Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 31-38 recibido 03-04-2018 revisado 02-05-2018 aceptado 04-05-2018

Aproximación a un modelo de recuperación de información personalizada basado en el análisis semántico del contenido

An Approach of a Personalized Information Retrieval Model based on Contents Semantic Analysis

Eric Utrera Sust[1], Alfredo Simón-Cuevas[2], Jose A. Olivas[3] and Francisco P. Romero[3]

[1]Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, CP.: 19370, La Habana, Cuba
ebutrera@uci.cu

[2]Universidad Tecnológica de La Habana "José Antonio Echeverría", Cujae
Ave. 114, No. 11901, CP: 19390, La Habana, Cuba
asimon@ceis.cujae.edu.cu

[3]Universidad de Castilla La Mancha
Paseo de la Universidad, 4, Ciudad Real, España
{JoseAngel.Olivas, FranciscoP.Romero}@uclm.es

Resumen: En este trabajo se presenta una primera aproximación de un modelo de recuperación de información personalizada basado en el procesamiento semántico del contenido. El modelo propuesto reduce la sobrecarga de información innecesaria para los usuarios y mejora los resultados recuperados mediante la combinación de un procesamiento semántico de contenido aplicado a las consultas y documentos indexados, y la información de los perfiles de usuarios. La aplicabilidad de la propuesta fue evaluada en el contexto de un motor de búsqueda real, a través de consultas diseñadas por expertos en diferentes dominios y la medición de su rendimiento. Los resultados obtenidos fueron comparados con los del motor de búsqueda puesto a prueba, lográndose mejoras en cuanto a la precisión y exhaustividad.

Palabras clave: recuperación de información personalizada, análisis semántico

Abstract: In this paper, an approach of a personalized information retrieval model based on the semantic processing of the content is proposed. The proposed model reduces the unnecessary information overload for users and improves the retrieval results through combining a content semantic processing applied to the queries and indexed documents, and information user processing from different perspectives. The applicability of the proposal was evaluated in the context of a real web search engine, through several queries designed by experts and associated to differents topics, and the measurement of their performance. The results were compared to those obtained by the search engine put to the test, achieving improvements the retrieval results.

Keywords: information retrieval systems, search engines, semantic processing

1 *Introducción*

Actualmente, los buscadores web no siempre ofrecen la información que el usuario necesita como resultado de una consulta. Algunos de los factores que inciden en esta situación son: el análisis del contenido (consulta – documento indexado) aún suele estar basado, fundamentalmente, en análisis sintáctico del contenido textual, sin tener en cuenta la semántica subyacente (Klusch, Kapahnke, Schulte, et al., 2016); no se tienen en cuenta los intereses de los usuarios (Jay, Shah, Makvana, et al., 2015); no existe tratamiento de la ambigüedad inherente al lenguaje natural (Shou, Bai, Chen, et al., 2014) y baja calidad en la formulación de consultas (Singh, Dey, Ashour, et al., 2017).

La personalización de la recuperación de información constituye una de las líneas que

ISSN 1135-5948. DOI 10.26342/2018-61-3

actualmente se está trabajando para incrementar la calidad de los resultados de los buscadores web (Singh, Dey, Ashour, et al., 2017). En este sentido, se reporta el uso de varias técnicas, tales como: la expansión de consultas, la desambiguación de consultas, el uso de bases de conocimiento (ej. taxonomías, ontologías, etc.) y modelos de datos enlazados para mejorar los resultados de búsqueda (Tanaka, Spyratos, Yoshida, et al., 2015). El proceso de personalización cuando se integra a un Sistema de Recuperación de Información (SRI) de dominio general, para diferentes consultas y en diferentes contextos de búsqueda, es consistentemente menos efectivo que cuando se centra en un solo dominio (Makwana, Patel y Parth, 2017). Por otra parte, las consultas ambiguas, incompletas y breves, provocan que en ocasiones estos modelos de personalización generen grados de interés sobre documentos que no son los correctos, debido a que el usuario introduce la misma consulta de diferentes maneras y los documentos tienen diferentes contextos de búsqueda. Todo esto dificulta predecir el comportamiento y grado de interés de los usuarios sobre los documentos de forma adecuada y obtener nuevos conocimientos a través de esta interacción (Sharma y Rana, 2017). Se ha demostrado que el procesamiento de consultas y documentos influye en el proceso de mejorar la relevancia de los resultados (Hahm, Yi, Lee, et al., 2014; Corcoglioniti, Dragoni y Rospocher, 2016). En este sentido varios investigadores han propuesto soluciones enfocadas a SRI de dominio general (Zhang, Yan-hong, Wei-jun, et al., 2013; Preetha y Shankar, 2014; Shafiq, Alhajj y Rokne, 2015; Zhou, Lawless, Wu1, et al, 2016; Makwana, Patel y Parth, 2017) pero aún el tratamiento de las consultas y documentos no es suficiente.

