Metodología para Asistir al Usuario Web en su Búsqueda de Información en un Dominio Específico

María Romagnano[1], Martín Marchetta[2], Pablo Dominguez[3]

[1] Instituto de Informática - FCEFyN - Universidad Nacional de San Juan (UNSJ)
Av. Ignacio de la Roza 590 (O) – Rivadavia – San Juan – Argentina.
[2] Centro Universitario – FI – Universidad Nacional de Cuyo (UNCuyo)
Mendoza – Argentina.
[3] Departamento de Informática - FCEFyN - Universidad Nacional de San Juan (UNSJ)
Av. Ignacio de la Roza 590 (O) – Rivadavia – San Juan – Argentina.
maritaroma@iinfo.unsj.edu.ar, mmarchetta@fing.uncu.edu.ar, ivandominguez@gmail.com

Abstract. *The search of information in the Web has become part from our daily routine. To read the newspaper, to know the climate, to look for information on a place where to travel. The Web, digital repository, it allows to count quickly and easy with a "fan of information". Nevertheless, this accessibility could become a problem if thousands of answers to the consultation appear and no of them is satisfactory (noise) or there is not answer (silence). Our contribution consists of a methodology to attend the usuary Web in its search of information in a determined dominion of application, reducing the difficulty in the exploration, improving the precision and the response time.*

Resumen. *La búsqueda de información en la Web se ha vuelto parte de nuestra rutina diaria. Leer el periódico, conocer el clima, buscar información sobre un lugar donde viajar. La Web, repositorio digital, permite contar con un "abanico de información" fácil y rápidamente. Sin embargo, esta accesibilidad podría convertirse en un problema si se presentan miles de respuestas a la consulta y ninguna de ellas es satisfactoria (ruido) o no se encuentra respuesta (silencio). 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 exploración, mejorando la precisión y el tiempo de respuesta.*

1. Introducción

La información con la que se cuenta es una de las herramientas más poderosa que tiene la sociedad. Se puede considerar como instrumento para tomar decisiones en la vida de un paciente, en el funcionamiento de una empresa, como estrategia de guerra, para realizar un viaje, etc.

Torres menciona "¿Vivimos en una época de cambios, o un cambio de época? ¿Cómo caracterizar las profundas transformaciones que acompañan la acelerada introducción en la sociedad de la inteligencia artificial y las nuevas tecnologías de la información y la comunicación (TICs)? ¿Se trata de una nueva etapa de la sociedad

industrial, o estamos entrando en una nueva era? Aldea global, era tecnotrónica, sociedad postindustrial, era o sociedad de la información y sociedad del conocimiento son algunos de los términos que se han acuñado en el intento por identificar y entender el alcance de estos cambios. Pero mientras el debate prosigue en el ámbito teórico, la realidad corre por delante y los medios de comunicación eligen los nombres que hemos de usar" [Torres, 2005].

Entonces, 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". Por lo tanto, 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. Se debería saber precisamente qué es lo que se está buscando y cuáles son las palabras claves con las que se debe realizar la búsqueda.

Por otra parte, 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. Al mismo tiempo, todavía existen recursos web que permanecen inexplorados [Mendez Duque, Chavarros Porras y Moreno Laverde, 2007]. Además, se suma a las problemáticas ya planteadas, la complejidad de ciertos dominios, donde la mayor parte de la información se encuentra distribuida en varios sitios web; almacenada usando formatos heterogéneos.

Por lo tanto, concretamente, a la hora de obtener información precisa de la Web se plantean ciertos inconvenientes desde distintas perspectivas:

- *Perspectiva del usuario*. Generalmente, el acceso a la información se hace a través de buscadores (Google, Yahoo, Bing, etc.) donde se ingresan palabras claves. En la mayoría de los casos el resultado es un conjunto de miles de documentos que contienen esas palabras, por lo que se hace difícil saber específicamente cuál o cuáles de esos documentos contemplan mejor la búsqueda. Uno de los principales factores que influye en la recuperación de la información tiene que ver con el usuario y su carencia a la hora de expresar en la consulta lo que él necesita. Así, cuando un usuario busca información en la web, debido a la diversidad del lenguaje natural e idiomas, quizás existan dos o más documentos que traten el mismo tema, lo cual es desconocido por parte de él [Scime, 2005]. Por otra parte, aún, podría sumarse la problemática de que si éste no es experto realizando búsquedas obtenga un número importante de resultados y el tiempo de respuesta sea considerable. Esto puede deberse a que los usuarios pocas veces ocupan las herramientas provistas por los buscadores para filtrar resultados, realizar búsquedas avanzadas o simplemente realizar búsqueda con palabras específicas de la temática.

[Cacheda y Viña, 2001] plantean que el principal problema que afecta al usuario en su interacción con los buscadores es la forma en la que ellos especifican su consulta y la forma en la que interpretan los resultados. Generalmente, los usuarios web realizan búsquedas "muy por encima", simplemente se limitan a abrir las páginas cuyos títulos y descripción se ajusten mejor a sus necesidades. Además, probablemente el usuario

desconoce las características de cada uno de los buscadores, usando el más popular socialmente, pero que posiblemente no es el que mejor información le brinda ante sus necesidades. Existen distintos tipos de buscadores: motores de búsqueda, directorios o índices temáticos, metabuscadores y multibuscadores; cada uno de los cuales establece distintas formas de realizar la búsqueda y sus respuestas son más o menos precisas, en menor o mayor cantidad, organizadas o clasificada de distinta forma, según determinados criterios.

Ya en 1998, Florescu, Levy y Mendelzon mencionaron: "El gigantesco crecimiento de la web, principalmente en cuanto al acceso a la información en Internet, se ha convertido en una cuestión que va más allá de determinar dónde está la información; pretendiendo saber de qué forma y cómo llegar a ella" [Florescu, Levy y Mendelzon, 1998].

Por los motivos planteados, toda la rapidez adquirida accediendo a la información a través de la Web, puede desvanecerse cuando se pierde tiempo tratando de discernir qué "de todo lo que se encontró" satisface correctamente las necesidades emergentes.

- *Perspectiva de la Web*. Tradicionalmente las consultas se realizaban a una base de datos local o remota con un esquema bien conocido. En los sistemas de información actuales las consultas implican fuentes de información heterogéneas, que no sólo pueden estar dispersas geográficamente, sino que también suelen tener esquemas diferentes. La información suele encontrarse como documentos, formularios, foros, listas de interés, imágenes, videos, etc. Más aún, estas fuentes plantean dificultades en cuanto a la relevancia, cantidad, calidad y disponibilidad de información; debido a que la dinamicidad de la Web hace que en un período de tiempo se cuente con ciertos documentos y en otro período éstos no estén disponibles [Baeza-Yates y Ribeiro-Neto ,1999].

De igual modo, el crecimiento en cuanto a cantidad de recursos web y los cambios de formatos de la información circulantes en la Web son enormes en poco tiempo, ocasionando que las aplicaciones desarrolladas se vuelvan rápidamente obsoletas.

Extractando, y teniendo en cuenta los inconvenientes planteados, una persona podría pasarse días leyendo cada uno de los miles de resultados o simplemente elegir al azar uno de los primeros y que quizás nada tenga que ver con sus intereses o que no le convence del todo. Con esta experiencia el usuario web queda desanimado para buscar nuevamente información en esta gran biblioteca digital. Este escenario 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. Es obvio que esta situación es indeseable para el interesado; por una parte sobrecargado con los resultados encontrados, visitando varios sitios web para reunir la información y/o productos que desea, y por otro lado quizás nada de lo que encontró satisface sus necesidades.

Existe una clara necesidad de contar con una "herramienta" que ayude al usuario en el proceso de búsqueda, una aplicación que le permita la recuperación automática de la información, brindando respuestas precisas, acordes a sus necesidades, sin tener que

pasar por la complejidad de discernir qué respuesta o respuestas son más convenientes de todas las que se le ofrecen; y a la vez en el menor tiempo posible.

