Acceso Preciso a los Recursos Web en un Dominio Específico. Estudio de Caso en el Área Turismo

María Romagnano[1], Martín Marchetta[2]

[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.
maritaroma@iinfo.unsj.edu.ar, mmarchetta@fing.uncu.edu.ar

Abstract. *When looking for information on the Web it can happen that you find what you want at first glance, or on the contrary you spend a good time trying to discern what of everything found is convenient. This contribution proposes a framework to assist the web user in his search for information in the domain of tourism, reducing the difficulty in the exploration, improving the accuracy and consequently the response time. Mainly focuses on the classification and grouping of web documents that come from heterogeneous and distributed environments. As a case study tourism was chosen, because it is considered an area that has experienced great growth, becoming a fundamental source of income for many regions and countries.*

Resumen. *Al buscar información en la Web puede suceder que se encuentre lo deseado en un primer vistazo, o por el contrario se pase un buen rato tratando de discernir qué de todo lo hallado es conveniente. Esta contribución propone un framework para asistir al usuario web en su búsqueda de información, en el dominio del turismo, reduciendo la dificultad en la exploración, mejorando la precisión y consecuentemente el tiempo de respuesta. Principalmente se focaliza en la clasificación y en el agrupamiento de recursos web que provienen de ambientes heterogéneos y distribuidos. Como caso de estudio se eligió al turismo por considerarse un área que viene experimentado gran auge, convirtiéndose en una fuente fundamental de ingresos para muchas regiones y países.*

1. Introducción

Con el afán de obtener información fácil y velozmente se considera a la Web como la primera, o quizás la única, fuente de información. Sin embargo, esta posibilidad que parece ser ventajosa para el usuario web podría convertirse en un problema si se pasan horas o quizás días tratando de obtener respuestas conforme a sus requerimientos. Tal vez ninguna de las respuestas es satisfactoria, o por el contrario no se encuentra respuesta alguna. 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. Es decir, dominar la diversidad del lenguaje natural, idiomas y el lenguaje de consulta de las bases de dato del sistema de recuperación que se está usando. Podría pensarse en la Web como una "desordenada biblioteca" debido a su continuo y acelerado crecimiento. Así, cuando un

usuario busca información en un ambiente distribuido y heterogéneo como lo es esta red, tal vez si no es experto realizando la búsqueda, podría estar 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 para tal fin. Asimismo, probablemente se desconocen cuáles son las características que presentan los sistemas de recuperación o buscadores que se usan actualmente, usando el más popular pero que posiblemente no es el que brinda mejor información. Si bien estos sistemas han evolucionado vertiginosamente, todavía cada una de ellos exhibe distintas formas de realizar la recuperación, lo cual lleva a encontrarse con respuestas más o menos precisas, en menor o mayor cantidad, organizadas o clasificadas de distinta forma o según determinados criterios.

Por lo tanto, este trabajo de investigación se vio motivado por la continua insatisfacción de los usuarios web a la hora de realizar una búsqueda rápida y personalizada de información, principalmente en un dominio como el turismo, que se caracteriza por poseer un cúmulo considerable y heterogéneo de información. Conjuntamente, el spamming es un problema significativo debido a que posiciona a recursos que pueden ser irrelevantes en primeras posiciones, perjudicando la calidad de los resultados de la búsqueda y la experiencia del usuario.

De igual modo, se considera un tema de relevancia social y que posee implicaciones prácticas. Así, en [Ferran Segura, 2011] se realizó un estudio de satisfacción de usuarios al efectuar una búsqueda. En éste se develó que el 40% de los encuestados estaban totalmente de acuerdo con la pregunta: *"…Google les ayuda a encontrar eficientemente la información que necesitan…"* y solo un 20% estaba totalmente de acuerdo con la consigna: *"…dependiendo de lo que estoy buscando, la información proporcionada generalmente es suficiente para satisfacer la necesidad…"*.

El resto del artículo se estructura de la siguiente forma: en la sección 2 se presentan trabajos relacionados. En la sección 3 se desarrolla la metodología seguida para abordar la investigación, se contextualiza brevemente al lector en el framework propuesto, el cual se sustenta en los trabajos [Romagnano y Marcheta, 2017], [Romagnano, Marcheta y Dominguez, 2017] y [Romagnano, Dominguez y Marcheta, 2017]. En la sección 4 se exponen resultados experimentales. En la sección 5 se somete a discusión nuestra contribución, comparándola con algunos trabajos precedentes. Finalmente, en la sección 6 se presentan las conclusiones logradas, luego de las pruebas experimentales y del análisis de resultados alcanzados en la sección 4; con la propuesta en el dominio del turismo.

2. Trabajos Relacionados

En cuanto a la recuperación de información, Li y Cao 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 términos generales, cuando el sistema de recuperación de información distribuida recibe una consulta del usuario se hará la selección de la colección de documentos de acuerdo con la correlación que exista [Li y Cao, 2014]. Solarte y Millán proponen extender el proceso de

recuperación de información usando tecnologías de la web semántica [Solarte y Millán, 2014]. En el trabajo de Evangelopoulos et al. se plantea una métrica para evaluar la recuperación de la información. Establecen que la mayoría de los enfoques tradicionales usan el juicio de expertos como métrica de evaluación. Aunque este enfoque es muy costoso y tedioso cuando se trata con un importante número de datos como lo es la web. Una solución al mismo consiste en usar el análisis del clic de los datos, donde el usuario refleja sus preferencias [Evangelopoulos et al., 2016]. En el trabajo presentado por Dahab, Kamel y Alnofaie se señala que la mayoría de los modelos de recuperación de información representan documentos como una bolsa de palabras que sólo tienen en cuenta la frecuencia de los términos, sin considerar la proximidad de éstos. También se menciona que mantener un índice de proximidades hace que se requiera más espacio. En este trabajo, para mejorar el puntaje de clasificación y mejorar la eficiencia de tiempo de ejecución para resolver la consulta y mantener un nivel razonable de espacio para el índice se presentan dos algoritmos diferentes usando la transformada de wavelet discreta [Dahab, Kamel y Alnofaie, 2017].

Con respecto a la clasificación, Baykan, Henzinger y Weber 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 [Baykan, Henzinger y Weber, 2013]. En el artículo de Liu et al. se propone un nuevo modelo de clasificación llamado Modelo de Gravitación (GM) para resolver el problema de clasificación cuando se tienen clases desbalanceadas [Liu et al., 2017].

Sobre agrupamiento, Sote y Pandey realizan 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 [Sote y Pandey, 2015]. Matsumoto y Hung, y Xiao y Hung, además de permitir el solapamiento entre clusters, proponen la idea de agrupar los resultados de búsqueda web ya generados por motores de búsqueda convencionales [Matsumoto y Hung, 2010], [Xiao y Hung, 2008] respectivamente. Kamjou y Ahmadzadeh, y Xing y Ha proponen un algoritmo Fuzzy C-Means 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 mejores resultados pero el tiempo de ejecución se incrementa considerablemente respecto al Fuzzy C-Means original [Kamjou y Ahmadzadeh, 2015], [Xing y Ha, 2014]. Ghosh y Kumar realizan una comparación entre los algoritmos K-Means y Fuzzy C-Means concluyendo que el primero es mejor en cuanto a que el segundo requiere más tiempo de cálculo debido a las métricas difusas necesarias para calcular las particiones [Ghosh y Kumar, 2013]. Hernández, Rivero, Ruiz y Corchuelo proponen un clasificador automático de links, realizando un rápido rastreo. Se usa una técnica de aprendizaje no supervisado, ya que el clasificador aprende de un conjunto de entrenamiento de páginas centrales no clasificadas que son recogidas automáticamente del sitio web, basado exclusivamente en las características de la URL y así no se requiere descargar una página para clasificarla [Hernández, Rivero, Ruiz, y Corchuelo, 2016]. Yan, Zhang, Ma y Yang proponen un método de clustering que factoriza conceptos imponiendo restricciones a los objetos que pertenecen a la misma clase y a clases diferentes, y así logran una capacidad más discriminatoria [Yan, Zhang, Ma, y Yang, 2017]. Por su parte, Yue, Zuo, Peng, Wang y Han, además de una ontología y

HowNet usan el modelo SVD para clusterizar [Yue, Zuo, Peng, Wang y Han, 2015]. Rekik y Kallel usan lógica difusa para evaluar sitios web [Rekik ya Kallel, 2013]. Romagnano, Aciar y Marchetta plantean un método que recupera y agrupa fuentes de información web de acuerdo a los servicios que ofrecen, reduciendo la complejidad de búsqueda [Romagnano, Aciar y Marchetta, 2015]. También, Romagnano, Domínguez y Marchetta proponen un agente de filtrado para localizar fuentes de información en la web, las agrupa y luego brinda al usuario información precisa, acorde a sus necesidades. Para determinar la relevancia de una página usan lógica difusa. Para agrupar las páginas similares proponen un algoritmo basado en los métodos K-means y Fuzzy C-means [Romagnano, Domínguez, Marchetta y Aciar, 2015].