En este trabajo se propone un Modelo de Recuperación de Información Personalizada (MRIP) basado en el procesamiento semántico del contenido para mejorar la eficacia de los SRI. MRIP está enfocado a un SRI de dominio general y se compone de 4 procesos: *Analizador de Contenidos*, *Generador de Perfiles*, *Personalizador y Generador de Ranking*. El *Analizador de Contenidos* procesa las consultas que se encuentran en el motor de consultas y los documentos almacenados en el índice del SRI. El *Generador de perfiles* recibe como entrada las acciones de los usuarios en el sistema y representa los perfiles de usuarios. El *Personalizador* es responsable de predecir qué documentos son interesantes para los usuarios y el *Generador de Ranking* es responsable de crear un ranking de los documentos recuperados por el MRIP. La aplicabilidad del MRIP fue evaluada por un total de 12 usuarios expertos y se integró en un motor de búsqueda web. Los resultados obtenidos fueron satisfactorios, mejorando la precisión y exhaustividad del motor de búsqueda.

El resto del documento está organizado de la siguiente manera: en la Sección 2 se analizan y caracterizan los trabajos relacionados con la problemática abordada; en la Sección 3 se presenta el modelo propuesto; en la Sección 4 se exponen y analizan los resultados del caso de estudio desarrollado para evaluar la aplicabilidad del modelo propuesto; y en la Sección 5 se presentan las conclusiones.

2 Trabajos relacionados

Los modelos de personalización tienen como objetivo caracterizar a los usuarios que interactúan con el sistema desde la generación de un perfil que está en correspondencia con sus gustos e intereses (Zhou, Lawless, Wul, et al., 2016). En (Zhang, Yan-hong, Wei-jun, et al., 2013) se propone un motor de búsqueda personalizado distribuido, que se utiliza para minar el historial web del usuario y crear una base de datos de patrones de interés. El modelo de interés de los usuarios se expresa mediante una tríada ordenada de la forma: palabra interesada, peso de la palabra y grado de estreno de la palabra. En (Shafiq, Alhajj y Rokne, 2015) se propone un enfoque para encontrar los intereses personales y los contextos sociales de los usuarios, basándose en las actividades de los usuarios en sus redes sociales, Desarrollan un mecanismo que extrae información de la red social de un usuario y la utiliza para volver a clasificar los resultados de un motor de búsqueda. Trabajo similar el de (Zhou, Lawless, Wu1, et al, 2016), donde se construyen perfiles de usuario mejorados a partir de un conjunto de anotaciones y recursos que los usuarios han marcado. Presentan dos modelos probabilísticos para incorporar simultáneamente anotaciones sociales, documentos y la base de conocimiento externa, y un modelo de expansión de consulta para mejorar la búsqueda. El proceso de expansión se hace a partir de un conjunto de palabras en el perfil de usuario, con el objetivo de devolver una lista ordenada de términos de perfil que se agregarán a la consulta. En

(Makwana, Patel y Parth, 2017) se analizan los clics del usuario y se crean grupos de usuarios similares utilizando la técnica de agrupamiento de C-means difusa. La consulta pasa por un proceso de eliminación de ambigüedades a partir de criterios que se han especificado en otras búsquedas. Para cada término de búsqueda relevante, se registran los enlaces en los que hizo clic ese usuario y también calcula el valor de interés.

