y Consultas

Dado que una de las partes álgidas de la Recuperación de la información, hace referencia a la representación de los documentos, se dedicará esta primera parte del estado del arte a conocer que propuestas existen en cuanto a la representación de los documentos.

Generalmente los documentos se representan a través de la frecuencia de aparición de palabras dentro del mismo y por ende se identificado un documento por sus palabras claves o términos. Esta representación se propuso por Luhn [3], quien se considera el pionero.

El uso de distribuciones estadísticas sobre las palabras claves, se introdujo con los trabajos realizados por Maron y kuhns [4] quienes permitieron introducir un diccionario como ayuda para la recuperación.

Jones [5] usa medidas de asociación entre las palabras claves, basada en sus frecuencias y su "co-ocurrencia", es decir la frecuencia con que una palabra clave ocurre dentro del documento y demuestra la eficiencia en el recalculo (proporción de documentos relevantes recuperados).

El trabajo de Rijsbergen [1], propone que un documento se represente por medio de una lista de clases, donde un documento podrá ser indexado a una clase, si la clase contiene una de las palabras claves del documento. Por lo que el procedimiento se realiza en tres partes: encontrar las palabras claves de más alta frecuencia, encontrar los sufijos y buscar las ramas.

En los años 90 se introducen técnicas de algoritmos genéticos con Klabbankoh [6] y Vrajitoru [7], donde el documento pasa a representar un cromosoma de

1s y 0s (presencia o ausencia de una palabra clave). El documento se presenta de acuerdo a un vector de palabras claves. Cada documento se representa como un vector de unos y ceros dependiendo de la presencia o ausencia de la palabra clave dentro del documento. Es claro que tanto el vector de palabras claves, como el de documentos tienen la misma dimensión. Tal representación corresponde a los cromosomas de un algoritmo genético.

En Shian [8], se propone la representación de documentos HTML, con una serie de parejas (término, peso). Cada peso se calcula en varias fases así: se le asigna un peso a cada uno de los marcadores del documento dependiendo de la relevancia del marcador, por ejemplo el titulo tendrá mucho más peso que un salto de línea (posiblemente no tenga peso). Por la frecuencia de una palabra dentro de cada uno de los parágrafos se le asigna un peso. El peso total del término, entonces, corresponderá a la combinación del peso del marcador y del peso del término dada su frecuencia. Dado que para representar un documento se pueden involucrar vectores de gran dimensión, se propone acotar de acuerdo a los valores más altos del peso.

Se ha trabajado la representación de las páginas que utilizan plantillas, incluyendo cada elemento de la página dentro de una base de datos como una tupla, tal como lo propone Arasu y Garcia [9]

Estructuras de Datos

Al tema de estructura de la información, realmente no se la ha puesto la importancia que se requiere, solo de manera experimental se ha mostrado que los tiempos de recuperación mejoran cuando se logran algunas estructuras. Generalmente, los documentos se almacenan de forma secuencial. Esta estructura se vuelve ineficiente en términos de tiempo, si varios usuarios simultáneamente realizan consultas.

Se buscan nuevas técnicas como la de Salton [10] que introduce el concepto de archivos invertidos y la técnica de Cluster[11] que ha dado resultados muy satisfactorios.

En Ruthven [12], se describe ampliamente diferentes estructuras de índices que se han empleado, multilistas, multilistas celulares, estructuras de anillos, árboles y tablas "hash". En Becker [13], se propone trabajar con índices invertidos incluyendo la ruta del documento.

Baldi [14] dentro de las estructuras para el manejo de índices profundiza sobre los índices invertidos y presenta una forma de reducir una de las entradas a estos, las localizaciones a las palabras. Una de las entradas de los índices invertidos es la localización de la palabra dentro del texto. Esta localización se puede representar por la diferencia que se da entre una y otra localización. Con relación a la misma temática de índices invertidos para los Sistemas de Recuperación de Información Melnik [15] va aún más lejos y propone trabajar con dichos índices pero distribuidos.