No obstante la existencia de varios trabajos de investigación, todavía quedan algunos aspectos abiertos a la discusión, como la precisión y el tiempo de respuesta al usuario. Considerando, además, la complejidad del dominio en cuestión debido a que algunos de estos presentan problemas para recuperar información en cuanto a calidad, cantidad, heterogeneidad, distribución, entre otros.

Se pone de manifiesto la necesidad de clasificar y agrupar recursos web de acuerdo a un criterio preestablecido. Por lo tanto, se presenta un framework para asistir al usuario web, en el cual implícitamente se propone un método, para recuperar información web, focalizándose en la clasificación y en el agrupamiento solapado. Consecuentemente, con esta contribución se aborda el problema de precisión en la recuperación de información web para el dominio del turismo, ampliando la labor de investigación a través de evaluaciones experimentales y del análisis de resultados de la contribución divulgada por los autores en [Romagnano y Marchetta, 2017].

3. Desarrollo

3.1. Metodología

El tipo de estudio que se llevó a cabo fue exploratorio-descriptivo, lo cual permitió describir qué técnicas se usaron y cómo fueron usadas. Además, si las mismas brindaban solución o no a la problemática expuesta.

En cuanto a la estrategia metodológica se decidió que sería mixta (cualitativa-cuantitativa), predominando la característica cualitativa. Es decir que se planteó un supuesto y luego a través de datos empíricos se fueron creando descripciones e interpretaciones de este supuesto.

En cuanto al contexto de la investigación se decidió que se realizarían estudios observacionales y experimentales. Es decir que en algunas instancias se realizaron observaciones de la unidad de análisis, en otras se realizaron experimentos y en otras ambas actividades.

Relacionado con la población, en este estudio, la misma estuvo formada por el conjunto de recurso web del dominio del turismo para la provincia San Juan, Argentina y las muestras representativas de esta población estuvieron formadas por aquellos recursos web seleccionados como relevantes, en el desarrollo experimental y proporcionados por personal a cargo de la Secretaría de Turismo.

Conjuntamente para probar la satisfacción de los usuarios web que requerían información turística de San Juan se realizaron pruebas con 25 participantes voluntarios, que en enero de 2018, actuando en el rol de usuarios web, realizaron búsquedas a través de un prototipo del sistema de recuperación propuesto y a través de un buscador popular.

Para la recolección de datos se utilizaron como instrumentos: encuestas a usuarios, entrevistas abiertas y sesiones grupales con expertos del dominio, revisión de archivos y/o documentos, observaciones no estructuradas de distintos recursos web y evaluación de experiencias personales.

3.2. Framework

El framework propuesto sistematiza las tareas del sistema de recuperación de información web que lo materializa. Se localizan recursos en la web, se agrupan de acuerdo a un criterio preestablecido y luego se brinda al usuario información precisa. En la Figura 1 se presentan tres gestiones principales: recuperación de recursos web (Recuperación), procesamiento y agrupamiento de estos (*Agrupamiento*). Por último se resuelve la consulta del usuario (*Vista Unificada*).

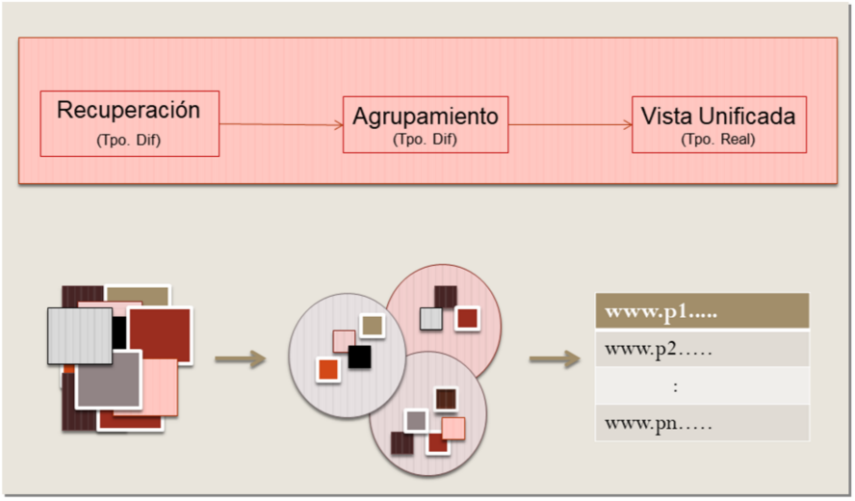


Figura 1. Acciones principales del framework [Romagnano, Marchetta y Domínguez, 2017]

Consecuentemente, los aportes del presente trabajo muestran la mejora que se logra al acceder precisamente a los recursos web y en menor tiempo, gracias a los métodos de clasificación y agrupamiento implementados en las fases del framework. Es decir, que el usuario no debe enfrentarse a una gran cantidad de resultados. Sólo debe realizar la consulta al sistema de recuperación, que brindará información clasificada y agrupada sobre subdominios específicos.

La Figura 2 muestra el diseño conceptual de este sistema de recuperación, donde D_i representa una generalización de la clase dominio, la cual se especificará con alguno

de los dominios preestablecidos. La Figura 3 muestra la arquitectura del sistema de recuperación de información, que resuelve las consultas siguiendo el framework propuesto. La consulta del usuario puede ser mediante palabras claves, usando operadores booleanos, a través de frases o usando la forma más simple para él; el lenguaje natural. En el módulo de la interfaz, el procesamiento de la consulta elimina stop words, 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 recursos 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 recursos indexados y almacenados para permitir la posterior recuperación. El módulo de rankeo calcula la relevancia de cada recurso 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.

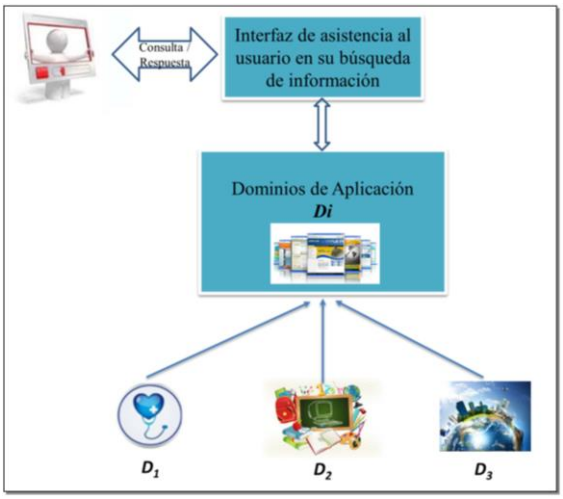


Figura 2. Diseño conceptual del sistema de recuperación

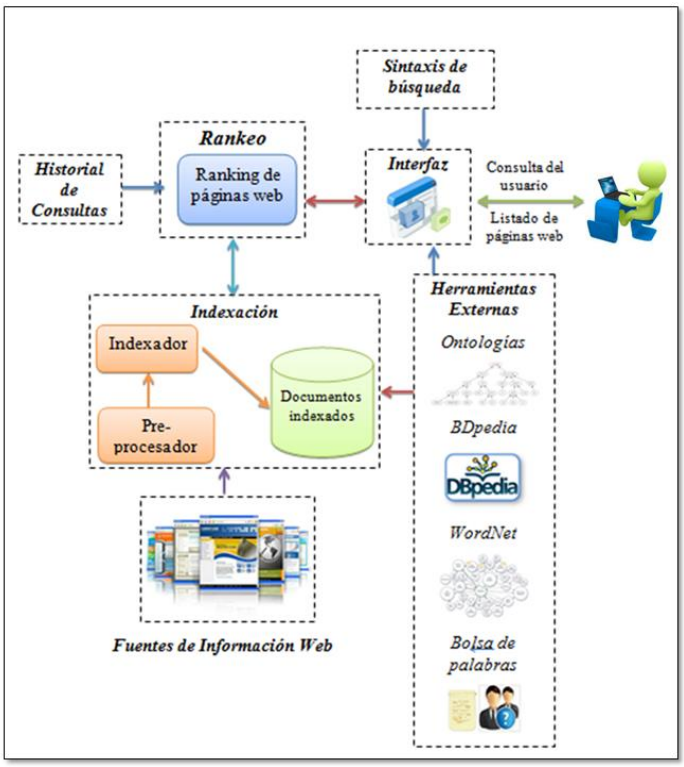


Figura 3. Arquitectura del sistema de recuperación propuesto [Romagnano, Marchetta y Domínguez, 2017]

3.3. Fases del Framework

El framework presenta seis fases, abarcando desde la recuperación, procesamiento e indexado de los recursos web hasta resolver la consulta del usuario.