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.

Si bien en trabajos recientes se han sugerido distintas formas de unificar y agrupar fuentes de información similares, tratando de dar solución a ciertos problemas planteados, todavía, existen dominios más complejos que otros, donde encontrar la solución es un problema aún más complejo; debido a la cantidad y variedad de información que se maneja y porque el conjunto de fuentes que se consultan es considerable. 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. Con esta contribución el usuario logra éxito en sus búsquedas y no se enfrenta a una gran cantidad de resultados.

El resto del artículo se estructura de la siguiente forma: en la sección 2 se presentan los trabajos relacionados. La sección 3 describe la metodología propuesta de recuperación de la información para un dominio de aplicación con características específicas. En la sección 4 se aplica la propuesta a un dominio concreto. Finalmente, la sección 5 presenta las conclusiones y trabajos futuros.

2. Trabajos Relacionados

Uno de los problemas a los que la humanidad se ha tenido que enfrentar desde la invención de la escritura es el almacenamiento y la posterior recuperación de la información. Con la aparición de las nuevas tecnologías de información y comunicación (NTICs) este problema se ha resuelto parcialmente. Es más fácil producir datos que guardarlos, administrarlos y recuperarlos [Li y Cao, 2014].

Para llevar a cabo la tarea de brindar respuesta a la consulta de los usuarios, algunos autores han usado diversas técnicas y/o herramientas relacionadas con web mining, recuperación de la información, clasificación, clustering, ontologías, web semántica, inteligencia artificial, entre otras. Así [Özel, 2011] y [Baykan, Henzinger y Weber, 2013] proponen clasificar páginas web a través de algoritmos genéticos, usando las URLs de las páginas, n-gramas, teniendo en cuenta el contenido y la semántica de las páginas. En [Chen, Bau y Tsai, 2010], [Ramage, Heymann, Manning y Garcia Molina, 2009], [Shelke, Sadavarte, Dhurjad y Pandit, 2012], [Patel y Zaveri, 2011], [Liu, Yu, Xu y Shi, 2012], [Ghosh y Kumar, 2013] y [Sote y Pande, 2015] se realiza clustering de páginas web considerando palabras claves, contenido semántico, ontologías y etiquetas como herramientas externas y mapas auto organizados, K-Means y C-Means como métodos estándares para realizar clustering hard y fuzzy respectivamente. En [Matsumoto y Hung, 2010] y en [Xiao y Hung, 2008] además de permitir el solapamiento entre clusters proponen la idea de agrupar los resultados de búsqueda web ya generados por motores de búsqueda convencionales. [Qu, Wei, Wang, y Liu, 2011],

proponen la arquitectura y el diseño de un sistema de recuperación basado en Web Semántica. [Hegde, 2011] trata de agrupar páginas web independizándose de los clásicos algoritmos de clustering, usando el diccionario de Internet para contextualizar las palabras claves. [Hernández, Rivero, Ruiz y Corchuelo, 2012] proponen usar una técnica estadística para clusterizar páginas web y [Yue et al., 2015], además de una ontología y HowNet, usa el modelo SVD para clusterizar. En el trabajo de [Rekik y Kallel, 2013] se usa lógica difusa para evaluar sitios web y en [García-Plaza, Fresno y Martínez, 2008] se propone una representación basada en lógica borrosa para clusterizar páginas web. El trabajo de [Kamath, Piraviperumal, Meena, Karkidholi, y Kumar, 2013] 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 [Romagnano, Dominguez, Marchetta, 2015] se propuso un agente de filtrado que localiza y agrupa fuentes de información de acuerdo a los servicios que ofrecen.

Aunque la problemática relacionada con la búsqueda web pareciera estar solucionada por varios trabajos existentes, todavía no se ha propuesto una aplicación que permita al usuario hacer una búsqueda, simulando un buscador convencional que brinde respuestas precisas, sin pérdida de información y en menor tiempo. Los trabajos mencionados anteriormente, en su mayoría, si bien plantean soluciones para agrupar páginas web haciendo mención al uso de distintas técnicas de clasificación y/o clustering, en algunos casos contemplan dominios muy específicos al cual dan una solución concreta. En otros casos, si bien se plantean soluciones más generales, se usan procedimientos complejos; llevando a mayor tiempo de respuesta. Además, en aquellos donde se propone clusterizar, los clusters no se forman desde la perspectiva de reducir la complejidad y el tiempo en la posterior búsqueda web.

3. Metodología Propuesta

Cuando se realiza una búsqueda, más allá de establecer palabras claves, usar comillas u operadores booleanos, existe la posibilidad de encontrarse con documentos que nada tienen que ver con la búsqueda deseada, como lo demuestra la Figura 1 donde al realizar la búsqueda, en febrero de 2017, con las palabras claves *Turismo San Juan Argentina*, puedo observarse que en la página N° 14 de Google se muestra un resultado que no tiene relación con lo que se estaba buscando. Por otra parte, existen documentos web que se repiten más de una vez; y aún más significativo, documentos que tienen información pertinente a la búsqueda realizada y los mismos se encuentran en las últimas páginas del buscador, como se expone la Figura 2 en la que el documento del Parque Nacional San Guillermo se encuentra en la página Nº 15 y cuyo contenido es abundante en información turística del lugar. Obviamente, este documento raramente será visitado por un interesado; seguramente abandona la búsqueda antes de llegar a él.

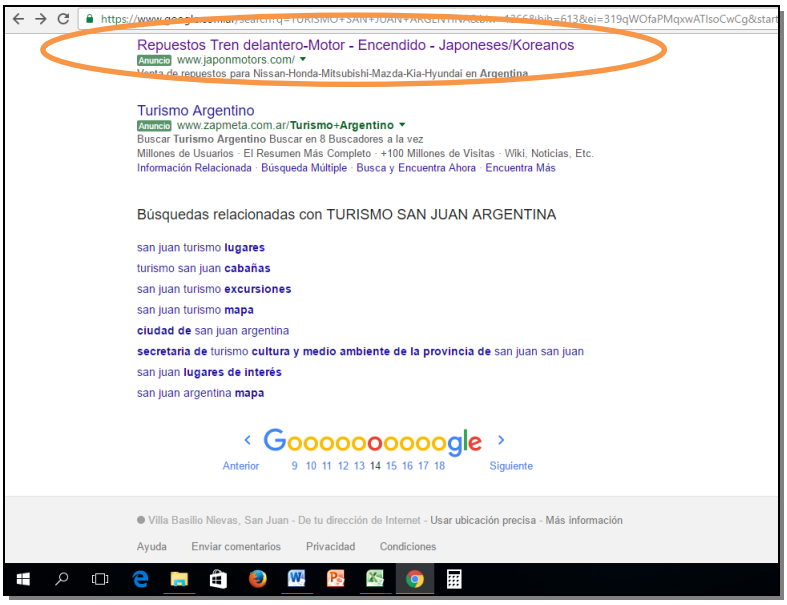


Figura 1. Resultado incorrecto de la búsqueda, en una posición considerable de ranking

Figura 2. Resultado correcto de la búsqueda, pero en una posición no considerable de ranking

Por lo tanto, la propuesta consiste en una metodología que localiza fuentes de información en la web por medio de los buscadores; las agrupa de acuerdo a los servicios que ofrecen y luego a través de una aplicación web brinda al usuario información precisa, acorde a sus necesidades; reduciendo la complejidad y mejorando la precisión y el tiempo de respuesta (Figura 3).