La mayoría de los trabajos presentados analizan el grado de interés de los usuarios sobre los documentos basándose en los términos claves con los que se relaciona, sin tener en cuenta que un documento puede tratar de "java" y sin embargo ese término puede tener una frecuencia de aparición baja. Por otra parte, el proceso de expansión y desambiguación de consultas se hace utilizando búsquedas anteriores, a través de un análisis léxico-sintáctico que no tiene en cuenta el sentido de las palabras.

3 Modelo de Recuperación de Información Personalizada (MRIP)

MRIP (ver Figura 1), se compone de 4 procesos: *Analizador de Contenidos*, *Generador de Perfiles*, *Personalizador* y *Generador de Ranking.*

3.1 Analizador de contenidos

Este proceso se encarga de analizar las consultas que se encuentran en el motor de consultas y los documentos indexados. Se compone de dos subprocesos: *Procesamiento de consultas* y *Procesamiento de documentos*.

3.1.1 Procesamiento de consultas

La gran parte de las consultas depende de la falta de contexto de búsqueda. Para solucionar este problema el subproceso para analizar las consultas está compuesto por 4 subprocesos: *Análisis*, *Expansión*, *Almacenamiento* y *Cálculo de similitud*. El subproceso *Análisis* apoyándose en Freeling (Padró, y Stanilovsky, 2012) y BabelNet (Navigli, y Ponzetto, 2012), identifica el idioma de la consulta, tokeniza, elimina palabras vacías, extrae las palabras claves, e identifica entidades nombradas. Para identificar el idioma se compara el texto entrado con los módulos disponibles para diferentes idiomas en Freeling y devuelve el idioma en el que está escrito el texto. Para extraer los tokens se utilizan las reglas de tokenización propuestas por Freeling y para identificar las entidades nombradas se utiliza BabelNet y sus servicios de máxima entropía para detectar personas, nombres y organizaciones sobre las palabras claves. El subproceso *Expansión* extrae las relaciones semánticas entre los términos utilizando BabelNet, extrayendo sinónimos, hipónimos e hiperónimos para cada término de la consulta. La tercera etapa *Almacenamiento* colecciona la consulta expandida para el usuario en los repositorios del modelo de la Figura 1.

El cálculo de similitud semántica es aplicado en diferentes áreas del conocimiento y en el caso específico de los SRI posibilita que se encuentren resultados de búsquedas similares a

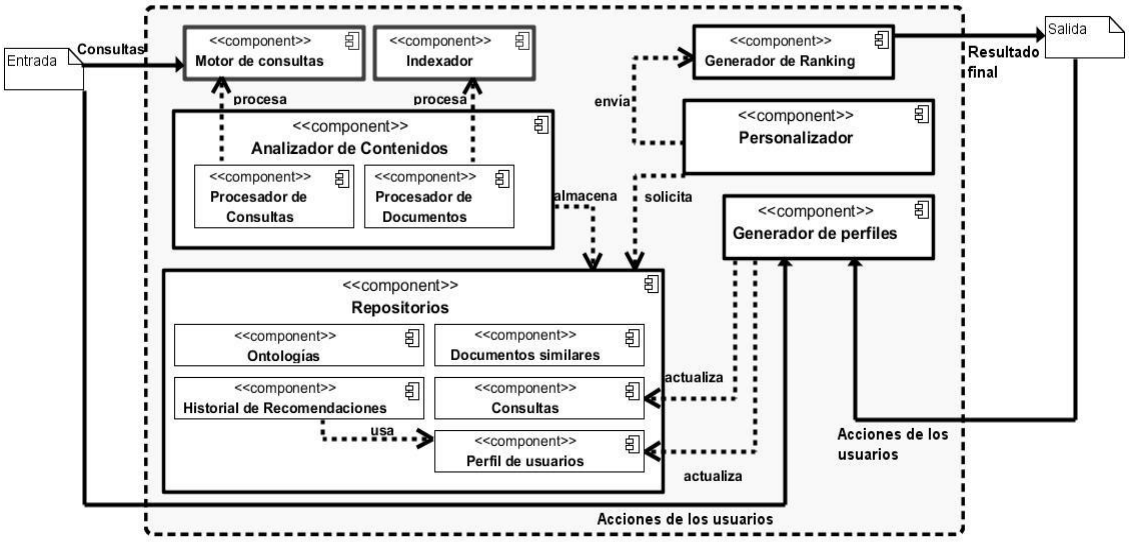

Figura 1. MRIP basado en el procesamiento semántico del contenido

las consultas de los usuarios. El subproceso *Cálculo de similitud* utiliza la medida propuesta por (Mihalcea, y Strapparava, 2006) por su utilidad en el cálculo de similitud entre cadenas de textos (ver fórmula (1)). Este cálculo se almacena en los Repositorios de la figura 1 como una matriz $M_{2x2}(\mathbb{R})$.