En la *primera fase* 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 metabuscadores, buscadores generales e índices temáticos, se realiza la búsqueda de la información con ciertas palabras claves.

En la *segunda fase* se realiza un análisis preliminar de los resultados obtenidos en la fase anterior, seleccionando aquellos recursos que sean relevantes al dominio en cuestión. Para esto ser aplican técnica de aprendizaje supervisado, donde un experto provee ejemplos de recursos relevantes y no relevantes del dominio. Además, se usa una ontología preexistente del dominio de aplicación para determinar sinónimos y/o relaciones entre términos y de esta forma poder contemplar la semántica en este paso.

En la *tercera fase* se determinan cuáles serán los subdominios relevantes del dominio preestablecido, los que posteriormente serán nombres de los agrupamientos. Para seleccionar dichos términos se emplean la bolsa de términos (producto de la consulta a un experto del dominio), BabelNet, WordNet y la ontología del dominio, bosquejadas en la Figura 3 como herramientas externas. Manualmente, el experto del dominio es quién determina qué subdominios (bolsa de 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 y/o diferencias semánticas (sinonimia o antonimia, polisemia u homonimia, hiperonimia o hiponimia, holonimia o meronimia) entre cada término de la bolsa de términos 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 establecidos como nombres de los futuros grupos. Este requerimiento asegura una mejor cobertura de la semántica.

En la *cuarta fase* la tarea es determinar el peso que tiene cada subdominio en cada recurso, por lo tanto el procedimiento a desarrollar consiste en analizar el contenido de cada recurso almacenado en esta base y a través de minería web remover las stop words y realizar stemming para luego poder determinar la frecuencia de aparición de cada término relevante y sus variantes, identificados en la fase 3. Nuevamente, se usan las herramientas externas propuestas en la Figura 3 para establecer la correlación semántica entre los términos de las diversas fuentes de información web y los subdominios definidos como relevantes en la fase anterior. Por último, se usa el esquema TF para calcular el peso de los subdominios relevantes en cada recurso.

En la *quinta fase* el framework propone agrupar los recursos. En algunas situaciones, donde los recursos pueden presentar información de varios subdominios relevantes, se debe considerar la posibilidad de solapamiento. Por lo tanto, necesariamente, se debe permitir que un recurso pueda corresponder a uno o a más grupos con un cierto grado de pertenencia.

En la *sexta fase*, ante la consulta del usuario, el sistema entrega un conjunto de recursos web. Se proporciona un modelo de consulta que no sólo aprovecha las ventajas del lenguaje natural para que los usuarios hagan sus consultas, sino que además agrega semántica en cuanto a que se tiene en cuenta el peso de cada término de la consulta en el recurso, cantidad de términos relevantes, densidad de información del recurso, aparición del recurso en el historial de búsqueda, etc; logrando sortear algunas de desventajas que presentan algunos sistemas de recuperación actuales. Así, los sistemas de recuperación de información basados sólo en palabras claves si bien son fáciles de usar, se limitan sólo a considerar términos específicos, con lo cual se pierde información. Otro caso es el de los sistemas de recuperación basados en formularios. Si bien presentan gráficamente al usuario partes de la estructura para que seleccione las clases con las cuales se realiza la búsqueda, son poco flexibles ya que sólo se pueden seleccionar los elementos que se muestran en el formulario. En cuanto a los sistemas basados en lenguaje formal presentan un nivel de complejidad superior para el usuario.

4. Experimentación y Resultados

Teniendo en cuenta la factibilidad de usar la propuesta tanto en el ámbito público como en el privado, se dispuso elegir una base de datos de recursos web del sector turismo y tomar como referencia los trabajos [Romagnano, Marchetta y Domínguez, 2017] y [Romagnano, Domínguez y Marchetta, 2017]. Esta decisión se tomó considerando que este dominio es un área que viene experimentado gran auge, convirtiéndose en una fuente fundamental de ingresos para muchas regiones y países. Dada la diversidad y

cantidad de información relacionada con este mercado, la dispersión y la desregulación del mismo, se dificulta cada vez más la labor de los agentes o portales de servicios turísticos para ofrecer al usuario web información completa y actualizada.

Para llevar a cabo la primera fase se realizaron pruebas integrando al prototipo del sistema de recuperación las APIs del buscador Google. En este caso se decidió usar la funcionalidad de rastreo de la web provista por el buscador debido a que dicha funcionalidad se encuentra ampliamente desarrollada y probada por otras aplicaciones que la han usado. Como provincia se eligió a San Juan de la República Argentina debido a que históricamente no presenta antecedentes en el mercado del turismo y actualmente políticas gubernamentales conjuntamente, y a través de la Universidad de San Juan, están apostando a este sector productivo [Carrizo, 2018].

Por recomendación del experto en turismo, las palabras claves usadas fueron TURISMO, SAN JUAN y ARGENTINA. Además, se propuso que debido a la dinamicidad de la web, la recuperación automática debía realizarse una vez a la semana.

La Figura 4 muestra una captura de pantalla donde el prototipo del sistema de recuperación va obteniendo sucesivamente el conjunto de URLs al invocar al buscador.

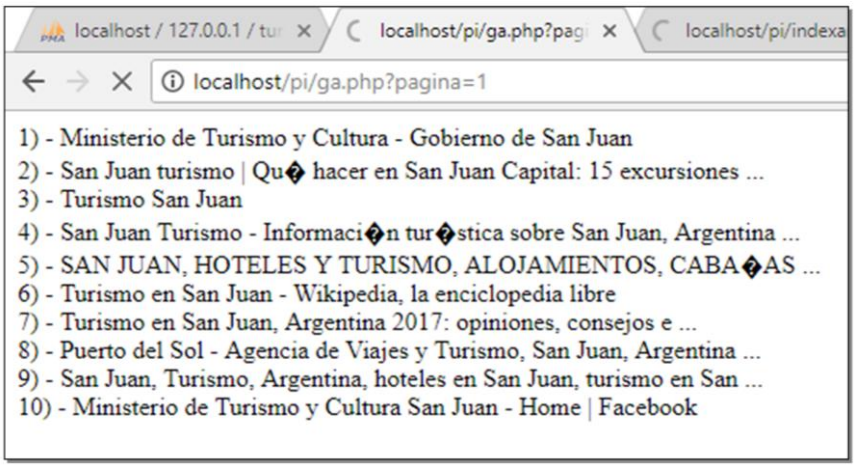


Figura 4. Imagen del prototipo del sistema de recuperación de información usando la API para invocar a Google y poder realizar la búsqueda

Para llevar a cabo la segunda fase se desarrolló una ontología del turismo, usando el software libre Protégé. Se tuvieron en cuenta las especificaciones dadas por el experto del dominio, se partió desde un modelo conceptual, es decir un modelo de clases (Figura 5).

Posteriormente se establecieron dos clases disjuntas y se aplicaron técnicas de aprendizaje supervisado, donde un experto en el turismo proporcionó ejemplos de recursos relevantes y no relevantes del dominio, para aprender qué términos eran discriminantes. Igualmente, se usó la ontología desarrollada para contemplar la semántica en este paso. Por otra parte, los expertos proporcionaron una base de 157 recursos para poder realizar las pruebas experimentales. Dichas pruebas fueron realizadas con Weka 3.9 (ya que se trata de una herramienta de uso libre y que ofrece

varios algoritmos de clasificación), bajo Windows 10 y procesador Core i7. En el siguiente paso se construyó un clasificador. Para analizar qué clasificador era más conveniente, en función de los datos, se probaron varias técnicas de clasificación. En el Cuadro 1 puede observarse que la técnica *Logistic*, en promedio, obtuvo mejores resultados en cuanto a porcentaje de correctamente clasificados, precisión, recall, verdaderos positivos; por ejemplo.