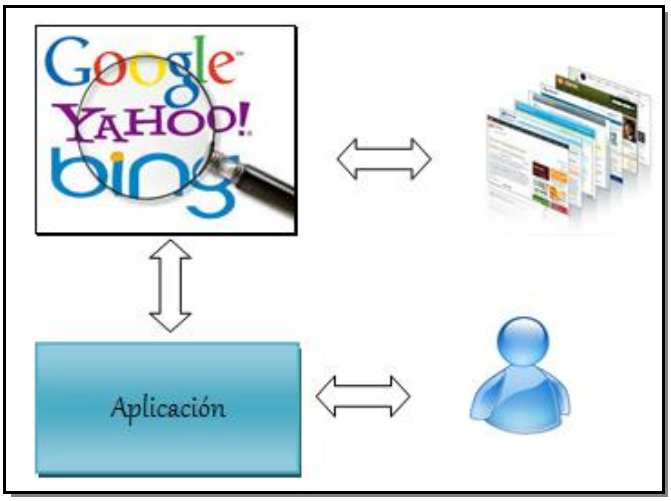


Figura 3. Interacción buscadores-aplicación-usuario

La metodología sugiere tres gestiones principales (Figura 4), abarcando desde el relevamiento de los documentos web (Recuperación – Tiempo Diferido), procesamiento y agrupamiento de los mismos (Agrupamiento – Tiempo Diferido), hasta resolver la consulta del usuario (Vista Unificada – Tiempo Real). Luego éstas se desglosan en etapas más específicas (Figura 5).

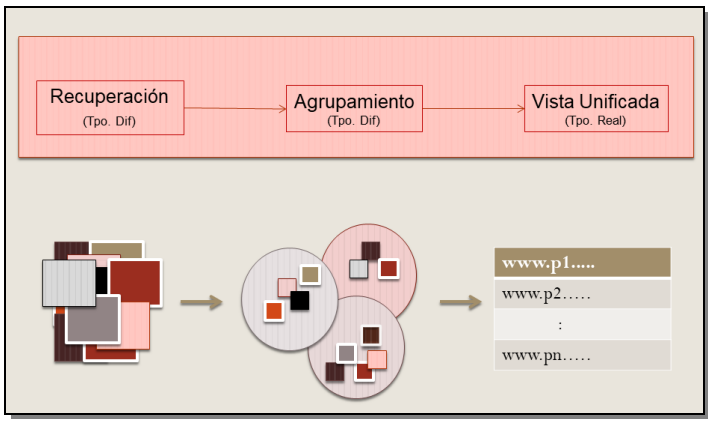


Figura 4. Acciones principales de la metodología

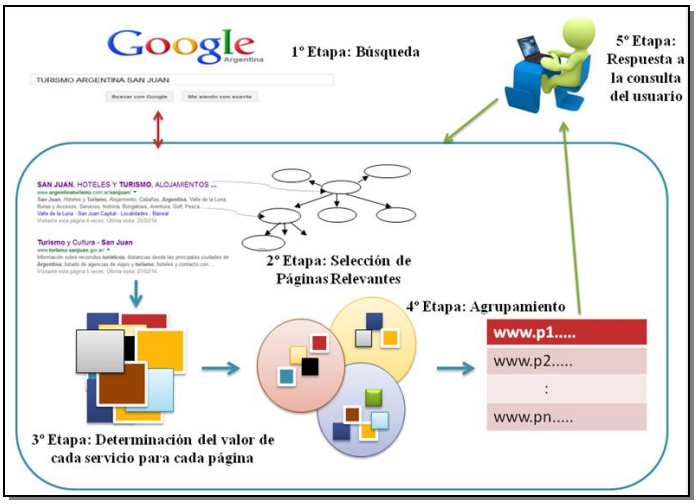


Figura 5. Etapas de la metodología

Entonces, unificando, la Figura 6 presenta una visión general de la propuesta.

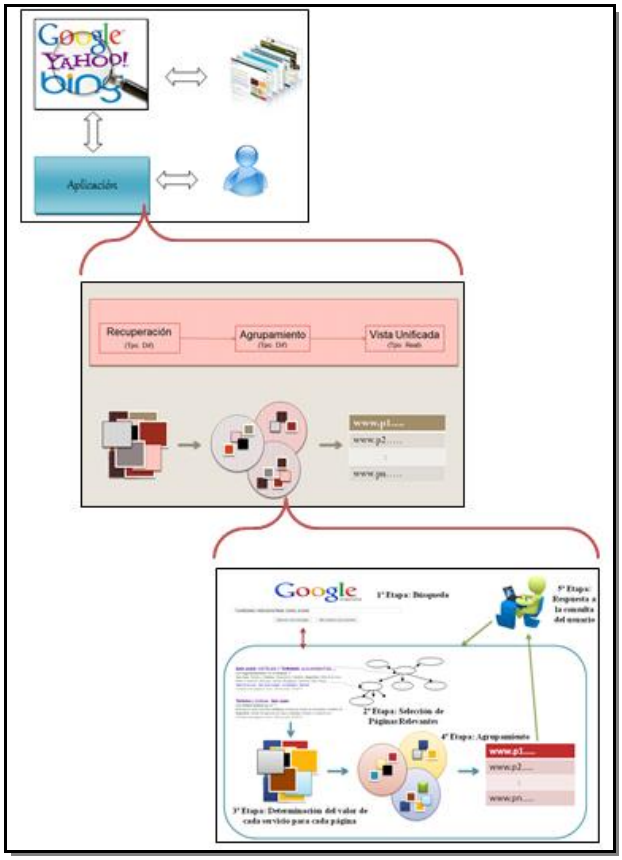


Figura 6. Visión general de la propuesta

Nuestra contribución consiste en 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. Es decir que, el sistema a través de las APIs de buscadores y/o índices temáticos realiza la búsqueda de los servicios de un dominio específico, con la ayuda de ciertas palabras claves. El motivo principal de por qué usar herramientas existentes para buscar en la Web, es porque nuestra propuesta no compite contra los buscadores, meta buscadores o índices existentes; sino que la idea es hacer uso de éstos logrando exactitud y a su vez ganancia en tiempo. Posteriormente, el sistema deberá seleccionar los documentos relevantes al dominio en cuestión. Para realizar este pre-procesamiento se toman los resultados arrojados por los buscadores y de cada uno de ellos se analiza su texto anclado y el URL (Figura 7).

El URL es una cadena de caracteres con la cual se asigna una dirección única a cada uno de los recursos de información disponibles en Internet. Existe un URL único para cada página de cada uno de los documentos de la WWW. El URL de un recurso de información es su dirección en Internet, la cual permite que el navegador la encuentre y la muestre de forma adecuada. El texto anclado es el texto del hipervínculo en el cual hacemos click para acceder a cada resultado arrojado por el buscador. Según [Brin y Page, 1998] el texto anclado, a menudo, provee una descripción más acertada de las páginas web que las páginas mismas. Además, pueden existir para documentos que no pueden ser indexados por un motor de búsqueda basado en texto, como ser imágenes, programas, y bases de datos.

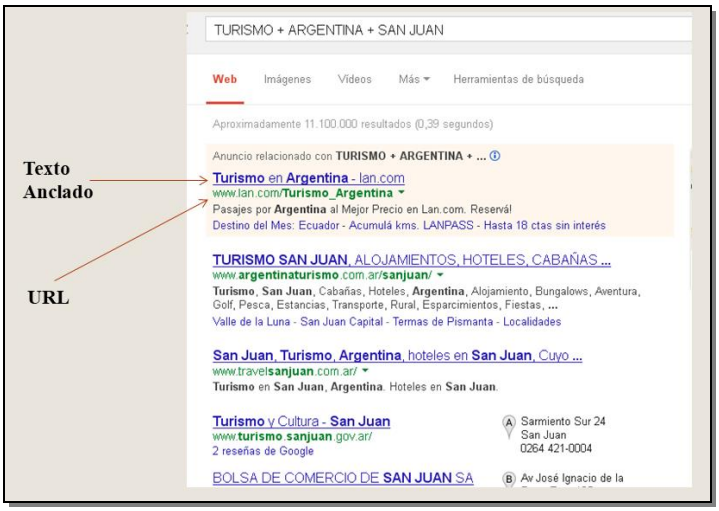


Figura 7. Texto anclado y URL en una búsqueda

En el análisis se usa una ontología preexistente del dominio como base de contrastación. Además se llevan dos contadores (positivo y negativo). Cada término que aparece en el texto anclado y/o en el URL del documento en cuestión, se contrasta con la ontología. 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.

