WIREE
Propuesta de una Metodología de Recuperación de Información Web Eficaz y Eficiente para un Dominio Específico

María R. Romagnano
Dpto. de Informática-Instituto de Informática
FCEFN – UNSJ
San Juan, Argentina
maritaroma@iinfo.unsj.edu.ar

Martín G. Marchetta
Centro Universitario
FI- UNCuyo
Mendoza, Argentina
mmarchetta@fing.uncu.edu.ar

Resumen

Las tecnologías de información han impulsado el uso masivo de Internet, principalmente de la Web como fuente de información. Este repositorio digital permite contar con un "abanico" de información fácil y rápidamente. Esta accesibilidad podría convertirse en un problema si se presentan miles de respuestas a nuestra consulta y ninguna de las ellas es satisfactoria (ruido); o no se encuentra respuesta (silencio). Esta situación es realmente verdadera si se tiene en cuenta que generalmente se busca información precisa sobre un dominio particular, usando buscadores que devuelven resultados sin distinción del dominio. En trabajos recientes se han sugerido distintas formas de agrupar fuentes de información similares, tratando de dar solución a los problemas planteados. No obstante, existen dominios más complejos que otros, donde encontrar la solución es un problema aún más engorroso. El desafío de una aplicación que brinde información va más allá de ubicar dónde está esa información, entregando información relevante para el usuario. Nuestra contribución consiste en una metodología para asistir al usuario web en su búsqueda de información en un dominio de aplicación determinado, reduciendo la dificultad en la búsqueda, mejorando la precisión y el tiempo de respuesta.

1. Introducción

La WWW se ha convertido en una de las mayores fuentes de información sobre prácticamente todas las áreas de interés. El usuario, en su afán de obtener información fácil y velozmente, actualmente recurre a la Web en primer orden. Este proceder podría convertirse en un problema si se encuentran miles de respuestas y que tal vez ninguna de ellas le es satisfactoria, lo cual se conoce comúnmente como "ruido". Por el contrario podría no encontrar respuesta alguna, "silencio". Podría decirse: "la Web se ha convertido en una biblioteca caótica debido a su continuo y acelerado crecimiento".

Por lo tanto, tratando de obtener respuestas acorde a sus necesidades, el usuario podría pasar un buen lapso de tiempo examinando cada uno de los miles de resultados o simplemente elegir al azar uno de los primeros y que quizás no lo convence del todo. Esto puede deberse a que pocas veces se ocupan las herramientas provistas por los buscadores para filtrar resultados, por ejemplo no se realizan búsquedas avanzadas o búsquedas con palabras específicas de la temática. Las respuestas suelen ser más o menos precisas, en menor o mayor cantidad, organizadas o clasificada de distinta forma o según determinados criterios. Los principales problemas que afectan al usuario en su interacción con los buscadores son la forma en la que ellos especifican su consulta y la forma en la que interpretan los resultados. Generalmente realizan búsquedas "muy por encima", simplemente se limitan a abrir las páginas cuyos títulos y descripción se ajusten mejor a sus necesidades [1].

Habitualmente, para lograr recuperar e integrar datos desde diferentes sitios, se requiere de aplicaciones especializadas, complejas y con dificultades en tiempo de desarrollo y permanente mantenimiento. El ideal sería poder recuperar esta información en la Web como si se estuviese trabajando en bases de datos, con la misma facilidad y transparencia de contenidos para los usuarios. Al mismo tiempo, todavía existen recursos web que permanecen inexplorados, lo que se conoce como Internet invisible, ya sea porque se trata de páginas de acceso restringido o porque estos recursos están embebidos dentro de sitios web individuales y los actuales buscadores no poseen la facilidad para acceder a ellos [2].