Para llevar a cabo la tercera fase se realizaron reuniones de expertos del turismo donde se propusieron términos relevantes, es decir la bolsa de palabras que contenía términos candidatos a ser seleccionados. Luego se realizaron las comparaciones semánticas entre cada uno de los términos de esta bolsa de palabras y las herramientas externas y así se determinaron los términos relevantes, los cuales serían los nombres de los futuros grupos. Para realizar pruebas en la cuarta fase, en primera instancia se eliminaron stop words teniendo en cuenta el listado proporcionado por el Proyecto Snowball [Snowball, 2017]. Luego se continuó con la tarea de stemming. Morfológicamente las palabras están estructuradas en prefijos, sufijos y la raíz. La técnica de stemming lo que pretende es eliminar las posibles confusiones semánticas que se puedan dar en la búsqueda de un concepto, para ello trunca la palabra y busca solo por la raíz [Porter, 2017].

Existen varios algoritmos y librerías para descartar prefijos y sufijos. Así por ejemplo entre los algoritmos se pueden mencionar: Porter, Paice/Husk, Stemmer-es, Snowball, PECL, entre otros. En este caso, para realizar la experimentación se consideraron el algoritmo Stemmer-es y la librería PECL; debido a que están programados en el mismo lenguaje del prototipo del sistema de recuperación. En el siguiente paso se determinó la frecuencia de aparición de cada término relevante y sus variantes en cada recurso (Figura 6), se calculó su peso y se le asignó uno o varios n-gramas. Estos n-gramas se formaron a partir de tres a cinco términos claves de turismo, proporcionados por los expertos en las sesiones grupales. Posteriormente con cada uno de estos grupos se hicieron comparaciones semánticas con BabelNet logrando obtener n-gramas más específicos.

Para continuar con la experimentación, siguiendo con la quinta fase, se tuvieron en cuenta las propuestas de algunos autores. Según Liu, generalmente, se calcula la similitud entre documentos a través de la similitud del coseno; más bien que la distancia entre los vectores que representan esos documentos [Li y Han, 2013]. Sin embargo algunos autores han encontrado algunas falencias al aplicar esta métrica. Liu propone extender la medida de similitud del coseno, considerando la distancia Mahalanobis [Liu, 2007]. Por su parte, Mikawa, Ishida y Goto plantean que la similitud del coseno tiende a sesgarse a aquellos términos con alta frecuencia y no se analiza cómo es el comportamiento cuando entre dos vectores existen términos cuyos valores son cero [Mikawa, Ishida, y Goto, 2011].

Así como lo hicieron algunos de los trabajos mencionados en la sección de introducción, en esta propuesta se realizaron pruebas con las tradicionales fórmulas del Coseno y Distancia Euclídea. En el caso de tener que establecer la similitud en cuanto a cantidad de información de cada término que ofrecen los recursos web, cualquiera de los métodos planteados para calcular similitud o distancia presentaron problemas al ser aplicados. Es decir, proveían una similitud (o distancia) general entre dos vectores, por

lo tanto cuando se necesitaba puntualizar la similitud (o distancia) para un término específico entre dos recursos no se obtuvieron verdaderos resultados. Este trabajo propone la siguiente fórmula para calcular la similitud, la cual es aún una mejora de la fórmula planteada en [Romagnano y Marchetta, 2017]:

$$S_{d_j r_i} = \frac{w_{ij}}{r_i} * 100 \qquad (1)$$

Dónde:

Sd_{jri}: similitud entre el recurso r_j y el representante r_i

w_{ij}: peso del término t_i en el recurso r_j, obtenido como la normalización del número de veces que el subdominio t_i aparece en el documento.

r_i: peso del representante r_i, obtenido como el máximo valor del término t_i

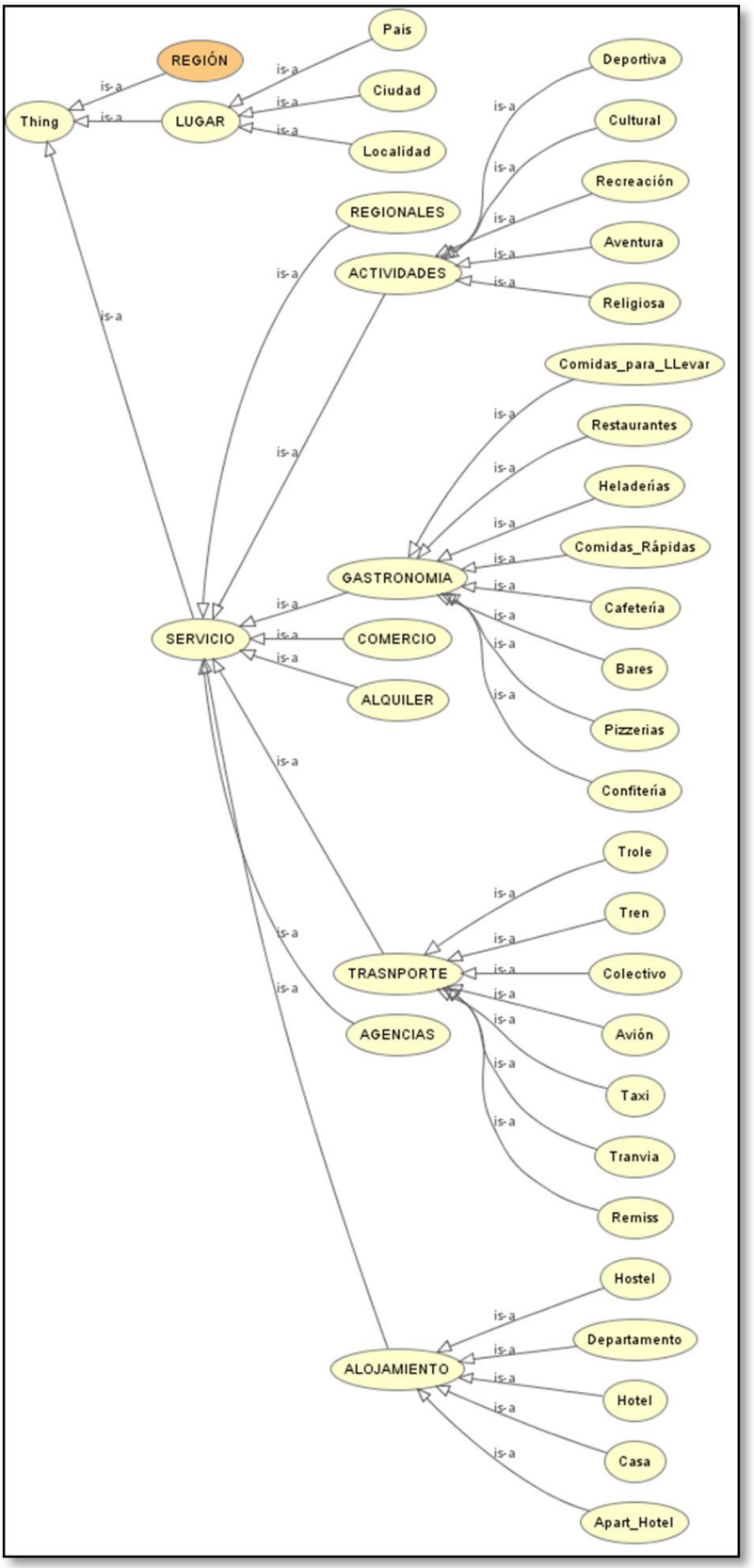


Figura 5. Diagrama de clases de la ontología creada

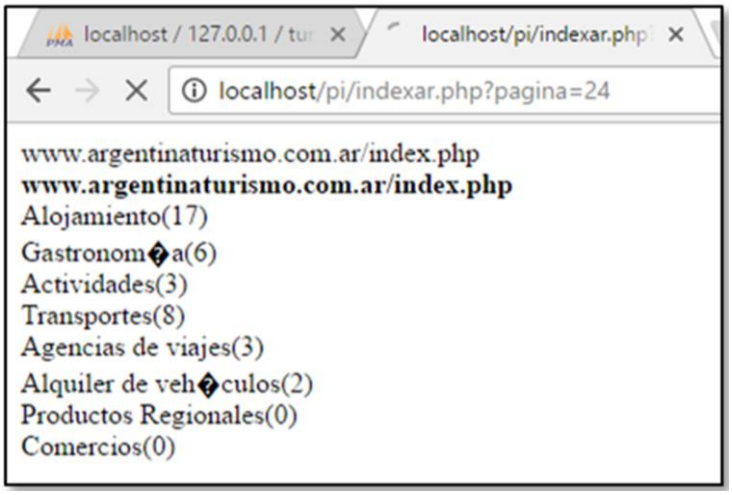


Figura 6. Frecuencia de cada término relevante en la página www.argentinaturismo.com.ar