Luego si el contador positivo es mayor que el contador negativo, entonces se selecciona el documento como relevante; de lo contrario el documento se descarta (Figura 8).

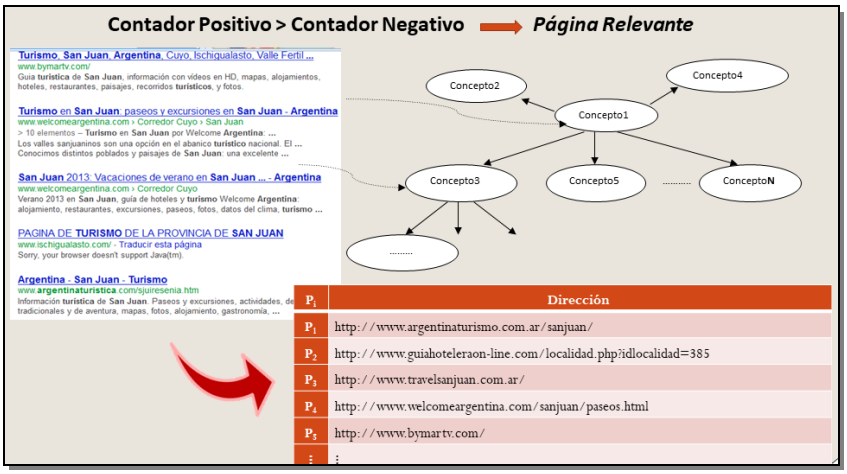


Figura 8. Procedimiento para determinar si un documento es relevante

Por otra parte, para reforzar la decisión de elegir un documento como relevante se calculó la cobertura como se muestra en la siguiente fórmula:

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

Cd_j: cobertura del documento d_j.
n: cantidad de términos.
t_{ij}: término i 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. UCd_j podrá variar de acuerdo al dominio en cuestión, dependiendo de la densidad de información del mismo.

De estos documentos almacenados se debe analizar su contenido para determinar qué servicios, productos y/o información ofrecen y cuál es la frecuencia de los mismos. Para un dominio concreto, se preestablecen los servicios y/o productos que se desean relevar. Para seleccionar dichos términos (servicios y/o productos) se emplean la bolsa de palabras (producto de la consulta a un experto del dominio). Luego estos términos serán los nombres de los grupos a los cuales, a partir de ahora, denominaremos G_i.

Luego, las frecuencias resultantes de cada término relevante deben almacenarse en registros que representan los valores de cada servicio y/o producto para cada documento. Estos registros se representan 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. Para calcular el peso w_{ij} se aplica el esquema TF (frecuencia del término) del Modelo Espacio Vectorial [Liu, 2007]:

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

Además se usa nuevamente la ontología para determinar la relación entre términos semejantes.

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}

Aquellos documentos que presenten información de un mismo servicio y/o producto deben agruparse. Sin embargo los documentos pueden presentar información de varios servicios y/o productos, por lo cual necesariamente la metodología debe permitir que un documento pueda corresponder a uno o a más grupos con un cierto grado de pertenencia.

Se decidió usar técnicas de clasificación y de clustering. Clasificación en el momento de determinar qué grupos o clases se desean tener. Esta tarea se realiza de ante mano y con ayuda del experto del dominio. Clustering, posteriormente, para permitir que un documento pueda pertenecer a más de una clase con un cierto grado de pertenencia. Esta opción está permitida por algunos métodos de clustering. Las técnicas de clasificación establecen que un objeto sólo puede pertenecer a una clase bien definida

de antemano. Por lo tanto, para llevar a cabo la tarea de clasificación se usa una técnica estándar tal como Árboles de decisión, K-vecinos más cercanos, Bayes, SVM, etc. En cuanto al agrupamiento, se propone un algoritmo, tomando como referencia métodos particionales y soft [Han y Kamber, 2006]. De los métodos particionales se adopta la simplicidad en cuanto a la idea de elegir un centro para cada grupo y concentrar los restantes elementos en función de la distancia de ellos con cada centro. Pero como éstos no permiten que un elemento pueda pertenecer a más de un cluster, se adopta esta posibilidad de los métodos soft.

Como puede observarse, nuestra propuesta adopta parte de estos dos métodos populares pero en nuestro caso se tiene definido de antemano cuáles serán los grupos o clusters que se desean tener, los cuales coinciden con los servicios preestablecidos como relevantes. Los centros se eligen una única vez y se mantienen durante todo el proceso. Para la elección se usa una ontología pre-existente del dominio y la herramienta léxica WordNet, disponible la web gratuitamente. Usando estos recursos externos se pueden reducir considerablemente los tiempos de procesamiento, logrando a su vez simplicidad y precisión en los cálculos. Se registra en una matriz y en una única iteración, la pertenencia de cada elemento a cada cluster y la dimensión de cada elemento coincide con la cantidad de clusters; disminuyendo la cantidad de operaciones a realizar por el algoritmo. De esta forma se arman o rearman los grupos y así se refleja el ABM (Alta, Baja y Modificación) de las fuentes de información web con las que cuenta el sistema para resolver una consulta.

El siguiente pseudocódigo del algoritmo bosqueja los pasos seguidos por la metodología para realizar las tareas de cálculo de similitudes y de agrupamiento.

3.1. Algoritmo

1. Para cada grupo (G_i) calcular el centro (c_i), como el máximo valor de cada columna. Si dos o más documentos tuviesen un valor máximo se calcula su densidad (De_j). 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 candidatos debido a que tienen máximo valor de columna y coinciden en densidad, se evalúa la información total (I_j) de cada documento candidato, como la cantidad total de términos relevantes, para desempatar.

Si aún siguen empatando, el sistema elige al azar uno de los documentos en competencia. La Tabla 2 brinda un esbozo de estos cálculos.

2. Mientras existan documentos por analizar:

> *2.1. Calcular la similitud/distancia de cada documento d_j con cada centro c_i. La similitud/distancia se puede calcular con alguno de los estándares conocidos: Similitud del Coseno, Distancia Euclídea, Distancia de Mahalanobis, Distancia de Hamming; o a través de la razón geométrica entre el término t_i en el documento d_j y el centro c_i.*

Esta última forma de cálculo de la similitud, propuesta por este trabajo de investigación, 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/distancia, como el promedio entre similitudes/distancias.
2.3. Por cada similitud/distancia 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 al umbral 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}. O si la distancia del documento d_j con el centro c_i es menor o igual al umbral 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.

Tabla 2. Cálculo de los centros c_i

D_j	URL	t_1	t_2	t_3	…	t_n	De_j	I_j
D_1	URL_1	w_{11}	w_{21}	w_{31}	…	w_{n1}	de_1	i_1
D_2	URL_2	w_{12}	w_{22}	w_{32}	…	w_{n2}	de_2	i_2
D_3	URL_3	w_{13}	w_{23}	w_{33}	…	w_{n3}	de_3	i_3
....	....	....	....	....	....	....	....	....
D_m	URL_m	w_{1m}	w_{2m}	w_{3m}	…	w_{nm}	de_m	i_m
Max	-	w_{11}	w_{22}	w_{3m}	…	w_{n1}		

Es importante destacar que toda esta tarea de pre-procesamiento web (web mining) y agrupamiento de documentos web se hace en tiempo *diferido*, significando en ganancia en tiempo de respuesta. Aunque probablemente, y debido a la dinamicidad de la web, se pierda en cantidad de respuestas, en cuanto a que los documentos pueden haber sido indexados y agrupados en un instante de tiempo y en otro puede que se hayan agregado a la web otros documentos; los cuales obviamente no fueron indexados por la aplicación. El inconveniente de que los documentos indexados no estén disponibles al momento de ser mostrados al usuario será comprobado por la aplicación antes de mostrar la respuesta al usuario.