$$\begin{aligned}&\mathrm{sim}(C_1, C_2)\\&= \frac{1}{2}\left(\frac{\sum_{w\in\{C_1\}}(maxSim(w, C_2) * idf(w))}{\sum_{w\in\{C_1\}} idf(w)}\right.\\&\left. + \frac{\sum_{w\in\{C_2\}}(maxSim(w, C_1) * idf(w))}{\sum_{w\in\{C_2\}} idf(w)}\right) \quad (1)\end{aligned}$$

donde:
C1 y C2: consultas a analizar.
maxSim (w, C1): similitud semántica máxima entre la palabra w en la consulta *C1* y cada una de las palabras en la consulta *C2*.
idf (w): frecuencia inversa en el documento de la palabra w.

3.1.2 Procesamiento de documentos

Este proceso, aplicado a los documentos indexados, incluyen las siguientes tareas: (1) categorización y (2) análisis de similitud.

La categorización de los documentos indexados consta de 5 pasos: pre-procesamiento, extracción de bi-gramas, construcción de la colección, construcción del categorizador y categorización del documento. Los documentos son categorizados teniendo en cuenta 150 categorías (ej. Economía, Justicia, Biología, Ciencias de la información, etc.), las cuales se tomaron de Wikipedia (en Español e Inglés). En la etapa de pre-procesamiento el primer paso es la conversión de palabras a minúsculas, luego se eliminan las palabras vacías y se eliminan los acentos. En el proceso de creación de la colección se le aplica el pre-procesamiento a cada documento de una categoría, se extraen los bigramas de ese documento y se agregan al vector de contenido del documento como una palabra unida. Luego se crea un vector formado por los términos y bigramas (unión de las palabras de un bigrama) y su frecuencia de aparición. La colección de entrenamiento se crea con todos estos vectores formados por las palabras claves y su frecuencia de aparición en el documento. Finalmente, el proceso de categorización consiste en cargar la colección de entrenamiento, realizarle el proceso de pre-procesamiento al texto que se desea categorizar, extraer los bigramas del texto, agregar la unión de un bigrama como término independiente al vector de texto, crear el vector con palabras y su frecuencia de aparición y la aplicación del algoritmo Naïve Bayes Multinominal para la predicción de la categoría a la que pertenece.

Similitud entre los documentos. La similitud semántica mide la fuerza de las relaciones semánticas entre los conceptos incluidos en un documento. En este proceso, los documentos a evaluar son representados en forma de vectores, los cuales están formados por las palabras claves presentes en dichos documentos, y su extensión con otros términos relacionados capturados de BabelNet. La construcción de los vectores se lleva a cabo a partir de: eliminación de los términos repetidos en los documentos, extracción de las palabras claves según frecuencia de ocurrencia de los términos en el documento, etiquetado POS para obtener únicamente nombres, verbos y adjetivos de los documentos de entrada, la desambiguación de los términos usando Babelfy (Moro, Cecconi, y Navigli, 2014), y finalmente, la captura de BabelNet de otros términos sinónimos, hipónimos e hiperónimos, relacionados con las palabras claves identificadas. La evaluación de la similitud entre dos documentos se lleva a cabo usando el coeficiente de similitud de Jaccard (Chunzi y Wang, 2017) sobre los vectores característicos de los documentos. Finalmente, la similitud entre documentos es almacenada en los *Repositorios* de la Figura 1 como una matriz $M_{2x2}(\mathbb{R})$.