Cuadro 1. Comparación de técnicas de clasificación con sus respectivos algoritmos, en Weka 3.9

	REGLAS	**ARBOL**	**BAYES**	**FUNCIONES**	**METAS**
	PART	LMT	BayesNet	Logistic	CostSensitiveClassifier
Tpo	11,00	0,47	0,00	0,08	0,00
% CC*	84,71	89,81	85,35	92,36	84,71
% IC*	15,30	10,19	14,64	7,64	15,28
TPR	0,97	0,98	1,00	0,98	0,97
FPR	0,50	0,33	0,50	0,23	0,50
Pre	0,84	0,90	0,84	0,92	0,84
Re	0,97	0,98	1,00	0,98	0,97

Referencias:
* % CC= % de clasificados correctamente, % CI*= % de clasificados incorrectamente, FPR= tasa de falso positivo, Tpo= tiempo para construir el modelo (en segundos), TPR= tasa de verdadero positivo, Pre= precisión, Re= recall.

En el Cuadro 2 se observa que el método de agrupamiento propuesto logra mejores resultados, fundamentalmente en cuanto a precisión que es el criterio de evaluación que más incumbe en este caso de estudio, debido a que en este caso interesan mayor cantidad de recursos relevantes que recuperados.

Cuadro 2. Comparación de métodos según las métricas

Método Métrica	Coseno	Distancia Euclídea	Propuesta
Precisión	0,591	0,604	0,959
Recall	0,580	0,848	0,944
F-Score	0,585	0,705	0,951

Para evaluar la última fase y poder medir la satisfacción de los usuarios se realizó una experimentación con un grupo de 25 personas interesadas en buscar información relacionada con el turismo, considerando lo establecido en [Hernández, Fernández y Baptista, 2010]. Se les pidió que abrieran una pestaña del navegador y a través del prototipo del sistema realizaran al menos cinco consultas diferentes relacionadas con el dominio del turismo. Luego se les pidió que en otra pestaña hicieran las mismas búsquedas en un buscador popular y de su agradado. Por último se les pidió que analizaran los resultados devueltos por ambas aplicaciones y que contestaran una encuesta de satisfacción. Las Figuras 7, 8 y 9 muestran las acciones realizadas por una persona que formaba parte del experimento; elegida al azar.

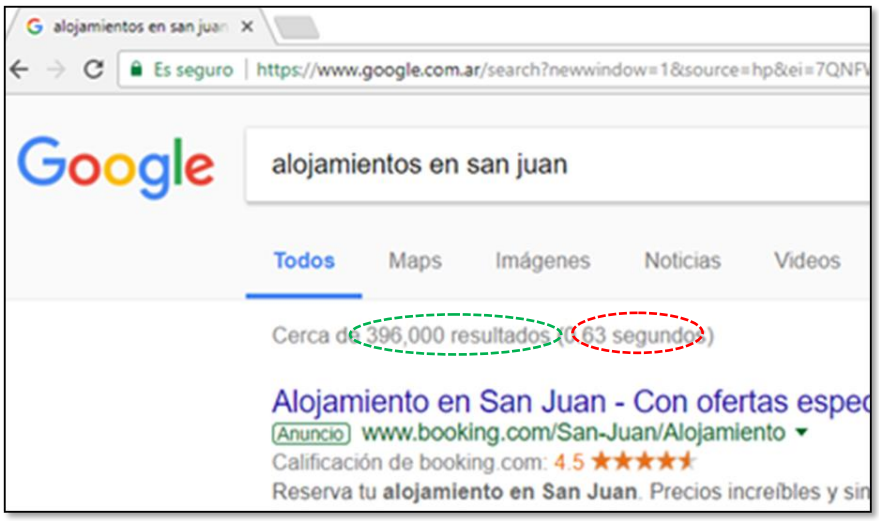


Figura 7. Pestaña del navegador

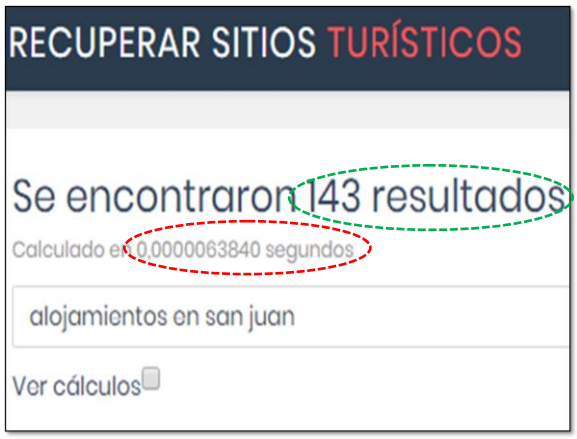
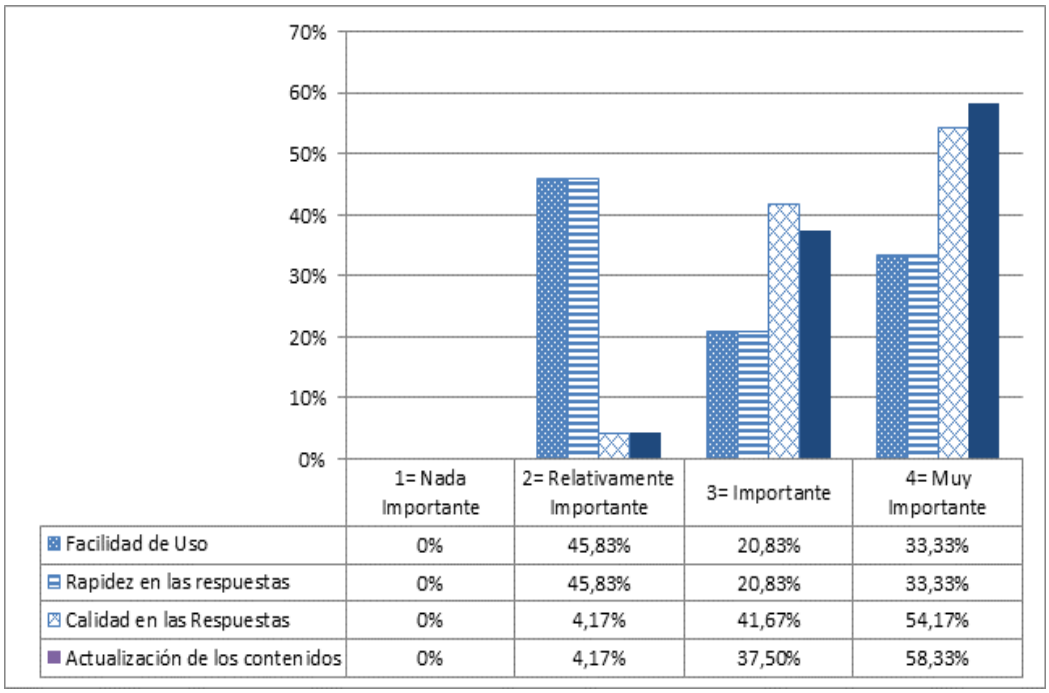


Figura 8. Interfaz del sistema de recuperación de información turística

Figura 9. Encuesta de uso y satisfacción (https://www.e-encuesta.com/r/YR8SMWEi_QYUT0_ZSNz8-g/)