La problemática planteada se pone claramente de manifiesto en algunos dominios donde un amplio espectro de información se encuentra distribuido en varios sitios web, almacenado usando formatos heterogéneos. Lo anhelado sería que el usuario se encuentre con los servicios, productos y/o la información que necesita sin demasiado esfuerzo. El desafío de una aplicación que brinde información va más allá de ubicar dónde está esa información, ofreciendo información precisa y relevante para el usuario. Consecuentemente,

nuestro aporte es una metodología para asistir al usuario web en su búsqueda de información en un dominio de aplicación determinado y caracterizado, reduciendo la dificultad de búsqueda, mejorando la exactitud y el tiempo de respuesta. En la sección 2 de este artículo se presentan los trabajos relacionados. La sección 3 describe la metodología propuesta y se menciona un ejemplo de aplicación. Finalmente, la sección 4 presenta las conclusiones y trabajos futuros.

2. Trabajos Relacionados

Según [3] la recuperación de la información, o IR; ayuda al usuario a encontrar información necesaria y relevante desde una gran colección de documentos de texto. En [4] se refuerzan la idea de que usar agentes para recuperar información de la Web mejora la eficiencia de los mercados on-line y además crea un clima donde las empresas agilizan sus relaciones. En [5] los autores se refieren a la recuperación de información masiva o recuperación de información distribuida como una técnica que consulta múltiples colecciones de documentos al mismo tiempo. En [6] se plantea una métrica para evaluar la recuperación de la información. Los autores proponen una métrica de evaluación estadística incorporando además de la relevancia, la popularidad del documento (cantidad de veces que se visitó el documento). En [7] proponen extender el proceso de recuperación de información usando tecnologías de la web semántica. Esta propuesta tiene en cuenta tanto las palabras clave expresadas en la consulta del usuario como también su significado, el cual se representa mediante una ontología. En [8] se plantea un clasificador para determinar si una página considera un determinado tema o no. Se usa un algoritmo genético basado en un sistema de clasificación de páginas web. En [9] se presenta un método de clúster relacionado con páginas web que no sólo considera las palabras claves de un dominio específico en una ontología, sino que también analiza el contenido semántico de las páginas web. En [10] se ha usado la información de las etiquetas para clusterizar, usando el método k-means. En [11,12] se propone un algoritmo de clustering difuso ponderado, el cual inicializa las características ponderándolas con la varianza de los términos y además actualiza la fórmula de Fuzzy C-Means y reformula la función objetivo. Se logran menores resultados pero el tiempo de ejecución se incrementa considerablemente respecto al algoritmo de clustering difuso original. En [13] se propuso un prototipo de buscador web que agrupa resultados en forma sólida y difusa, logrando mejoras en la clasificación o rankeo de los resultados y etiquetado de los clusters en relación con los buscadores convencionales. En [14] se introduce el uso del diccionario de Internet para obtener los contextos en los cuales una palabra clave puede aparecer y a su vez poder agrupar los resultados basados en este contexto. En [15] los autores proponen usar una técnica estadística para clusterizar páginas web por medio de su URL. En [16] se realiza una comparación entre los algoritmos K-Means y Fuzzy C-Means para agrupar páginas web. El trabajo de [17] tuvo como objetivo presentar una arquitectura de un motor de búsqueda semántico, basado en un enfoque bottom up para incorporar semántica en la búsqueda, centrado en la construcción de una base de datos semántica para almacenar el contenido web y luego poder llevar a cabo las consultas, teniendo en cuenta una alta precisión y un menor recall. En [18] se plantea la arquitectura de un buscador semántico para indexar y recuperar documentos de texto relevantes. Se usa el modelo clásico de recuperación TF/IDF, se realiza la indexación de los documentos usando un archivo invertido y luego se remueven las stopwords y se realiza stemming y capitalización. El motor de búsqueda contrasta la consulta con un índice compuesto por las palabras de cada documento, además de un puntero a la ubicación de cada palabra en cada documento. También, pueden mencionarse los sistemas de clustering de resultados webs llamados motores de clustering web los cuales agrupan los resultados en función de un criterio determinado y muestran los resultados en clases bien definidas. Ejemplos de estos son: Carrot2, Vivísimo, Yippy, Clusty, etc [19]. En [20] se propone un enfoque para clusterizar resultados de los motores de búsqueda teniendo en cuenta la similitud semántica y sintáctica entre los documentos. Plantea un marco de trabajo comenzando con la consulta del usuario enviada a un motor de búsqueda. Los documentos resultantes son procesados para obtener las características significativas de cada documento. Los términos significativos son tokenizados y lematizados haciendo uso de WordNet, a la vez que se remueven las stopwords. En [21] desarrollaron un modelo de dominio específico que soporta múltiples dominios proporcionando así una alta escalabilidad. Este prototipo está diseñado principalmente para algunos productos tales como libros, móviles, medicamentos, etc. El modelo de búsqueda almacena el contenido RDF de páginas web para un dominio específico. Ofrece respuestas más completas y entendibles. En [22] los autores han propuesto un agente de filtrado que localiza fuentes de información y las agrupa de acuerdo a los servicios que ofrecen. En [23] se propone construir una ontología para E-Turismo y así proveer servicio turístico inteligente. El algoritmo se diseña para integrar datos desde fuentes confiables, estructurarlos adecuadamente en una base de conocimiento del turismo y así poder realizar una búsqueda eficiente.