3.2 Generador de perfiles

Los perfiles se generan tanto para usuarios registrados como no registrados. Cuando el usuario no está registrado, se crea un perfil utilizando las cookies del navegador. El Generador de Perfiles recibe como entrada las acciones implícitas (información capturada a través de las acciones que realiza el usuario sobre los resultados de la búsqueda) y explícitas (temas de interés y datos personales registrados por los usuarios en el sistema, por ejemplo: datos demográficos, fecha de nacimiento, sexo, educación y temáticas favoritas). El perfil del usuario P_u es descrito por los siguientes elementos:

Id_Usuario, Profesión, Fecha_Nacimiento, Localidad, Consulta, Documentos_Consultados, Grado_Interés, Temáticas_Preferidas, Factor_Olvido, Expiración_Cookie.

Para la generación de los perfiles de los usuarios se aplica la técnica de aprendizaje automático basada en el Modelo Espacio-Vectorial, donde las variables de P_u se representan como vectores.

3.3 Personalización de la Recuperación

El proceso de la personalización de los resultados de la recuperación en el MRIP está basado en un análisis que combina la modelación de las relaciones Usuario-Consulta, Usuario-Documentos y Usuario-Temática, a partir de las preferencias capturadas del usuario. Para generar *Grado_Interés*, P_u necesita de tres variables P_u (u_id_j, u_iq_j, u_it_j), donde u_id_j representa las acciones realizadas por el usuario u_i sobre los documentos consumidos d_j, u_iq_j representa las consultas q_j realizadas por el usuario u_i y u_it_j representa el porciento de búsqueda del usuario u_i por la temática t_j.

Correlación Usuario-Documento (u_id_j). Se encarga de predecir cuáles de los documentos d_i puede ser interesante para el usuario u_i, basándose en un entorno de conocimiento que representa las acciones de u_i sobre d_i. Este entorno es creado basándose en el grado de interés que tiene un usuario sobre un documento recuperado, capturado mediante un conjunto de reglas. En estas reglas se modela el comportamiento del usuario sobre un documento considerando acciones tales como: *Like (L), Dislike (DL), Share (S), Do not share (NS), Visit (V),* y *Do not visit (NV)* y se infiere su grado de interés (GI). Las reglas definidas son:

1. R1: *Si L ∧ NV ∧ NS, entonces GI = Irrelevante*;
2. R2: *Si DL ∧ NV ∧ NS, entonces GI = Irrelevante*;
3. R3: *Si S ∧ NV, entonces GI = Irrelevante;*
4. R4: *Si S ∧ DL ∧ NV, entonces GI = Irrelevante;*
5. R5: *Si DL ∧ V ∧ NS, entonces GI = Relevancia baja;*
6. R6: *Si S ∧ L ∧ NV, entonces GI = Relevancia baja;*
7. R7: *Si S ∧ V, entonces GI = Relevancia Media;*
8. R8: *Si S ∧ DL ∧ V, entonces GI = Relevancia Media;*
9. R9: *Si S ∧ L ∧ V, entonces GI = Relevancia Alta; y*
10. R10: *Si NS ∧ L ∧ V, entonces GI = Relevancia Alta.*

El subproceso por otra parte utiliza una matriz donde se representan las reglas aplicadas por los usuarios sobre los documentos a partir de una consulta. Luego predice a partir del grado de similitud entre los documentos, las reglas que puede tener en cuenta un usuario con respecto a un documento que no ha visitado.

Finalmente, para buscar la similitud entre los vectores de calificación de dos usuarios *u* y *v* se utiliza el coeficiente de correlación de Pearson (Desrosiers y Karypis, 2011). La fórmula 1 muestra como calcular el Coeficiente de correlación de Pearson.

$$\mathrm{PC}(u,v) = \frac{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_u)(r_{vi}-\bar{r}_v)}{\sqrt{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_u)^2\sum_{i\in I_{uv}}(r_{vi}-\bar{r}_v)^2}} \quad (1)$$

donde:
r_{ui} : valoración que ha dado el usuario u documento i.
r_{vi}: valoración que ha dado el usuario v al documento i.
$\bar{r}_u$ y $\bar{r}_v$: total de documentos valorados en común por los usuarios u y v respectivamente.

Correlación Usuario-Consulta (u_iq_j). Este subproceso es el encargado de medir los usuarios similares al usuario actual a partir de una matriz donde se almacenan la similitud entre las consultas. Posteriormente al igual que el subproceso visto anteriormente busca la similitud entre el usuario actual y los demás usuarios utilizando el Coeficiente de correlación de Pearson.