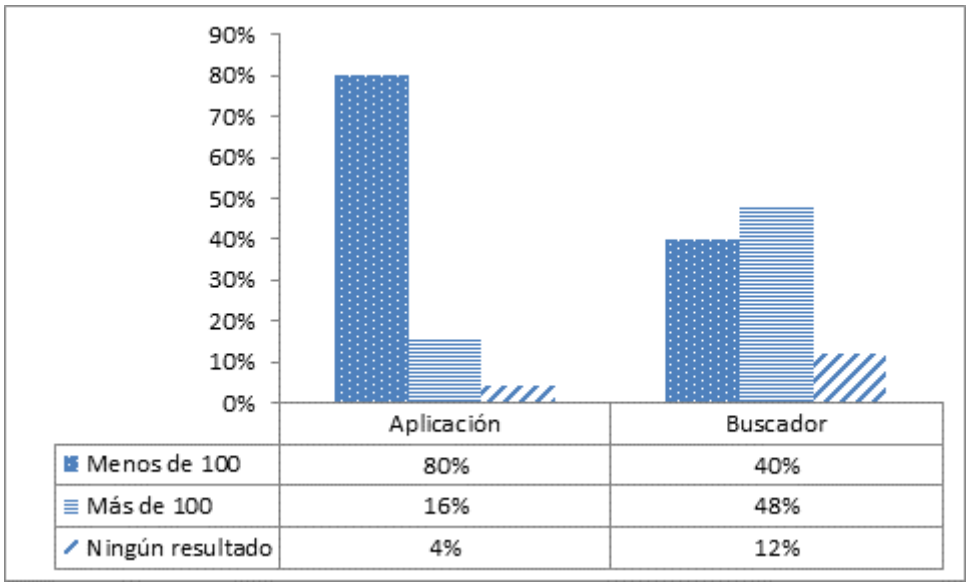
Por otra parte, el análisis estadístico de las encuestas demostró que el 45,83% de los participantes consideró relativamente importante los aspectos: facilidad de uso y rapidez en las respuestas, al realizar la búsqueda en la Web (Figura 10).



	1= Nada Importante	2= Relativamente Importante	3= Importante	4= Muy Importante
Facilidad de Uso	0%	45,83%	20,83%	33,33%
Rapidez en las respuestas	0%	45,83%	20,83%	33,33%
Calidad en las Respuestas	0%	4,17%	41,67%	54,17%
Actualización de los contenidos	0%	4,17%	37,50%	58,33%

Figura 10. Evaluación del grado de importancia de la facilidad de uso y rapidez de las respuestas

El 80% de los participantes contestó que obtuvo menos de 100 resultados al realizar la búsqueda con el sistema de recuperación de información web o aplicación; denominación que se le ha dado en la encuesta (Figura 11).



	Aplicación	Buscador
Menos de 100	80%	40%
Más de 100	16%	48%
Ningún resultado	4%	12%

Figura 11. Evaluación de la aplicación y del buscador en cuanto a la cantidad de resultados obtenidos

El 82,61% prefirió usar este sistema de recuperación para realizar otra búsqueda en un dominio concreto (Figura 12).

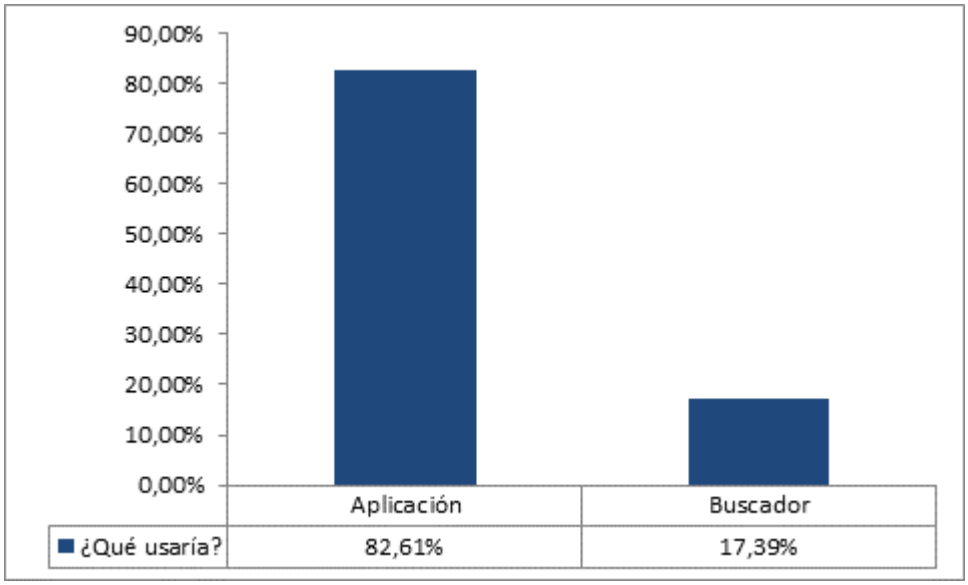


Figura 12. Evaluación de la aplicación y del buscador en cuanto a qué usaría para volver a realizar una búsqueda en un dominio de aplicación concreto

En cuanto al nivel de satisfacción al realizar la búsqueda en el sistema de recuperación propuesto, el 60% de los participantes estuvo satisfecho (Figura 13).

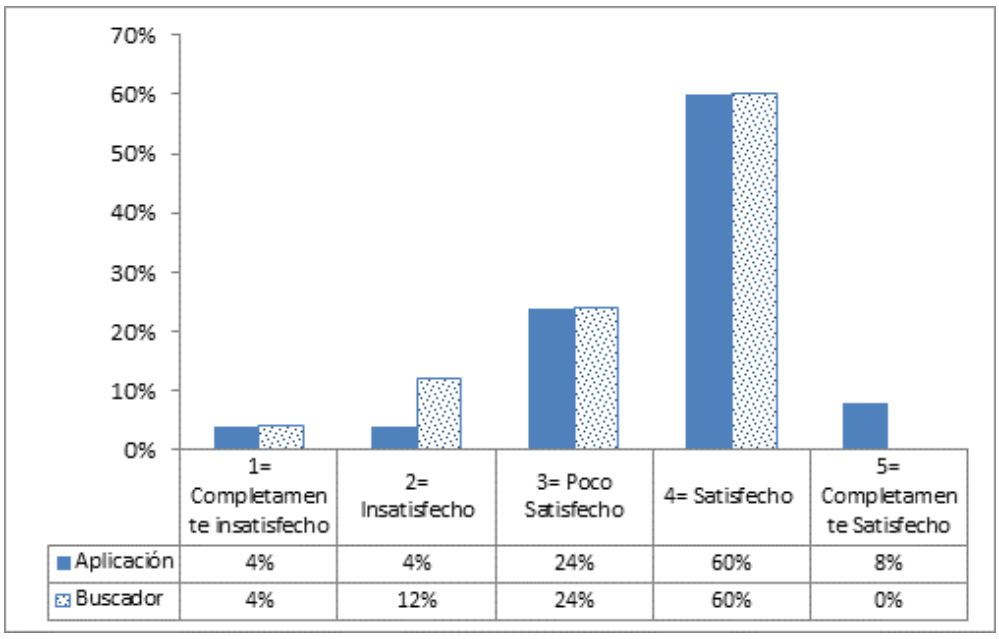
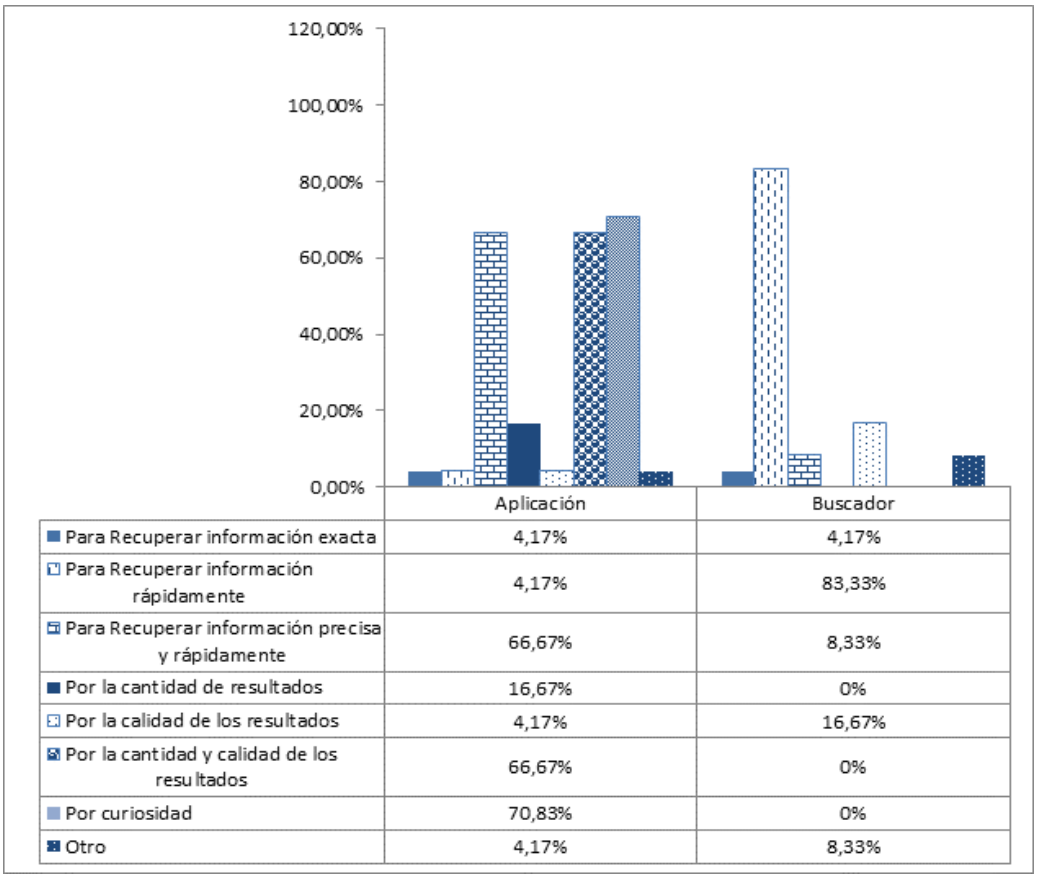