Dos temas que van muy relacionado con la estructura de los archivos es el método de clasificación que se utilice para lograr dicha estructura y las posibles transformaciones que puedan sufrir los datos antes se lograr una clasificación y por ende una estructura. Por lo que en muchas ocasiones se encuentran trabajos donde se unen dichos temas y es muy difícil hablar de uno solo por separado.

En cuanto a los métodos de clasificación, estos de han logrado por muy variados métodos. Como pioneros cabe resaltar a Good [16] y a Fairthorne [17] quienes sugirieron que la clasificación automática era muy útil para la recuperación de información. Desde entonces se han presentado varios trabajos que permiten crear grupos de documentos a través de clasificaciones, sin embargo, la mayoría se han dado a pequeña escala. En Rijsbergen [1] se puede encontrar todo un capítulo que trata sobre la clasificación a través de agrupamientos, dándole un manejo bastante riguroso, desde el punto de vista matemático, mostrando las diferentes propiedades ventajas y desventajas.

El tema se ha abordado desde un sin numero de teorías como por ejemplo, clasificación a través de métodos Bayesianos en Lewis [18] y Tzeras[19], Técnicas del vecino más cercano en Lam [20] y Masand [21], redes neuronales en Wiener [22], árboles de decisión en Apte [23], reglas de aprendizaje inductivo en Cohen [24], "Support Vector Machines" en Joachims [25] y Godbole [26], Técnicas de máxima entropía con Nigam [27] y Técnicas de "boosting" en Schapire [28].

En shian [8], se propone clasificar los documentos en clases jerárquicas, a través de máquinas de aprendizaje supervisado. La máquina parte de un grupo de documentos clasificados por humanos, y luego a través de una serie de subsistemas automáticamente hace la clasificación. En el trabajo se plantea refinar la clasificación a través de la técnicas de la minería de reglas de asociación para descubrir los términos asociados dentro de cada clase y así aumentar el conocimiento aprendido. Por último plantea un nuevo algoritmo Perfect Term Support (PTS) para eliminar términos no representativos explorando sus relaciones.

En Syan[29], se aborda el tema de la recuperación de documentos de hipermedia e hipertexto tal como HTML y XML. El trabajo plantea el problema de la recuperación de documentos almacenados físicamente separados y unidos a través de hipervínculos. El trabajo incluye dentro de su solución una estructura

"Web-strucutre" en conjunción con las palabras claves y grafos.

En Srikanth [30], se parte de trabajos previos que se han hecho para la recuperación de información a través de lenguaje natural ("NLP"), en lo que compete a la indexación de términos que representen los documentos y los trabajos recientes sobre "Statistical Language Models" (SLM), para representar los documentos a través de calificaciones generadas a partir de la consulta dada por el usuario. El trabajo en particular presenta una técnica basada en lenguaje nartural y SLM que permite analizar gramaticalmente una consulta y de acuerdo a este análisis se recuperan los documentos. En SkillSoft[31] se plantean construir esquemas de representación a partir del lenguaje natural. Con técnicas como maquinas de aprendizaje se puede elaborar una base de conocimiento y con redes neuronales aplicadas a la base del conocimiento se crean representaciones matemáticas que faciliten no solo el almacenamiento si no la búsqueda.

Becker [13] propone una estructura específica para almacenar los dominios y representar los documentos por palabras claves y árboles para documentos en formato XML. El diccionario de dominios es concebido como un conjunto de descriptores (o conceptos) conectados por relaciones jerárquicas, relaciones de equivalencia o relaciones de asociación y almacenados en un catálogo global de recursos. Cada documento representado a través de un árbol, es indexado por medio de sus hojas y el diccionario de dominios. Por otro lado, propone utilizar consultas especializadas de la especificación XML-QL, para la recuperación de documentos.