Como lo muestra la Figura 9, en *tiempo real* el usuario tendrá la posibilidad realizar la consulta eligiendo dominio, país, provincia, localidad, servicio o servicios y/o producto o productos de los cuales requiere información.

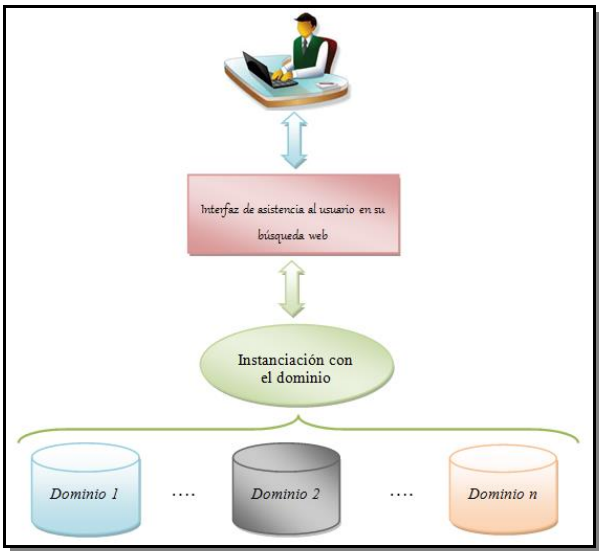


Figura 9. Instanciación de la aplicación con un dominio específico en función de la consulta del usuario

El sistema deberá determinar cuáles son los servicios que se le están solicitando. En función de esto deberá seleccionar el o los agrupamientos apropiados para brindar información calificada, precisa y relevante. Así se devolverá una lista de páginas web, sin que el interesado tenga que realizar él mismo la búsqueda, invirtiendo gran cantidad de tiempo y enfrentándose a numerosos resultados. Es decir, en esta etapa la metodología propone un prototipo de sistema que ante una consulta de un usuario, en lenguaje natural, se deberán realizar las siguientes tareas:

1. *Remover los sotpwords*.
2. *Realizar stemming*.
3. *Análisis del dominio*. Como el sistema tiene preestablecido que dominios puede cubrir, se debe determinar cuál de ellos es el que se encuentra en cuestión (Figura 9). Para realizar esta tarea el sistema deberá comparar los términos de la consulta y la ontología de cada uno de los dominios.
4. *Instanciar la sub aplicación del dominio*. Seleccionar la aplicación con el dominio comprometido.
5. Establecer, a través de *herramientas externas (DBpedia, WorNet, ontología, bolsa de palabras)*, 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. Es decir, si solo se requiere un término específico, el sistema deberá acceder al grupo que contiene información de ese término. Si se solicita información de más de un término, el sistema deberá acceder a la intersección de los grupos implicados.
6. Otorgar un listado de URLs.

El interesado, es decir quién hace uso de este sistema de recuperación de información, podría ser un usuario web u otro sistema mayor, digamos un sistema recomendador que requiere los documentos para armar paquetes de servicios y así poder realizar recomendaciones a los usuarios web.

La Figura 10 muestra la arquitectura del sistema de recuperación de información web propuesto. La *consulta del usuario* puede ser mediante palabras claves, usando operadores booleanos, a través de frases, por consulta de documentos completos o usando la forma más simple para el usuario; el lenguaje natural. El módulo de *sintaxis de consulta* coteja el orden y relación de los términos de la consulta, realiza el pre-procesamiento como la eliminación de stopword y stemming, etc. transformando la consulta del usuario en una consulta entendible para el sistema. Para realizar esta tarea dicho módulo se basa en el aporte del módulo de *herramientas externas*. El módulo de *indexación* es quien toma los documentos de la web (*fuentes de información web*), realiza el pre-procesamiento, con ayuda de las *herramientas externas*, para determinar cuáles son relevantes al dominio en cuestión, los agrupa obteniendo los documentos indexados y almacenados para permitir la posterior recuperación. El módulo de *rankeo* calcula la relevancia de cada documento indexado para la consulta solicitada. Además se tiene en cuenta el *historial de consulta* para establecer si se trata de una consulta repetida o no y así poder aportar más información para realizar el ordenamiento y listado de los resultados a entregar.

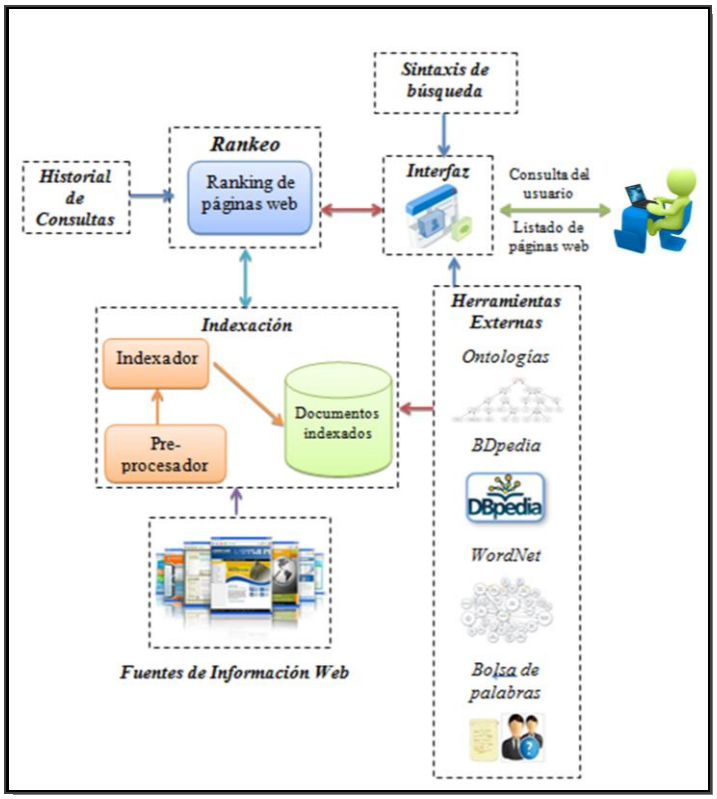


Figura 10. Arquitectura del sistema de recuperación

4. Aplicación al dominio del Turismo

Como se ha señalado en la Figura 9 el sistema puede instanciarse para buscar en un dominio específico. Como ejemplos, se pueden mencionar el dominio salud, educación o turismo. En esta sección se presentan los resultados de aplicar la propuesta al dominio

del turismo, en la provincia de San Juan. Cada una de las actividades principales planteadas en la metodología fue desglosada en etapas más específicas, que se describen a continuación.

En una primera etapa, una vez a la semana, por recomendación del experto en turismo, automáticamente se realizó la búsqueda de la información con las palabras claves TURISMO, SAN JUAN y ARGENTINA.

Posteriormente, en la segunda etapa se seleccionaron páginas relevantes. Se obtuvo una muestra de 486 páginas siguiendo los criterios establecidos de diferencia entre contadores y cobertura de términos.