Figura 13. Evaluación del nivel de satisfacción al realizar una búsqueda en la aplicación y en el buscador

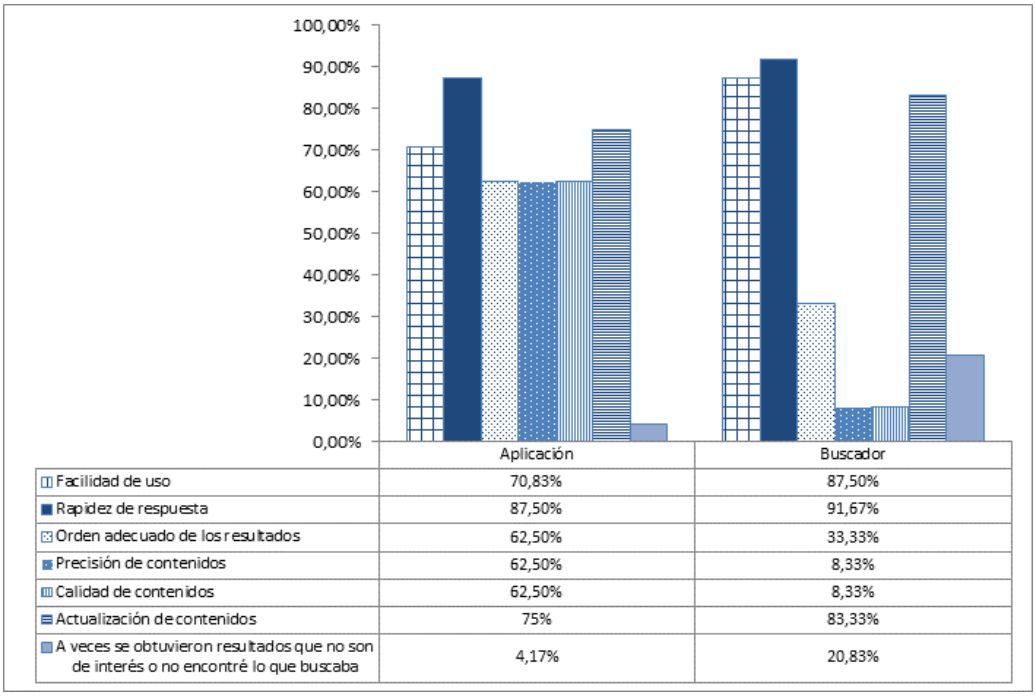
El 66,67% de los participantes contestó que usaría el sistema propuesto para recuperar información precisa y rápidamente. Además el mismo porcentaje de participantes respondió que lo usaría por la calidad y cantidad de resultados obtenidos (Figura 14).



	Aplicación	Buscador
Para Recuperar información exacta	4,17%	4,17%
Para Recuperar información rápidamente	4,17%	83,33%
Para Recuperar información precisa y rápidamente	66,67%	8,33%
Por la cantidad de resultados	16,67%	0%
Por la calidad de los resultados	4,17%	16,67%
Por la cantidad y calidad de los resultados	66,67%	0%
Por curiosidad	70,83%	0%
Otro	4,17%	8,33%

Figura 14. Evaluación de la aplicación y del buscador en cuanto a la finalidad con la cual lo usarían para realizar una búsqueda

Por último, en cuanto al ítem experiencia de búsqueda, el 62,50% cree que el sistema cumplió con el orden, calidad y precisión de contenidos (Figura 15).



	Aplicación	Buscador
Facilidad de uso	70,83%	87,50%
Rapidez de respuesta	87,50%	91,67%
Orden adecuado de los resultados	62,50%	33,33%
Precisión de contenidos	62,50%	8,33%
Calidad de contenidos	62,50%	8,33%
Actualización de contenidos	75%	83,33%
A veces se obtuvieron resultados que no son de interés o no encontré lo que buscaba	4,17%	20,83%

Figura 15. Evaluación de la aplicación y del buscador en función de los criterios orden, calidad y precisión de los contenidos

5. Discusión

Para someter a discusión esta contribución se realiza una comparación con algunos trabajos mencionados anteriormente; seleccionados en cuanto su relación y actualidad.

En [Hernández, Rivero, Ruiz y Corchuelo, 2016] se clasifican las páginas sin analizar su contenido, lo cual lo hace ventajoso en cuanto a tiempo y almacenamiento. Con esta propuesta, al realizar una doble selección de relevantes (fase 2 y 6), se asegura más la pertinencia de los resultados.

En [Yue, Zuo, Peng, Wang y Han, 2015], al igual que esta contribución, se construye una ontología del dominio y luego se analizan documentos web basándose en esta ontología y en HowNet. Se calcula la similitud usando un procedimiento más complejo que el propuesto en este trabajo. Se realiza el depurado de los stopword, pero no el proceso de stemming. El proceso de clustring es algo más lento y complicado que el que se propone en este artículo. Otras limitaciones observadas son: dependencia de la comprensión de la ontología usada, la cual además es muy básica y escaso procesamiento del lenguaje natural para identificar características. En este aporte también se comparte la limitación en cuanto a dependencia de la ontología, no así el problema del lenguaje natural, lo cual es contemplado en la consulta del usuario.

En [Solarte y Millán, 2014] propone un sistema de recuperación de información web donde las palabras claves expresadas en la consulta del usuario son extendidas por medio de anotaciones que se obtienen a partir de una ontología del dominio. En este trabajo también se tienen en cuenta las palabras claves y se extienden por medio de sus

sinónimos obtenidos de la ontología, pero además se usan n-gramas, BabelNet y WordNet para ampliar semánticamente. Se realiza stemming y se eliminan los stopword. Sin embargo el sistema que se propone en este trabajo es más completo respecto del trabajo de Solarte y Millán, debido a que no sólo se expone cómo trabajar con la consulta del usuario, sino que además se explica cómo se llevan a cabo las fases de recuperación, selección y agrupamiento de documentos.

En [Romagnano y Marchetta, 2017], [Romagnano, Marchetta y Domínguez, 2017] y en [Romagnano, Domínguez y Marchetta, 2017] se presenta 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. En el presente trabajo se evalúa esa propuesta brindando resultados concretos a través de un caso de estudio en el dominio del turismo.

6. Conclusiones

La investigación fue motivada por la posibilidad de aportar precisión, simplicidad y rapidez ante la búsqueda de recursos web relacionados con el dominio del turismo. 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 de recuperación de información web, el cuál como ya cuenta con información clasificada obtiene resultados exactos y en menor tiempo. Esta ganancia en precisión y tiempo se logra debido a que se analiza la semántica tempranamente y durante la mayor parte de las fases.

Se planteó un framework para asistir al turista web, en el cual implícitamente se propone un método, para recuperar información web, focalizándose en la clasificación y en el agrupamiento solapado. Clasificación debido a que de ante mano y con ayuda del experto del turismo se establecen cuáles serán los grupos. Agrupamiento soft en cuanto a que se permite que un recurso pueda pertenecer a varios grupos con determinado grado de pertenencia.