Baldi [14] y Hofmann[32] incluye el análisis estadístico de la semántica latente "Latent semantic indexing" (LSI) para estimar las estructuras que se esconden tras los conceptos. Dichos análisis se basan en las técnicas de descomposición de valores singulares "singular value descoposition" (SVD) que permiten descubrir los más importantes patrones asociativos entre palabras y conceptos. Hofmann[33] bajo esta misma técnica plantea realizar indexación

Ya recientemente en Li [34] se introduce el uso del análisis discriminante para encontrar transformaciones que reflejen las similitudes propias de los datos y así poder categorizar los textos.

Selección de Documentos Relevantes

Desde 1954, se han venido trabajando las medidas de asociación clásicas con Goodman en el 54 [35], Goodman en el 59 [14], Kuhns [36], Cormack [37], Sneath [38], Maron [4] y Salton [39]. Dichas medidas incluyen métricas como Dice´s, Jaccard's, Cossine, Overlap. A partir de entonces, estas medidas han sufrido muy pocas modificaciones, dentro de estas cabe desatacar la mencionada en Rijsbergen [1] que propone trabajar con un índice simple: el documento y la consulta se representa con ceros y unos (presencia o ausencia de una palabra clave dentro del documento X y presencia o ausencia de una palabra clave dentro de la consulta Y) es decir $|X \bigcap Y|$, sin tener en cuenta el tamaño de los vectores.

Existen otras medidas de asociación de tipo probabilística Maron [4] y Jardine [40], donde se tiene en cuenta la probabilidad de la presencia de una palabra clave dentro de un documento.

Las medidas de similitud trabajadas hasta el momento involucran la hipótesis de similitud entre un documento y una pregunta, en Becker [13] se cuestionan tal hipótesis a partir de los trabajos realizados por Simonnt [41] que presenta la sustentación de que la pregunta y el documento no juegan un papel simétrico en la búsqueda de la información y el trabajo de Nie[42] donde se expone que es necesario que el usuario exprese en su pregunta sólo dos criterios importantes: la exhaustividad de la pregunta en el documento y la especificidad del documento con respecto a la pregunta. Becker [13] extienden estas reflexiones e introduce dos nuevas medidas: una basada en la exhaustividad que estima el grado de pertenencia de la pregunta Qi en el documento Di, la otra medida basada en la especificidad, que tiene en cuenta el grado de pertenencia del documento Di a la pregunta Qi.

El tema en particular se ha tratado a través de programación y algoritmos genéticos. Klabbankoh2000[43] incluye tres nuevas funciones de aptitud ("fittnes") como medidas de distancia para determinar los documentos relevantes. Pathak [44] propone resolver la función de distancia a través de algoritmos genéticos y una combinación lineal de los ya existentes: Cosine, Jaccard, Dice y Overlap. Vrajitoru [45] introduce la utilización del operador crossover de algoritmos genéticos para recuperar los documentos relevantes y hace una comparación con los demás operadores mostrando la eficiencia del operador. Praveen [46], al igual que el anterior, resuelve la función de relación o distancia a través de un algoritmo genético.

Se han aplicado otras técnicas para sofisticar la búsqueda de los documentos relevantes como la minería de datos como Wang [47] en donde a través de estas técnicas se plantea extraer estructuras típicas de una colección de objetos semi-estructurados y Thuraisingham[48], que no solamente aborda algunas técnicas de minería para extraer características de los documentos, si no que plantea varios de los temas que tienen que ver con la recuperación de información.

Eficiencia de los Sistemas de Recuperación

Existe una gran cantidad de trabajos que proponen medidas de eficiencia de los sistemas de recuperación de información. Como pionero se cita a Cleverdon [49] en el trabajo presentado para el proyecto Cranfield en 1966. Posteriormente en Cuadra [50] se sugiere que dado un conjunto de documentos ordenados, la relevancia puede ser medida en una escala ordinal según la posición; posición que depende de factores externos y que no puede ser controlada en laboratorio. Salton [37] a través de las medidas de precisión (proporción de documentos recuperados que son relevantes) y recalculo (proporción de documentos relevantes recuperados) muestra que la utilización de variables dicotomicas con una probabilidad de error, permite caracterizar la relevancia o no relevancia de un documento. Medidas que aún hoy se utilizan para evaluar la eficiencia de un sistema de recuperación.