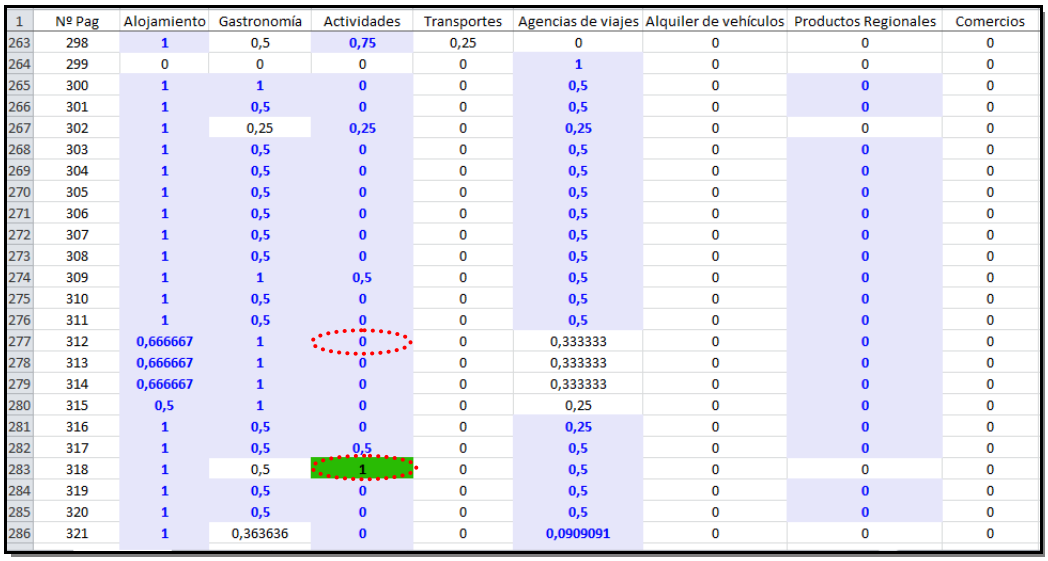
En la tercera etapa se seleccionaron términos relevantes con el asesoramiento de expertos de la Secretaría de Turismo, de la provincia de San Juan. Se estableció que los términos relevantes serían *Actividades, Agencias de Viajes, Alojamiento, Alquiler de Vehículos, Comercios, Gastronomía, Productos Regionales y Trasporte*; nombres de los futuros grupos que se formen y los que agruparan a los documentos seleccionados como relevantes.

En la cuarta etapa, se determinó el valor de cada servicio en cada documento. Es decir que para cada uno de los documentos seleccionado en la segunda etapa se calculó la frecuencia de cada uno de los términos relevantes. Luego se completaron los registros con los pesos calculados con dichas frecuencias y de acuerdo al esquema TF; teniendo en cuenta la aproximación entre términos similares.

En la quinta etapa, se realizó el agrupamiento. Para cada grupo se determinaron los centros, como lo establece la metodología. Posteriormente se calcularon las similitudes y el umbral de similitud de los documentos con cada uno de los centros.

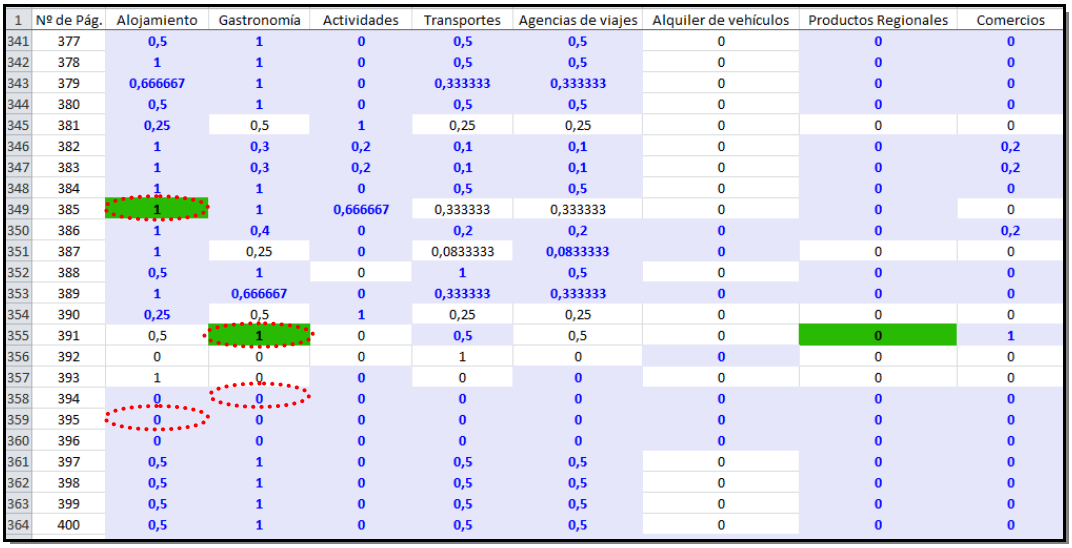
Para comparar la similitud (o distancia) entre dos documentos se realizaron pruebas con la fórmula del coseno y con la distancia euclídea, exponiendo una situación poco real. En ambos casos se provee una similitud (o distancia) general entre dos vectores, por lo tanto cuando se necesita puntualizar la similitud (o distancia) para un término específico entre dos documentos no se obtienen verdaderos resultados dado que un documento que no contiene un término determinado podría ubicarse en el grupo de ese término. Así, en la Figura 11, el sistema exporta hacia Excel una tabla con las similitudes calculadas, donde puede observarse que el centro del grupo Actividades es la página Nº 318 (celda sombreada en verde) y que entre los documentos considerados como similares (celdas sombreadas en celeste) se observan frecuencias de cero. Es decir, que se están contemplando como similares al centro ciertos documentos que no tienen términos coincidentes con ese grupo.

Lo mismo puede observarse en la Figura 12, donde se calcula la distancia euclídea entre el centro y el resto de los documentos. Para observar que esta problemática se reitera en varios casos, en dicha figura, esta vez, se muestra que sucede cuando el centro es Gastronomía o Alojamiento, por ejemplo.



1	Nº Pag	Alojamiento	Gastronomía	Actividades	Transportes	Agencias de viajes	Alquiler de vehículos	Productos Regionales	Comercios
263	298	1	0,5	0,75	0,25	0	0	0	0
264	299	0	0	0	0	1	0	0	0
265	300	1	1	0	0	0,5	0	0	0
266	301	1	0,5	0	0	0,5	0	0	0
267	302	1	0,25	0,25	0	0,25	0	0	0
268	303	1	0,5	0	0	0,5	0	0	0
269	304	1	0,5	0	0	0,5	0	0	0
270	305	1	0,5	0	0	0,5	0	0	0
271	306	1	0,5	0	0	0,5	0	0	0
272	307	1	0,5	0	0	0,5	0	0	0
273	308	1	0,5	0	0	0,5	0	0	0
274	309	1	1	0,5	0	0,5	0	0	0
275	310	1	0,5	0	0	0,5	0	0	0
276	311	1	0,5	0	0	0,5	0	0	0
277	312	0,666667	1	0	0	0,333333	0	0	0
278	313	0,666667	1	0	0	0,333333	0	0	0
279	314	0,666667	1	0	0	0,333333	0	0	0
280	315	0,5	1	0	0	0,25	0	0	0
281	316	1	0,5	0	0	0,25	0	0	0
282	317	1	0,5	0,5	0	0,5	0	0	0
283	318	1	0,5	1	0	0,5	0	0	0
284	319	1	0,5	0	0	0,5	0	0	0
285	320	1	0,5	0	0	0,5	0	0	0
286	321	1	0,363636	0	0	0,0909091	0	0	0

Figura 11. Calculo de la similitud entre documentos, a través de la fórmula del coseno