3. Metodología Propuesta

La metodología propuesta forma parte de una aplicación automática que cada un determinado tiempo y en forma organizada recupera y analiza documentos web. Luego agrupa y almacena sólo los documentos relevantes para un dominio específico. La Figura 1 muestra una idea general de la aplicación propuesta, donde D_i representa una generalización de la clase dominio, la cual se especifica con los dominios preestablecidos.

Posteriormente el sistema de recuperación de información asistirá eficaz y eficientemente al usuario en su búsqueda dentro del espectro de servicios públicos como educación, salud y turismo. La Figura 2 muestra la estructura interna del sistema de recuperación propuesto.

La metodología sugiere seis etapas, abarcando desde el relevamiento de los documentos web, procesamiento e indexado de los mismos, hasta resolver la consulta del usuario. Estas etapas se describen a continuación.

3.1. Buscar

Cada cierto período de tiempo (cuya frecuencia variará dependiendo del dinamismo con el cual cambie el dominio en cuestión), automáticamente y a través de las APIs provistas por los buscadores generales e índices temáticos, se realiza la búsqueda de la información con las palabras claves DOMINIO, PROVINCIA y PAÍS.

3.2. Seleccionar documentos relevantes – Pre procesamiento

En esta etapa se realiza un análisis preliminar de los resultados obtenidos en la etapa anterior para seleccionar aquellos documentos que sean relevantes al dominio en cuestión. De los resultados arrojados por los buscadores se analiza su texto anclado y el URL. Para realizar el análisis se usa como base de contrastación una ontología preexistente del dominio de aplicación. Por cada término que aparece en el texto anclado y/o en el URL del documento en cuestión se contrasta con la ontología (Figura 3). Si el término (o su sinónimo) aparece en la ontología se lo llamará t_p y se le asignará un peso de 1. Si el término (o su sinónimo) no se encuentra en la ontología se lo denominará t_n y se le asignará un peso de 1. Los términos t_p (términos positivos) se irán sumando por su parte y los términos t_n (términos negativos) se irán sumando por otra parte, considerando la fórmula:

$$R_{d_j} = \sum_{p=1}^{m} t_p - \sum_{n=1}^{m} t_n \quad (1)$$

Rd_j: relevancia del documento d_j
m: cantidad de términos del texto anclado y del URL.
t_p: término que aparece en la ontología.
t_n: término que no aparece en la ontología.

Sí $R_{d_j} \geq 0$; entonces d_j se considera relevante.

Sí $R_{d_j} < 0$; entonces d_j se considera poco relevante.