Correlación Usuario-Temática (u_it_j). Tiene la responsabilidad de identificar los usuarios u_i que son similares al usuario actual a partir del porciento de temáticas T (Economía, Justicia, Biología, Ciencias de la información, etc.), preferidas por cada uno. Al igual que los subprocesos anteriores, busca la correlación lineal entre el usuario actual y los demás usuarios utilizando el Coeficiente de correlación de Pearson.

Decisor. Se encarga de decidir finalmente el grado de interés de un usuario sobre un documento. Luego de obtener la similitud entre el usuario actual y los demás por cada uno de los subproceso anteriores, este calcula un promedio de similitud general $sim(u,v)$. Posteriormente selecciona los *k-vecinos* usando la técnica *Máximo Número de Vecinos* la cual consiste en seleccionar los *k* usuarios que son más similares

al usuario activo, donde *k* será un parámetro del algoritmo. Finalmente, para predecir el Grado de Interés que un usuario *u* tendría sobre un documento d_i que no ha visitado, fue usado la fórmula de la *Media Ponderada* (fórmula (2)) (Ricci, Rokach, Shapira, et al., 2010):

$$\mathrm{MP(u,i)} = \frac{\sum_{v \in G_{u,i}} sim(u,v) * r_{v,i}}{\sum_{v \in G_{u,i}} sim(u,v)} (2)$$

donde:
v: grupo de usuarios que han valorado un documento *i*.
$r_{v,i}$: voto del usuario *v* al documento *i*.
sim(u,v): valor de correlación calculado anteriormente entre el usuario *u* y *v*.

Debido a que el interés del usuario sobre un documento puede disminuir, se incluye el cálculo del factor olvido puesto en práctica por (Wang, Li., Lin, et al., 2017).

$$F(d) = e^{\frac{\log_2(\Delta t)}{f}} \quad (3)$$

donde:
f: cantidad de días
Δt: período de tiempo entre la última actualización del grado de interés sobre el documento di y el grado de interés actual.

Este valor es actualizado en el campo *Factor_Olvido* registrado en el perfil del usuario. El nuevo grado de interés (NGI) sobre el documento d_j se calcula:

$$NGI(d) = VGI(d) * F(d) \quad (4)$$

donde:
VGI(d): viejo grado de interés sobre d_j.
F(d): factor olvido calculado en la fórmula 2.

3.4 Generador de ranking

El componente *Generador de Ranking* es responsable de hacer un nuevo ranking entre los documentos recuperados por el MRIP a partir de una consulta. Posteriormente, consulta la matriz de documentos con sus similitudes (ya almacenadas en los repositorios) y finalmente devuelve los documentos de mayor similitud en la parte superior de los resultados.

4 Caso de estudio: Red Cuba

El modelo propuesto fue implementado e integrado a un buscador web denominado Red Cuba (https://www.redcuba.cu/) con el objetivo de evaluar su aplicabilidad en un escenario real. Red Cuba está basado en un modelo de Recuperación de Información Híbrido (Booleano-Espacio vectorial) que se basa en el análisis de la frecuencia de los términos en los documentos. Está llamado a ser la principal fuente de acceso a información cubana en Internet y actualmente cuenta con más de 2 millones de contenidos indexados, de los cuales más del 90% no son indexados por buscadores internacionales. Se realizó un experimento en el que participaron 12 expertos en diferentes temáticas, (Educación, Turismo, Comercio, Juegos, Cultura, Cocina, Deporte, Política, Tecnología, Moda, Dirección, y Pedagogía), cada uno de los cuales definió una consulta asociada a su área de experiencia. Se propuso de esta forma para lograr tener un control de las consultas asociadas a documentos que responden a temáticas de interés para un experto.

Cada usuario seleccionó los documentos más relevantes (DR) para dicha consulta (respuesta deseada), según lo indexado en el buscador; de ellos solo 6 crearon su perfil (P) en el buscador. En la Tabla 1 se describen y caracterizan dichas consultas.