En cuanto al cálculo del representante y de la similitud, se propone una forma que se diferencia de las usadas por la mayoría de los trabajos actuales. Al proponer grupos solapados, se escogen rápidamente aquellos recursos que sólo brindan información explícita de un término solicitado por el usuario o se seleccionan los que ofrecen información de más de un término ubicando la intersección de los grupos comprometidos.

El aporte significativo de este trabajo se evidenció con el análisis de los resultados obtenidos por el framework propuesto al ser aplicado a un complejo dominio como lo es el turismo, en la provincia de San Juan.

7. Referencias

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). V. 7, no 1, article 3, pp. 1-27. DOI: 10.1145/2435215.2435218.

Carrizo, G. (2018). “Proyecto: San Juan Capital Nacional del Turismo Astronómico”. Disponible: https://exactas.unsj.edu.ar/2018/04/19/capital-turismo-astronomico.

Dahab, Y., Kamel, M. and, Alnofaie, S. (2017). “An Empirical Study of Documents Information Retrieval Using DWT”. In: Shaalan K., Hassanien A., Tolba F. (eds) Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence, vol 740. Springer, Cham, DOI: https://doi.org/10.1007/978-3-319-67056-0_13, ISSN: 978-3-319-67056-0, pp.251-264.

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 Map Reduce”. International Journal Engineering Applications of Artificial Intelligence. vol. 51, pp. 16-23, Mayo de 2016. Disponible: https://doi.org/10.1016/j.engappai.2016.01.023.

Ferran Segura, A. (2011). “Evaluación de la experiencia Google: temas de satisfacción y mejoras al diseño de la búsqueda web”. Pp. 40-42, 2011. Disponible: https://repositori.upf.edu/handle/10230/13010.

Ghosh, S. and, Kumar, D. (2013). “Comparative Analysis of K-means and Fuzzy C-means Algorithms”. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 4(4), pp. 35-39, 2013. Disponible: http://dx.doi.org/10.14569/IJACSA.2013.040406.

Hernández, S., Fernández, C. and, Baptista, L. (2010). “Metodología de la Investigación”. 5ta Edición. McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V. ISBN: 978-607-15-0291-9. México, 2010.

Hernández, I., Rivero, C., Ruiz, D. and, Corchuelo, R. (2016). “CALA”. Journal of Systems and Software. DOI: 10.1016/j.jss.2016.02.006.Vol. 115(C), pp. 130-143, Mayo de 2016, Elsevier Science Inc. New York, NY, USA.

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). ISSN: 2395-5317. Vol. 2(2), pp. 59-62, Junio de 2015.

Li, B. and, Han, L. (2013). “Distance weighted cosine similarity measure for text classification”. 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2013) – Vol 8206, pp. 611-618, 20 al 23 de Octubre, 2013, Hefei, China.

Li, M. and, Cao, S. (2014). “A Serie Method of Massive Information Storage, Retrieval and Sharing”. Mechatronics and Automation (ICMA), IEEE International Conference on, pp. 1171-1175, 3 al 6 de Agosto de 2014, Tianjin, China, 2014.

Liu, B. (2007). “Web Data Mining – Exploring Hyperlinks, Contents and Usage Data”. Springer-Verlag Berlin Heidelberg,139. Department of Computer Science University of Illinois at Chicago 851 S. Morgan Street Chicago, IL 60607-7053,USA, 2007.

Liu, C., Wang, W., Tu, G., Xiang, Y., Wang, S. and, Lv, F. (2017). “A New Centroid-Based Classification Model for Text Categorization”. Journal Knowledge-Based

Systems. Vol. 136(C), pp. 15-26, Noviembre de 2017. Disponible: https://doi.org/10.1016/j.knosys.2017.08.020.

Matsumoto, T. and, Hung, E. (2010). "Fuzzy Clustering and Relevance Ranking of Web Search Results with Differentiating Cluster Label Generation". In Fuzzy Systems (FUZZ), 2010 IEEE International Conference on, 18-23 de Julio, Barcelona, 2010, pp.1-8.

Mikawa, K., Ishida, T. and, Goto, M. (2011). "A proposal of extended cosine measure for distance metric learning in text classification". 2011 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1741-1746, 9-12 de Octubre de 2011, Anchorage, AK, USA.

Porter, S. (2017). "The Porter Stemming Algorithm". Disponible: https://tartarus.org/martin/PorterStemmer/.

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. ISSN 2150-7988. Vol. 5, pp. 126-136, 2013. Disponible: http://mirlabs.org/ijcisim/regular_papers_2013/Paper88.pdf.

Romagnano, M., Aciar, S. and, Marchetta, M. (2015). "Method to Reduce Complexity and Response Time in a Web Search". International Journal of Information Technologies and Systems Approach (IJITSA). vol. 8(2), pp. 32-46, July 2015.

Romagnano, M. and, Marchetta, M. (2017). "Improving the Overlapping Clustering of Web Resources for a Specific Domain". In Proceeding of 2017 XLIII Latin American Computer Conference (CLEI). 4 al 8 de Septiembre, Córdoba, Argentina. DOI: 10.1109/CLEI.2017.8226430. ISBN: 978-1-5386-3057-0.

Romagnano, M., Marchetta, M. and, Dominguez, P. (2017). "Metodología para Asistir al Usuario Web en su Búsqueda de Información en un Dominio Específico". Revista Electrónica Argentina-Brasil (ReABTIC). Vol. 1(7), Agosto de 2017. ISSN 2446-7634.

Romagnano, M., Dominguez, P., Marchetta, M. and, Aciar, V. (2015). "Reduciendo la Complejidad de Búsqueda Web en Base a las Necesidades del Usuario". In Proceedings of the 3º Congreso Nacional de Ingeniería Informática y Sistemas de Información (CONAIISI2015), 19 y 20 de Noviembre de 2015, Bs. As. Argentina. Disponible: http://conaiisi2015.utn.edu.ar/memorias.html.

Romagnano, M., Dominguez, P. and Marchetta, M. (2017). "Mejorando la Búsqueda Web en un Dominio Específico". 8º Simpósio de Tecnologia da Informação da Região Noroeste do Rio Grande do Sul. 5º Seminario Argentina-Brasil de Tecnologías de la Información y de la Comunicación. 18º Fórum de Informática, Facultad Três de Maio - Três de Maio - Rio Grande do Sul – Brasil. Disponible: http://websites.setrem.com.br/stin/programacao.

Snowball, P. (2017). "Proyecto Snowball". Disponible: http://snowball.tartarus.org/algorithms/spanish/stop.txt.

Solarte, P. and, Millán, G. (2014). “Propuesta para Extender Semánticamente el Proceso de Recuperación de Información”. Revista Escuela de Ingeniería de Antioquia (EIA), ISSN 1794-1237, Año XI, V. 11, Edición N° 22, pp. 51-65, 2014.

Sote, A. and, Pandey, S. (2015). “Web Page Clustering Using Self-Organizing Map”. International Journal of Computer Science and Mobile Computing, vol. 4, no 1, pp. 78-84, Enero de 2015. Disponible en https://pdfs.semanticscholar.org/618e/621c03e34c141e8397a194a87df933d8ad29.pd.

Xiao, L. and, Hung, E. (2008). “Clustering Web-Search Results Using Transduction-Based Relevance Model”. In IEEE 1st pacific-asia workshop on web mining and web-based application, Osaka, Japón, 2008.

Xing, H. and, Ha, M. (2014). “Further Improvements in Feature-Weighted Fuzzy C-means”. Information Sciences, Elsevier. Vol.267, pp. 1-15, 20 de Mayo de 2014. Disponible: https://doi.org/10.1016/j.ins.2014.01.033.

Yan, W., Zhang, B., Ma, B. and, Yang, Z. (2017). “A Novel Regularized Concept Factorization for Document Clustering”. Journal Knowledge-Based Systems, vol. 135(C), pp. 147-158, Noviembre de 2017. Disponible: https://doi.org/10.1016/j.knosys.2017.08.010.

Yue, L., Zuo, W., Peng, T., Wang, Y. and, Han, X. (2015). “A Fuzzy Document Clustering Approach Based on Domain-Specified Ontology”. Journal Data & Knowledge Engineering. Vol. (A), pp. 148-166, November, 2015. Disponible: https://doi.org/10.1016/j.datak.2015.04.008.