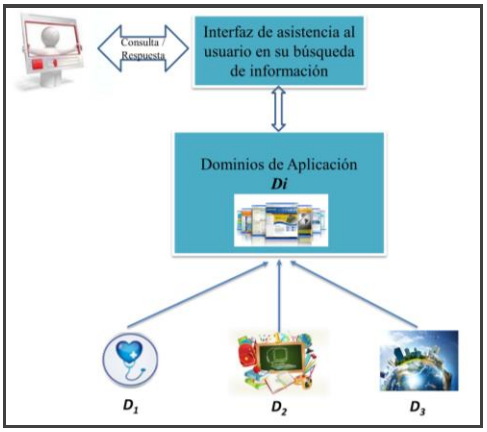

Figura 1. Idea general de la aplicación

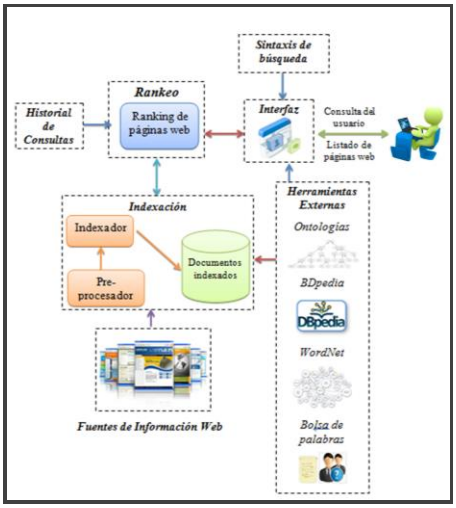

Figura 2. Estructura del sistema de recuperación propuesto

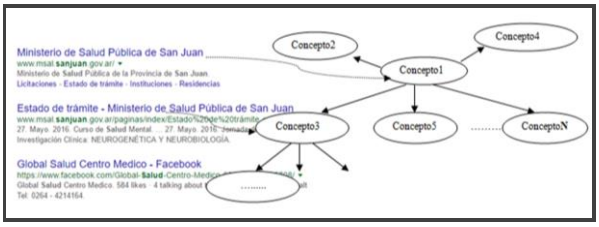

Figura 3. Selección de documentos relevantes

Para finalizar esta etapa, de aquellos documentos que hayan sido considerados como relevantes se calcula la cobertura como se muestra en la siguiente fórmula:

$$Cd_j = \frac{\sum_{i=1}^{n} t_i}{ct_o} * 100 \quad \textbf{(2)}$$

Cd_j: cobertura del documento d_j
n: cantidad de términos
t_i: cantidad de términos relevantes del documento d_j
ct_o: cantidad de términos de la ontología

Luego, aquellos documentos cuya cobertura (Cd_j) sea mayor o igual a un umbral de cobertura preestablecido (UCd_j) se confirma su relevancia y se almacenan en una base de relevantes.

3.3. Seleccionar términos relevantes

En esta etapa se determina cuáles serán los términos relevantes. Para seleccionar dichos términos se emplean la bolsa de palabras, BDpedia, WorNet y la ontología del dominio, bosquejadas en la Figura 2 como herramientas externas. Manualmente, el experto en el dominio es quién determina qué términos serán los candidatos a ser seleccionados y posteriormente en forma automática se realiza una comparación de estos con los términos de las restantes herramientas. Es decir, se realiza un análisis semántico estableciendo relaciones entre cada término de la bolsa de palabras y los términos de cada una de las restantes herramientas externas. Aquellos términos dados por el experto que sean similares o que coincidan con los términos de al menos dos de las tres herramientas restantes serán considerados como relevantes y por consiguiente serán establecidos como nombres de los futuros grupos.

3.4. Determinar el valor de cada término en cada documento

El objetivo de esta etapa es obtener la cantidad de información que ofrece un documento sobre términos discriminantes, es decir términos relevantes al dominio comprometido. La base que contiene los documentos relevantes se representa en la Tabla 1, donde D_j representa la j-ésimo documento, t_i representa el i-ésimo término relevante y w_{ij} representa la normalización del número de veces que el término t_i aparece en el Documento D_j.

El procedimiento consiste en analizar el contenido de cada documento y remover las stopwords y realizar stemming para determinar la frecuencia de aparición de cada término relevante y sus variantes. Nuevamente, se usan las herramientas externas para establecer la correlación semántica entre los términos de las diversas fuentes de información web y los términos relevantes. Luego, se calcula el peso w_{ij} de los términos relevantes en cada documento, según la ecuación:

$$w_{ij} = \frac{f_{ij}}{\max\left\{f_{1j}, f_{2j}, ..., f_{nj}\right\}} \quad \textbf{(3)}$$

$f_{ij:}$ es la frecuencia del término relevante t_i en $D_{j.}$
$n:$ es cantidad de términos.