Ruthven [1] en sus trabajos aportan de manera experimental y formal que la razón por la cual un usuario selecciona un documento como relevante no necesariamente es la presencia de palabras claves. El trabajo se enfoca a determinar el por qué un documento se marca como relevante y demostrar que para cada consulta en particular se deben seleccionar un conjunto de características diferentes; utiliza la teoría de la combinación de la evidencia de Dempster-Shafer como herramienta para evidenciar las diferentes características del uso de términos en los documentos. Los pesos a los términos se da en términos de relaciones con otros términos.

III. TRABAJOS FUTUROS

El crecimiento desmesurado de documento en el WEB, ha llevado a que cada vez nos enfrentemos con un cliente más exigente, en el sentido que a la hora de realizar un consulta a través de cualquiera de los buscadores existentes, se espera encontrar solamente y únicamente aquellos documentos que realmente le aportan a su necesidad. A pesar del sinnúmero de técnicas estudiadas e implementadas para satisfacer al usuario, aún nos encontramos con el problema de poder interpretar a través de una consulta expresada en unas pocas palabras (relacionadas o no a través de una estructura semántica) el requerimiento del usuario con fidelidad.

Es así que se requiere buscar alternativas basadas en la semántica como por ejemplo búsquedas de meta datos contextuales o técnicas de semántica latente, que permitan encontrar exactamente lo que el usuario quiere buscar en un sentido o significado estricto. Las técnicas y metodologías que permitan extractar el significado real de unas palabras relacionadas es un tema de investigación. El desarrollo en esta área, no solo cambia las estructuras para almacenar y manejar la información, sino que, las medidas que permitan relacionar un documento con una consulta no se darían en términos de correlación por palabra sino correlación por estructura semántica o concepto.

REFERENCIAS BIBLIOGRÁFICAS

[1] Rijsbergen Van C.J. Information Retrievel. Departamente of Computing Science, University of Glasgow Second edition . 1.979

[2] Huang Lan, A survey on Web Information Retrieval Technologies, Computer Science Department, State University of New York at Stony Brook, Stony Brook, NY 11794-4400, 2003.

[3] Luhn, h.p "A statistical approach to mechanized encoding an seraching of library information" IBM journal and Research and Development,1,309-317 (.1957)

[4] Maron, M.E. and KuhnS, J.L., 'On relevance, probabilistic indexing and information retrieval', Journal of the ACM, 7, 216-244 (1960).

[5] Sparck Jones, K. "Automatic Keyword Classification for Information Retrieval", Butterworths, London 1971

[6] Klabbankoh Bangorn, Pinngern PH.D. Applied Genetic Algorithms in Information Retrievel Faculty of Information Technology, King Mongkut´s Institute of Techology Ladkrabang, 2000.

[7] Vrajitoru Dana, Crossover Improvement For The Genetic Algorithm in Information Retrieval, Universite de Nauchatel, Intitui interfacultaire d'informatique, 1998.

[8] Shian_Hua Lin. Member, IEEE, Meng Chang Chen, Jan-Ming, "ACIRD: Intelligent Internet Document Organization and Retrieval." IEEE Transactions on Knowledge and Data Engineering, Vol 14 No 3 May/June 2002.

[9] Arazu Arvid, Garcia-Molina Hector Extracting Structured Data From Web Pages, Project Stanford InfoLab; Database Group, Stanford University, 2002

[10] Salton, G. "Advanced Study Institute for on-line mechanized information retrieval systems", Nato (1972)

[11] Kobayashi Mei, Takeda Koichi, Information Retrieval on the Web, IBM Research, ACM Computing Surveys. Vol. 32, No2, june 2000.