Id.	Palabras claves	P	DR
Q1	Libro, autor, P. J. Deitel	Si	35
Q2	Hotel, La Habana, 5 estrellas	Si	30
Q3	Tienda, camisas, blancas	Si	39
Q4	Juegos, Android, ciencia	Si	48
Q5	Escritor, poemas, cubanos	Si	28
Q6	Pasta, Bocaditos, helado	Si	39
Q7	Director, equipo, beisbol, cubano, sub23	No	28
Q8	Aborígenes, Cuba	No	32
Q9	Creadores, computadora, cubana	No	38
Q10	Pelo, plancha, suavizador	No	12
Q11	Leyes, gaceta, oficial	No	28
Q12	Modelo, formación, integral, estudiantes	No	45

Tabla 1. Consultas y tipos de usuarios para el experimento

Las métricas de precisión y exhaustividad, fueron utilizadas para medir y comparar la eficacia del buscador Red Cuba, aplicando el modelo propuesto y sin aplicarlo, tomando de referencia los documentos relevantes identificados por los expertos. En la medición de los resultados solo se tuvo en cuenta los 50

primeros documentos recuperados y los resultados se muestran en la Figura 2 y 3.

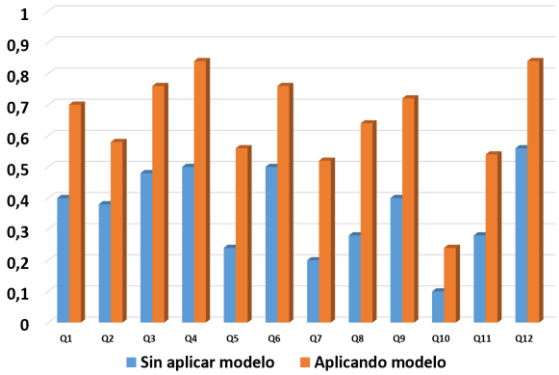

Figura 2. Resultados de la precisión

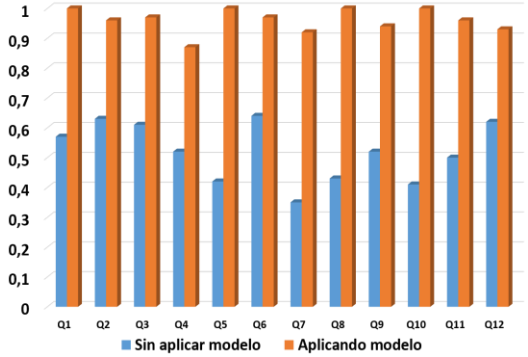

Figura 3. Resultados de la exhaustividad

Se puede observar como los resultados de precisión y exhaustividad aplicando el modelo al motor de búsqueda son mejores que sin aplicarlo, lo que demuestra que la integración de todos estos componentes contribuye a mejorar los resultados. Se observa además como los resultados de exhaustividad en el buscador de las consultas Q1, Q5, Q8 y Q10 aplicando el modelo llegan a tomar valores de 1. Esto constituye un resultado positivo ya que se logra maximizar el nivel de satisfacción del usuario en estas consultas.

Se está trabajando en la publicación de un subconjunto de datos que permita replicar los experimentos. Mientras tanto se podría replicar el experimento rastreando el buscador e indexando sus contenidos.

5 *Conclusiones*

Con los elementos teóricos y prácticos más actuales en el campo de los buscadores web y los sistemas de personalización, se desarrolló un modelo para la recuperación de información personalizada basada en el procesamiento semántico del contenido. El modelo propuesto además de ser un sistema que utiliza técnicas características de sistemas de personalización y de web semántica, posee mecanismos que permiten personalizar los resultados basándose en el perfil del usuario y el significado que tiene el contenido de su preferencia con el que interactúa. El uso de la semántica propicia tener la disponibilidad de un importante cúmulo de conocimiento multi-dominio. La evaluación final comprobó que la personalización de los resultados, guiando al usuario desde el inicio con sus intereses y conociendo el sentido semántico de las consultas y documentos, se logran mejores resultados de búsqueda, disminuyendo en gran medida la sobrecarga de información innecesaria.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto METODOS RIGUROSOS PARA EL INTERNET DEL FUTURO (MERINET), financiado por el Fondo Europeo de Desarrollo Regional (FEDER) y el Ministerio de Economía y Competitividad (MINECO), Ref. TIN2016-76843-C4-2-R.