Tabla 1. Términos con sus pesos para cada documento relevante, en un determinado dominio

Dj	*URL*	t_1	t_2	t_3	…	t_n
D_1	URL_1	w_{11}	w_{21}	w_{31}	…	w_{n1}
….						
D_m	URL_m	w_{1m}	w_{2m}	w_{3m}	…	w_{nm}

3.5. Agrupar

En esta etapa la metodología propone que aquellos documentos que presenten información de un mismo término deben agruparse. La intensión de agrupar consiste en disminuir los tiempos y aumentar la precisión de la posterior búsqueda. Sin embargo, los documentos pueden presentar información de varios términos relevantes. Necesariamente la metodología debe permitir que un documento pueda corresponder a uno o a más grupos con un cierto grado de pertenencia.

3.5.1. Algoritmo

Para desarrollar el algoritmo se tomaron como base métodos de clustering particionales y soft [24].

Los grupos se eligen una única vez y se mantienen durante todo el proceso. La pertenencia de cada elemento a cada grupo se registra en una matriz y en una única iteración. La dimensión de cada vector fila coincide con la cantidad de grupos. Se asume que a cada grupo (G_i) se le asignará el nombre de cada columna t_i de la Tabla 1 y que el mismo tendrá un valor central al cual se lo llama centro. Los pasos del algoritmo se detallan a continuación:

1. Para cada grupo calcular el centro como el máximo valor de cada columna. Si dos o más documentos tuviesen un valor máximo se calcula su densidad. Se considerará aquel documento cuya densidad sea al menos de un 30%, es decir que al menos cubre el 30% de la cantidad total de términos. En este caso no sólo se contempla que sea el máximo de la columna sino también cuántos términos relevantes son cubiertos por este documento. Además, si dos o más documentos resultan empatados como candidatos, tienen máximo valor de columna y coinciden en densidad, se evalúa la información total de cada documento candidato como la cantidad total de términos relevantes. Si aún siguen

empatando el sistema elige al azar uno de los documentos en empate.

2. Mientras existan documentos por analizar:

2.1. Calcular la similitud de cada documento d_j con cada centro c_i. La similitud se calcula como la razón geométrica que viene dada por la siguiente expresión:

$$S_{d_j c_i} = \frac{w_{ij}}{w_{ci}} * 100 \quad \textbf{(4)}$$

Sd_jc_i: similitud entre el documento d_j y el centro c_i
w_{ij}: peso del término t_i en el documento d_j.
w_{ci}: peso del centro c_i.

Esto dará la idea de qué porcentaje del máximo (centro) está cubriendo el documento analizado, para un término en particular.

2.2. Determinar el umbral de similitud como se establece en la fórmula:

$$U_S = \frac{1}{n} \sum_{i=1}^{n} Sd_j c_i \quad \textbf{(5)}$$

U_s: umbral de similitud
Sd_jc_i: similitud entre el documento d_j y el centro c_i
n:cantidad de centros

2.3. Por cada similitud del documento d_j con cada centro c_i

2.3.1. Si la similitud del documento d_j con el centro c_i es mayor o igual a U_S entonces el documento d_j se selecciona y ubica en el grupo G_i con su respectivo grado de pertenencia, es decir el peso w_{ij}.

2.3.2. Sino el documento se considera poco similar y se descarta como miembro del grupo G_i.

3.6. Responder a la consulta del usuario

En esta etapa ante la consulta de un usuario, el sistema deberá realizar las siguientes etapas:

1. Remover los sotpwords.
2. Realizar stemming.
3. Análisis del dominio en cuestión.
4. Instanciar la aplicación con el dominio comprometido.
5. Establecer relaciones semánticas entre los términos de la consulta y el nombre del o los grupos a los cuales se debe acceder para responder a la consulta.
6. Otorgar un listado de URLs. Para establecer el orden (o ranking) en el cual se van a mostrar las direcciones web se propone la función de relevancia y cada una de sus partes integrantes, como se muestra en las fórmulas:

$$Rd_j = dr_j + A_j + dc_j \quad \textbf{(6)}$$

Rd_j: relevancia del documento d_j.