[12] Ruthven Ian, Lalmas Mounia Lalmas, Using Dempster-Shafer´s Theory of Evidence to Combine Aspects of Informations Use, Department of Computing Science, University of Glasgow, 2001.

[13] Becker Shirley, "Effective Databases for Text & Document Management" IRM Press. Publisher of innovative scholarly and professional information technology titles in the cyber age 2003.

[14] Baldi Pierre, Frasconi Paolo, Smyth Padhraic, Modeling the Internet and the Web, Johm Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester West Sussex PO198SQ. England , ISBN 0-470-84906-1,2003.

[15] Melnik, Sergey; Raghavan, Sriram; Yang, Beverly; Garcia-Molina, Hector, Building a Distributed Full-Text Index for the Web, Databases and the Web; Digital Libraries, Stanford University, 2000

[16]. Goodman, L. and Kruskal, W., 'Measures of association for cross-classifications II:Further discussions and references', Journal of the American Statistical Association, 54, 123-163 (1959).

[17] Fairthorne, R.A, "The mathematics of classification" Towards Information Retrievel, Butterworths, London, 1-10 1961

[18] Lewis, D. D. (1998). Naive (Bayes) at forty: "The independence assumption in information retrieval. ECML-98.

[19] Tzeras, K., & Hartmann, S. (1993). Automatic indexing based on Bayesian inference networks. SIGIR-93 (pp. 22–34).

[20] Lam, W., & Ho., C. (1998). Using a generalized instance set for automatic text categorization. SIGIR-98 (pp. 81–89).

[21] Masand, B., Linoff, G., & Waltz., D. (1992). Classifying news

stories using memory based reasoning. SIGIR-92 (pp.59–64).

[22] Wiener, E. D., Pedersen, J. O., & Weigend, A. S. (1995). A neural network approach to topic spotting. 4th nnual Symposium on Document Analysis and Information Retrieval (pp. 317–332)."

[23] Apte, C., Damerau, F., & Weiss, S. Text mining with decision rules and decision trees. Proceedings of the orkshop with Conference on Automated Learning and Discovery: Learning from text and the Web. (59–64), 1998

[24] Cohen, W. W., & Singer, Y. Context-sensitive learning methods for text categorization. Proceedings of he 19th Annual International ACM SIGIR Conference on Research and Development in Information (pp. 307–315). 1996

[25]Joachims, T, A statistical learning model of text classification with support vector machines. SIGIR-01 pp. 128–136), 2001

[26] Godbole, S., Sarawagi, S., & Chakrabarti, S,Scaling multi-class support vector machine using inter-class confusion. SIGKDD-02 (pp. 513–518), 2002.

[27] Nigam, K., Lafferty, J., & McCallum, A. Using maximum entropy for text classification. In IJCAI-99 workshop on Machine Learning for Information Filtering (pp. 61–67).", 1999

[28], Schapire, R. E., & Singer, Y. Boostexter: A boosting-based system for text categorization. Machine Learning, 39, 135–168, 2000.

[29] Syan-Wen Li, Candan K. Selkuk, Vu Quoc, Agrawal Divyakant. Query Relaxation by Structure and Semantics for Retrieval of Logical Web Documents. IEEE Transactions On Knowledge and Data Enginering Vol 14, 2002.

[30] Srikanth Munirathnam, Srihari Rohini. Exploiting Systactic Structure of Queries in a Language Modeling Approach to IR, State University of New York at Buffalo, 2003 ACM 1-58113-723-0/03/0011

[31] SkillSoft Corporation, Natural Language Processing ReferencePoint Suite, Published by SkillSoft Corporation, 20 Industrial Park Drive, Nashua, NH 03062 (603) 324-3000, 2004.

[32] Hofmann Thomas. Probabilistic Latent Semantic Analysis. Eecs Departament, Computer Science Division, University of California, Berkeley & International Computer Science Institute, Berkeley, CA. 1999.