6 *Bibliografía*

Chunzi, W. y B. Wang. 2017. Extracting Topics Based On Word2vec and Improved Jaccard Similarity Coefficient. En *Proceedings of the IEEE 2nd International Conference On Data Science in Cyberspace*, páginas 389-397.

Corcoglioniti, F., M. Dragoni y M. Rospocher. 2016. Knowledge extraction for information retrieval. En *Proceedings of the International Semantic Web Conference. Springer, Cham*, páginas 317-333.

Desrosiers, Ch. y G. Karypis. 2011. A Comprehensive Survey of Neighborhood-Based Recommendation Methods. *Recommender Systems Handbook. Springer*, páginas 107-144.

Hahm, G. J., Yi, M. Y., Lee, J. H., Suh, H. W. 2014. A personalized query expansion approach for engineering document retrieval. *Advanced Engineering Informatics*, 28(4): 344-359.

Johnson., M.S. 2016. Personalized Recommendation System for Custom Google Search. *International Journal of Computer & Mathematical Sciences*, 5:2347–8527.

Jay, P., P. Shah, K. Makvana y P. Shah. 2015. Review On Web Search Personalization Through Semantic Data. En *Proceedings of the IEEE International Conference On Electrical, Computer and Communication Technologies*, páginas 1-6.

Klusch, M., P. Kapahnke, S. Schulte, F. Lecue y A. Bernstein. 2016. Semantic Web Service Search: A Brief Survey. *Ki-Künstliche Intelligenz*, 30(2):139-147.

Mihalcea, R. y C. Strapparava. 2006. Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. AAAI, páginas 775-780.

Moro, A., F. Cecconi, R. Navigli. 2014. Multilingual Word Sense Disambiguation and Entity Linking for Everybody. En *Proceedings of the International Semantic Web Conference*, páginas 25-28.

Makwana, K., J. Patel y S. Parth. 2017. An Ontology Based Recommender System to Mitigate the Cold Start Problem in Personalized Web Search. En *Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems. Springer*, páginas 120-127.

Navigli, R., Ponzetto, S.P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193: 217-250.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards Wider Multilinguality. En *Proceedings of the International Conference On Language Resources and Evaluation Lrec2012*.

Preetha, S. y V. Shankar. 2014. Personalized search engines on mining user preferences using click through data. En *Proceedings of the Information Communication and Embedded Systems, IEEE*, páginas 1-6.

Ricci, F., L. Rokach, B. Shapira y P.B. Kantor. 2010. Recommender Systems Handbook. *Springe*r.

Singh, A., N. Dey, A. Ashour y V. Santhi. 2017. Web Semantics for Personalized Information Retrieval. *IGI Global. Information Science Reference*, páginas 166-186.

Shou, L., H. Bai, K. Chen y Ch. Chen. 2014. Supporting Privacy Protection in Personalized Web Search. *IEEE Transactions On Knowledge and Data Engineering*. 26(2): 453-467.

Sharma, S. y V. Rana. 2017. Web Personalization through Semantic Annotation System. *Advances in Computational Sciences and Technology*, 10(6):1683-1690.

Shafiq, O., R. Alhajj. y J. G. Rokne. 2015. On personalizing Web search using social network analysis. *Information Sciences*, 314: 55-76.

Tanaka, Y., Spyratos, N., Yoshida, T., Meghini, C. 2015. En *Proceedings of the Information Search, Integration and Personalization*. páginas 1–2.

Verma, D. y B. Kochar. 2016. Multi Agent Architecture for Search Engine. *International Journal of Advanced Computer Science and Applications*, 7(3): 224-229.

Wang, M., Q. Li., Y. Lin, y B. Zhou. 2017. A personalized result merging method for metasearch engine. En *Proceedings of the 6th International Conference on Software and Computer Applications. ACM*, páginas 203-207.

Zhou, D., S. Lawless, X. Wu1, W. Zhao y J. Liu. 2016. Enhanced Personalized Search Using Social Data. En *Proceedings of the Conference On Empirical Methods in Natural Language Processing*, páginas 700-710.

Zhang, H., M. Yan-hong, M. Wei-jun, y B. Zhong-xian. 2013. Study of Distributed Personalized Search Engine. *Advanced Materials Research. Trans Tech Publications*, páginas 1035-1039.