$$dr_j = \sum_{i=1}^{n} \frac{w_{ij}}{wc_i} ; \; 0 < dr_j <= 1 \quad \textbf{(7)}$$

dr_j: densidad relativa de información del documento d_j.
w_{ij}: peso del término t_i en el documento d_j.
wc_i: peso del centro c_i.
n: cantidad de términos relevantes del documento d_j.

$$A_j = \begin{cases} 0 \text{ si el documento no aparece en el historial de consultas} \\ 1 \text{ si el documento aparece en el historial de consultas} \end{cases}$$

A_j: aparición del documento d_j en el historial de consultas.

$$dc_j = \frac{ctcd_j}{ctc} ; \; 0 < dc_j <= 1 \quad \textbf{(8)}$$

dc_j: densidad de cobertura del documento d_j.
$ctcd_j$: cantidad de términos de la consulta que aparecen en el documento d_j.
ctc: cantidad de términos de la consulta.

En este trabajo se eligió como ***ejemplo de aplicación*** el dominio de la salud en la provincia de San Juan. Para llevar a cabo las pruebas, durante enero de 2016, se realizaron distintas búsquedas con las palabras claves establecidas por la metodología y se obtuvo una muestra de 21 documentos. Ante la consulta *"hospitales Públicos y Privados en San Juan"* el sistema seleccionó los documentos que se encontraban en los grupos implicados y arrojó un listado de direcciones, siguiendo el orden de méritos establecido por la función de relevancia.

4. Conclusiones y Trabajos Futuros

En este trabajo se presentó una metodología que permite recuperar documentos web de interés para el usuario. El interesado no tiene que realizar por sí mismo la búsqueda en la web, enfrentándose a una gran cantidad de resultados. Sólo debe realizar la consulta al sistema, el cuál como ya cuenta con información clasificada obtiene resultados precisos y en menor tiempo. Esta ganancia en precisión y tiempo se logra debido a que la metodología analiza la semántica tempranamente y durante la mayoría de las etapas. Además, para realizar la tarea de selección y agrupamiento de documentos relevantes se presentan novedosas fórmulas. Necesariamente los documentos deben agruparse en función de la similitud que presentan respecto a un determinado término y no teniendo en cuenta toda la información del documento. Al proponer grupos solapados, se pueden escoger rápidamente documentos que sólo brindan información explícita de los términos solicitados. En trabajos futuros se espera

definir valores numéricos para cada uno de los umbrales, en función del cúmulo de información y complejidad semántica que se maneje en cada uno de los dominios que sean cubiertos por la aplicación. Además se desea que la aplicación sea escalable a dominios que presenten otro tipo de características y a otras ciudades.

5. Referencias

[1] Cacheda, F., and Vina, A. (2001). "Understanding how people use search engines: a statistical analysis for e-business". In Proceedings of the e-Business and e-Work Conference and Exhibition. Pp. 319-325.

[2] Mendez Duque N., Chavarros Porras J. C. and Moreno Laverde R. (2007). "Integrando información de fuentes heterogeneas. Enfoques y tendencias". Scientia Et Technica, mayo, año/vol. XIII. Nro. 034. Universidad Tecnológica de Pereira. Colombia. Pp. 397 - 402.

[3] Liu B. (2007)."Web Data Mining – Exploring Hyperlinks, Contents and Usage Data". Springer-Verlag Berlin Heidelberg, ISBN-10 3-540-37881-2, ISBN-13 978-3-540-37881-5. Pp.183-187.

[4] Bohórquez, Y. E., Téllez, M., and Rodríguez, J. E. (2012). "Software basado en agentes inteligentes y servicios web para búsqueda de productos en la web". Tecnura, 16(31). Pp.114-125.

[5] Li, M. and Cao, S. (2014). "A serie method of massive information storage, retrieval and sharing". In 2014 IEEE International Conference on Mechatronics and Automation. IEEE. Pp. 1171-1175.

[6] Evangelopoulos, X., Giannakouris-Salalidis, V., Iliadis, L., Makris, C., Plegas, Y., Plerou, A., and Sioutas, S. (2016). "Evaluating information retrieval using document popularity: An implementation on MapReduce". Engineering Applications of Artificial Intelligence, 51. Pp. 16-23.