1	Nº de Pág.	Alojamiento	Gastronomía	Actividades	Transportes	Agencias de viajes	Alquiler de vehículos	Productos Regionales	Comercios
341	377	0,5	1	0	0,5	0,5	0	0	0
342	378	1	1	0	0,5	0,5	0	0	0
343	379	0,666667	1	0	0,333333	0,333333	0	0	0
344	380	0,5	1	0	0,5	0,5	0	0	0
345	381	0,25	0,5	1	0,25	0,25	0	0	0
346	382	1	0,3	0,2	0,1	0,1	0	0	0,2
347	383	1	0,3	0,2	0,1	0,1	0	0	0,2
348	384	1	1	0	0,5	0,5	0	0	0
349	385	1	1	0,666667	0,333333	0,333333	0	0	0
350	386	1	0,4	0	0,2	0,2	0	0	0,2
351	387	1	0,25	0	0,0833333	0,0833333	0	0	0
352	388	0,5	1	0	1	0,5	0	0	0
353	389	1	0,666667	0	0,333333	0,333333	0	0	0
354	390	0,25	0,5	1	0,25	0,25	0	0	0
355	391	0,5	1	0	0,5	0,5	0	0	1
356	392	0	0	0	1	0	0	0	0
357	393	1	0	0	0	0	0	0	0
358	394	0	0	0	0	0	0	0	0
359	395	0	0	0	0	0	0	0	0
360	396	0	0	0	0	0	0	0	0
361	397	0,5	1	0	0,5	0,5	0	0	0
362	398	0,5	1	0	0,5	0,5	0	0	0
363	399	0,5	1	0	0,5	0,5	0	0	0
364	400	0,5	1	0	0,5	0,5	0	0	0

Figura 12. Calculo de la distancia entre documentos, a través de la distancia euclídea

Una alternativa para calcular la diferencia entre dos documentos en función de un término específico, más simple aún, podría consistir en tomar el peso de cada término en cada documento y compararlo con el peso del centro correspondiente. Si el valor absoluto de la diferencia entre ambos es menor o igual a un umbral, preestablecido como el promedio de diferencias, ese documento se ubica en el grupo de ese centro. Si bien esta forma proporciona resultados más reales, no subsana el inconveniente de contar en un grupo con documentos que no tengan información de un término determinado; debido a que al calcular la diferencia entre ambos ésta podría ser menor o igual a la del umbral y sin embargo el documento en cuestión no contiene información de ese grupo.

Entonces, en este trabajo se propone calcular la razón geométrica entre el peso del término del documento que se está analizando y el respectivo centro. Luego expresar

dicha razón en porcentaje. Esto proporcionará la idea de qué porcentaje del máximo (centro) está cubriendo el documento analizado, para un término en particular. Aquel documento cuya razón (en porcentaje) sea mayor o igual al umbral se ubica en el grupo.

Esta última forma planteada resultó la más apropiada debido a que cuando se busca información y se calcula la razón entre dos documentos solo interesa la información de un determinado término y no del resto con los que cuenta ese documento. Fundamenta, aún más, esta última propuesta el hecho que puede darse el caso de que dos documentos resulten similares por la cantidad de información que presentan en promedio, pero que para un término en particular sean muy poco o nada similares; con lo cual se puede cometer el posterior error de mostrar un documento que no presente nada o casi nada de información de un término determinado.

La Figura 13 muestra como con la razón geométrica sólo se consideran similares aquellos documentos que realmente contienen información del término en cuestión.

1	Nº de Pág.	Alojamiento	Gastronomía	Actividades	Transportes	Agencias de viajes	Alquiler de vehículos	Productos Regionales	Comercios
271	306	1	0,5	0	0	0,5	0	0	0
272	307	1	0,5	0	0	0,5	0	0	0
273	308	1	0,5	0	0	0,5	0	0	0
274	309	1	1	0,5	0	0,5	0	0	0
275	310	1	0,5	0	0	0,5	0	0	0
276	311	1	0,5	0	0	0,5	0	0	0
277	312	0,666667	1	0	0	0,333333	0	0	0
278	313	0,666667	1	0	0	0,333333	0	0	0
279	314	0,666667	1	0	0	0,333333	0	0	0
280	315	0,5	1	0	0	0,25	0	0	0
281	316	1	0,5	0	0	0,25	0	0	0
282	317	1	0,5	0,5	0	0,5	0	0	0
283	318	1	0,5	1	0	0,5	0	0	0
284	319	1	0,5	0	0	0,5	0	0	0
285	320	1	0,5	0	0	0,5	0	0	0
286	321	1	0,363636	0	0	0,0909091	0	0	0
287	322	1	0,166667	0	0	0,166667	0	0	0
288	323	1	0,0666667	0	0	0,0666667	0	0	0
289	324	1	0,166667	0	0	0,166667	0	0	0
290	325	1	0,0625	0	0	0,0625	0	0	0
291	326	1	0,166667	0	0	0,166667	0	0	0
292	327	1	0,0625	0	0	0,0625	0	0	0
293	328	1	0,166667	0	0	0,166667	0	0	0
294	329	1	0,0555556	0	0	0,0555556	0	0	0

Figura 13. Calculo de la similitud entre documentos, a través de la razón geométrica

Finalmente, en la última etapa, brindar una respuesta ante la consulta del usuario, por ejemplo: "Hoteles y Restaurantes en San Juan" el sistema selecciona los documentos que se encuentran en el grupo Alojamiento, en el grupo Gastronomía y en la intersección de ambos; exhibiendo un listado de direcciones de los documentos electos. Previamente a la entrega, el sistema examina la disponibilidad de los documentos en la Web. De esta forma el sistema no sólo asegura que éstos sean relevantes para el usuario, sino que además se pueda acceder a los mismos. La Figura 14 muestra un primer prototipo de la interfaz de sistema de recuperación para el dominio del turismo.

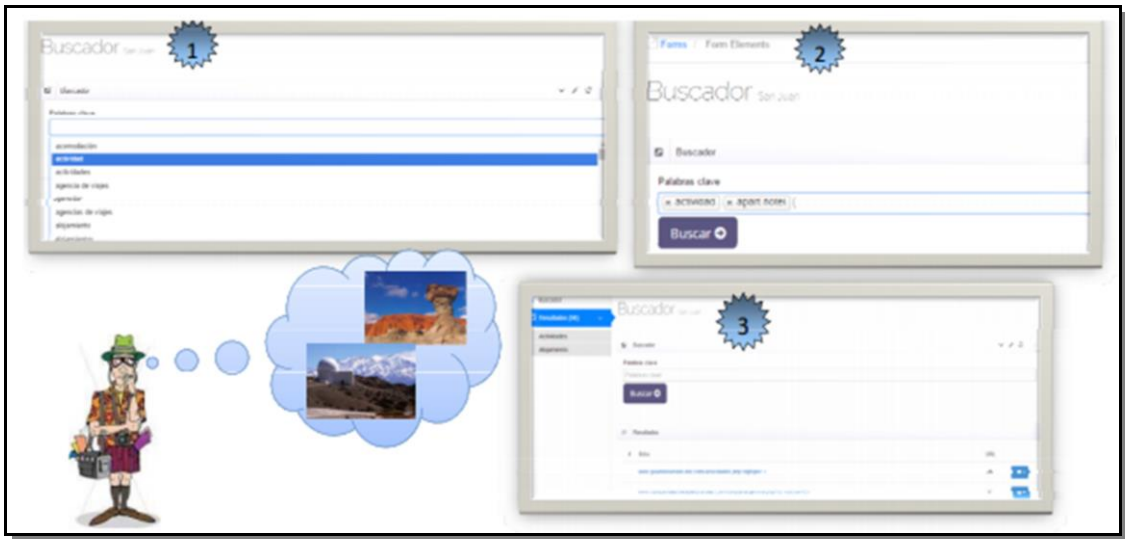


Figura 14. Interfaz del prototipo del sistema de recuperación de información turística

5. 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, desde la selección de documentos relevantes hasta la presentación del listado de URLs.

En cuanto a las fuentes de información web, se puede mencionar que presentan varias dificultades, en cuanto a heterogeneidad de contenido, ya que se encuentra diversidad de información; falta de estructura, en cuanto a que las fuentes no presentan un único formato; disponibilidad, porque en un instante de tiempo pueden estar disponibles y en otro no; distribución, en el sentido de que se encuentran dispersas por la web; cantidad, por el gran número de fuentes a las que se puede acceder; y calidad, en cuanto a que la información en la web no siempre es confiable. La propuesta presentada en este trabajo puede subsanar algunos de estos problemas. Así por ejemplo, al realizar un pre filtrado de los documentos web se logra homogeneidad de contenidos; debido a que se especifica un dominio de aplicación. En cuanto a la disponibilidad de los documentos, la aplicación deberá comprobar la disponibilidad y vigencia de los mismos antes de ser expuestos al usuario web. Además, se asegura una cantidad y calidad razonable de documentos. Lo que la propuesta no puede asegurar es la solución a la pérdida de información, debido a que existe gran cantidad de información a la cual no se puede acceder ya que los buscadores no tienen acceso a ésta por estar ocultas.