[33] Hofmann Thomas. Probabilistic Latent Semantic Index. Eecs Departament, Computer Science Division, University of Caligornia, Berkeley & International Computer Science Institute, Berkeley, CA. 1999.

[34] Li Tao, Zhu Shenghuo, Ogihara Mitsunori, Efficient Multi-Way Text Categorization via Generalized Discriminant Analysis, Computer Science Dept. University of Rochester. ACM 2003.

[35] Goodman, L. and Kruskal, W., 'Measures of association for cross-classifications', Journal of the American Statistical Association, 49, 732-764 (1954).

[36] Kuhns, J.L., 'The continuum of coefficients of association'. In Statistical Association Methods for Mechanised Documentation, (Edited by Stevens et al.) National Bureau of Standards, Washington, 33-39 (1965).

[37] Cormack, R.M., 'A review of classification', Journal of the Royal Statistical Society, Series A, 134, 321-353 (1971).

[38] Sneath, P.H.A. and Sokal, R.R., Numerical Taxonomy: The Principles and Practice of Numerical Classification, W.H. Freeman and Company, San Francisco (1973).

[39] Salton, G. Relevance assessments and Retrieval system evaluation, Information Storage an retrieval (1969),

[40] Jardine, N. and Sibson, R., Mathematical Taxonomy, Wiley, London and New York (1971).

[41] Simonnot, B. and Smail, M. (1996). Modele flexible por la recherche interactive de documents multimedias. Proceedings of Inforsid (pp. 165–178) Bordeaux, 1996

[42] Nie, J. An outline of a general model for information retrieval systems. Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (pp. 495–506). 1988

[43] Klabbankoh Bangorn, Pinngern PH.D. Applied Genetic Algorithms in Information Retrievel Faculty of Information Technology, King Mongkut´s Institute of Techology Ladkrabang, 2000.

[44] Pathek98 Praven, Gordon Michael, Fan Weiguo. Effective Information Retrievl using Genetic Algorithms bases Matching Functions Adaptation Departement of Computer & Information Ssystem. University of Michigan School 701 Tappan Street; Ann Arbor, 1998.

[45] Vrajitoru Dana, Crossover Improvement For The Genetic Algorithm in Information Retrieval, Universite de Nauchatel, Intitui interfacultaire d'informatique, 1998.

[46] Praveen Pathak, Michael Godon, Weiguo Fan, "Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation" Department of Computer & Information Systems, University of Michigan Business School. 1998

[47] Wang Ke, Liu Huiqing, Discovering Strucutral Association of Semistructured Date. IEEE. Transactions on Knowledge and Data Engineering, Vol 12 No 3 , 2000.

[48]Thuraisingham Bhavani, "Web Data Mining and Applications in Business Intelliegence and Counter –Terrerism" CRC PRESS Boca Raton London New York Washington, D.C, ISBN 0-8493-1460-7, 2003.

[49] Cleverdon, C.W., Mills, J. and Keen, M., Factors Determining the Performance of Indexing Systems, Vol. II, Test Results, SLIB Cranfield Project, Cranfield, 1966

[50] Cuadra, A.C. and Katter, R.V., Opening the black box of "relevance", Journal measures, Journal of Documentation, 25, 93-107. 1969

Sonia Ordóñez Salinas

Estadista, Universidad Nacional de Colombia. Ingeniera y Especialista en Teleinformática, Universidad Distrital Francisco José de Caldas. Directora Grupo de Investigación Gesdatos - Facultad de Ingeniería, Universidad Distrital. Profesora Universidad Distrital. **soniaord@neutel.net.co**

Fabio A. González O.

Ingeniero de Sistemas Universidad Nacional de Colombia. M.Sc. en Matemáticas, Universidad Nacional de Colombia. Ph.D & M.Sc. in Computer Science University of Memphis. Profesor asociado Universidad Nacional de Colombia. **fagonzalezo@unal.edu.co**