[7] Solarte P., O. and Millán G., M. (2014). "Propuesta para extender semánticamente el proceso de recuperación de información". Revista EIA, ISSN 1794-1237 / Año XI / Volumen 11 / Edición N.22 / Julio-diciembre 2014.

[8] Özel, S. A. (2011). "A web page classification system based on a genetic algorithm using tagged-terms as features". Expert Systems with Applications, 38(4). Pp. 3407-3415.

[9] Chen, R., Bau C. and Tsai, M. (2010)."“Web pages cluster based on the relations of mapping keywords to ontology concept hierarchy". International Journal of Innovative Computing, Information and Control, 6(6). ISSN 1349-4198, pp. 2749–2760.

[10] Shelke, M., Sadavarte, K., Dhurjad R., and Pandit, N. (2012). "Improved web page clustering using words and tags". 1° International Conference on Recent Trends in Engineering & Technology. Special Issue of International Journal of electronics, Communication & Soft Computing Science & Engineering. Pp. 25-28.

[11] Kamjou, M., and Ahmadzadeh, M. (2015). "Improvement of Fuzzy C-Means by using variance-based weighted Feature". Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org, 2(2).

[12] Xing, H. J., and Ha, M. H. (2014). "Further improvements in feature-weighted fuzzy c-means". Information Sciences, 267.

[13] Matsumoto, T., and Hung, E. (2010). "Fuzzy clustering and relevance ranking of web search results with differentiating cluster label generation". In 2010 IEEE International Conference on Fuzzy Systems (FUZZ). Pp. 1-8.

[14] Hegde, V. (2011). "Web Pages Clustering: A New Approach". International Journal of Innovate Technology & Creative Engineering, 1(4). ISSN:2045-8711.

[15] Hernández I., Rivero C., Ruiz D. and Corchuelo R. (2012). "A statistical approach to URL-based web page clustering". Proceedings of the 21st international conference companion on World Wide Web.ACM. Pp. 525-526.

[16] Ghosh, S., and Dubey, S. K. (2013). "Comparative analysis of k-means and fuzzy c-means algorithms". International Journal of Advanced Computer Science and Applications (IJACSA) 4(4). Pp. 35-39.

[17] Kamath, S. S., Piraviperumal, D., Meena, G., Karkidholi, S., and Kumar, K. (2013). "A semantic search engine for answering domain specific user queries". In International Conference on Communications and Signal Processing (ICCSP). IEEE. Pp. 1097-1101.

[18] Jayalakshmi, T. S., and Chethana, C. (2016). "A Semantic Search Engine for Indexing and Retrieval of Relevant Text Documents". International Journal, 4(5).

[19] Lin, T. and Chi, Y. (2014). "Application of web page optimization for clustering system on search engine v google study". In International Symposium on Computer, Consumer and Control (IS3C). IEEE. Pp. 698-701.

[20] Soliman, S. S., El-Sayed, M. F., and Hassan, Y. F. (2015). "Semantic Clustering of Search Engine Results". The Scientific World Journal. Volume 2015 (2015), Article ID 931258, p. 9.

[21] Sinha, S., Dattagupta, R., and Mukhopadhyay, D. (2012). "Designing an ontology based domain specific web search engine for commonly used products using RDF". In Proceedings of the CUBE International Information Technology Conference. ACM. Pp. 612-617.

[22] Romagnano, M., Dominguez, P., Marchetta, M., and Aciar, S. (2015). "Reduciendo la Complejidad de Búsqueda Web en Base a las Necesidades del Usuario". In Proceedings of the 3rd National Congress of Computer Engineering / Information Systems (CONAIISI, 2015), November 19-20. ISBN: 978-987-1896-47-9.

[23] Agarwal, J., Sharma, N., Kumar, P., Parshav, V., Srivastava, A., Rathore, R., and Goudar, R. H. (2014). "Semantic Search in E-Tourism Services: Making Data Compilation Easier". In Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012 Springer India. Pp. 811-820.

[24] Han J. and Kamber M. (2006). "Data Mining: Concepts and Techniques". Segunda Edición. Elsevier. ISBN 13: 978-1-55860-901-3 ISBN 10: 1-55860-901-6. Pp. 402-408.