La metodología propuesta realiza doblemente la tarea de recuperación de la información. En una primera etapa se debe recuperar información desde la web usando el URL y el texto anclado de cada documento, seleccionando aquellos documentos que surjan del análisis de su contenido y de la cobertura de los mismos respecto a una ontología preestablecida del dominio. Durante la etapa de consulta del usuario, también

se debe recuperar información desde los agrupamientos de los documentos, teniendo en cuenta la similitud de las palabras de la consulta del usuario con los grupos establecidos. Debido a que en ambas etapas se propone usar herramientas externas preexistentes se aprovechan las ventajas semánticas que se derivan de éstas. Es decir que la metodología recupera información teniendo en cuenta la semántica no sólo en el momento de realizar la búsqueda en la web sino además al responder la consulta del usuario.

Así mismo, la metodología plantea usar la idea de clasificación y clustering. Clasificación debido a que de ante mano y con ayuda del experto del dominio se establecen cuáles serán los grupos. Clustering particional en cuanto a la simplicidad de elección de los centros y el cálculo de similitud entre documentos y clustering soft en cuanto a que se requiere que cada documento se ubique en un cluster con determinado grado de pertenencia. En cuanto al cálculo del centro y de la similitud, la metodología propone fórmulas que se diferencian de las usadas por la mayoría de los trabajos actuales. Pudo observarse que necesariamente los documentos debían agruparse en función de la similitud que presentan respecto a un determinado término y no teniendo en cuenta toda la información del documento. Esto solucionó el inconveniente de que se ubicaran en un grupo aquellos documentos que tuvieran poca información o que no tuvieran, respecto a un término específico.

La idea de agrupar los documentos por términos relevantes es precisamente porque la propuesta pretende ofrecer asistencia al usuario web en su búsqueda de información para un dominio específico (salud, educación, turismo, etc.).

Al proponer grupos solapados; es decir admitir que un documento pueda encontrarse en la intersección de dos o más grupos, permite escoger rápidamente aquellos documentos que sólo brindan información explícita de un término solicitado por el usuario o seleccionar aquellos documentos que ofrecen información de más de un término ubicando la intersección de los grupos comprometidos.

Por lo tanto, nuestra principal contribución es una metodología que asiste 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. Es decir, que el usuario no debe enfrentarse a una gran cantidad de resultados. Sólo debe realizar la consulta al sistema que le brindará información clasificada sobre subdominios específicos.

Al mismo tiempo, nuestra propuesta no compite contra los buscadores, meta buscadores o índices existentes; sino que la idea es hacer uso de éstos logrando exactitud y a su vez ganancia en tiempo.

En trabajos futuros se espera implementar la metodología a través de un sistema que posteriormente sea validado realizando comparaciones con actuales sistemas de recuperación de información. Además, probar dicha metodología en otros dominios de aplicación, y así poder observar su escalabilidad a otros dominios que presenten otro tipo de características.

7. Referencias

Baeza-Yates, R. and, Ribeiro-Neto, B. (1999). “Modern Information Retrieval”. Essex, UK: Addison-Wesley.

Baykan E., Henzinger M. and Weber I. (2013). “A Comprehensive Study of Techniques for URL-Based Web Page Language Classification”. ACM Transactions on the Web (TWEB) 7(1): 3.

Brin S. and Page L. (1998). “The Anatomy of a Large-Scale Hypertextual Web Search Engine”. Computer Science Department, Stanford University, Stanford, CA 94305. Recuperado de http://infolab.stanford.edu/~backrub/google.html, Julio de 2014.

Cacheda, F. and Viña, 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.

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.

García-Plaza, A. P., Fresno, V., and Martínez, R. (2008). “Web Page Clustering Using a Fuzzy Logic Based Representation and Self-Organizing Maps”. In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society. Pp. 851-854.

Ghosh, S., and Kumar, 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.

Florescu D., Levy A. and Mendelzon A. (1998). Database Techniques for the World Wide Web: A Survey. *ACM SIGMOND Record, 27(3), Sept. 1998.*

Han J. and Kamber M. (2006). “Data Mining: Concepts and Techniques”. Second Edition. Elsevier. ISBN 13: 978-1-55860-901-3 ISBN 10: 1-55860-901-6. Pp. 402-408.

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

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.

Kamath, S. S., Piraviperumal, D., Meena, G., Karkidholi, S., and Kumar, K. (2013). “A Semantic Search Engine for Answering Domain Specific User Queries”. In Communications and Signal Processing (ICCSP), 2013 International Conference on. Pp. 1097-1101. IEEE.

Li, M. and Cao, S. (2014). “A Serie Method of Massive Information Storage, Retrieval and Sharing”. En Mechatronics and Automation, (ICMA, 2014) IEEE International Conference on. Pp. 1171-1175. IEEE.

Liu, B. (2007). “Web Data Mining – Exploring Hyperlinks, Contents and Usage Data”. ISBN-10 3-540-37881-2. Springer-Verlag Berlin Heidelberg. P.189.

Liu, J., Yu, C., Xu, W., and Shi, Y. (2012).”Clustering Web Pages to Facilitate Revisitation on Mobile Devices”. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces. Pp. 249-252.

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.

Mendez Duque N., Chavarros Porras J. C. and Moreno Laverde Ricardo (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.

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.

Patel, D., and Zaveri, M. (2011). “A Review on Web Pages Clustering Techniques”. In Trends in Network and Communications Springer Berlin Heidelberg. Pp. 700-710.

Qu, J., Wei, C., Wang, W., and Liu, F. (2011). “Research on a Retrieval System Based on Semantic Web”. In Internet Computing & Information Services (ICICIS), 2011 International Conference on. Pp. 543-545. IEEE.

Ramage, D., Heymann, P., Manning, C. D., and Garcia Molina, H. (2009). “Clustering the Tagged Web”. In Proceedings of the Second ACM International Conference on Web Search and Data Mining. Pp. 54-63.

Rekik, R., and Kallel, I. (2013). “Fuzz-Web: A Methodology Based on Fuzzy Logic for Assessing Web Sites”. International Journal of Computer Information Systems and Industrial Management Applications, 5. Pp.126-136.

Romagnano, M., Dominguez, P., and Marchetta, M. (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.

Scime, A. (2005). “Web Mining: Applitacitons and Techniques”. State University of New York College at Brockport, USA. , p.2. ISBN 1-59140-414-2, ISBN 1-59140-415-0 (ppb), ISBN 1-59140-416-9 (ebook).

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.

Sote, A. M. and Pande S. R. (2015). “Web Page Clustering Using Self-Organizing Map”. International Journal of Computer Science and Mobile Computing, 4(1). Pp. 78-84.

Torres, M. (2005). “Sociedad de la Información / Sociedad del Conocimiento”. Pp. 1-9. Recuperado de

http://www.ub.edu/prometheus21/articulos/obsciberprome/socinfsoccon.pdf, Mayo de 2015.

Xiao, L., and Hung, E. (2008). “Clustering Web-Search Result Susing Transduction-Based Relevance Model”. In IEEE 1st Pacific Asia Workshop on Web Mining and Web-based application 2008.

Yue, L., Zuo, W., Peng, T., Wang, Y., and Han, X. (2015). “A Fuzzy Document Clustering Approach Based on Domain-Specified Ontology”. Data & Knowledge Engineering, 100 (A). Pp.148-166.