Semantic Recommender System for the Recovery of the Preserved Web Heritage

Omar Portilla, José Aguilar*
Departamento de Computación
Universidad de los Andes
Mérida, Venezuela
aguilar@ula.ve, *oportillajaimes@yahoo.es*
*Prometeo researcher, Universidad Técnica Particular de Loja, Ecuador

Claudia León
Universidad Central de Venezuela
Caracas, Venezuela
claudia.leon@ciens.ucv.ve

***Abstract*— This paper presents a prototype of a semantic personalized recommender system for a repository of preserved web files. To do this, we design and implement a semantic repository of preserved web files, containing metadata associated with each preserved site. The knowledge stored in the metadata of the semantic repository is used for the recommender system, in order to give prioritized recommendations of the different preserved web files (or web heritage) that meet certain search criteria. The proposed recommender also considers semantic associations, in order to recommend not only the websites matched to the search criteria, but also semantically related.**

Keywords*— *Semantic recommender, patrimony web, web service, web archiving

I. INTRODUCCION

Desde hace más de dos décadas se ha venido trabajando en la preservación del patrimonio web (a lo cual se le denomina algunas veces preservación del patrimonio digital), que consiste en preservar sitios web seleccionados del ámbito cultural, técnico, educativo, entre otros. Según [1] el proceso de preservación se realiza mediante la migración de los sitios web a preservar a un nuevo formato, almacenándose en servidores de preservación en un ambiente seguro. De esta forma se garantiza que los contenidos de los sitios web seleccionados estén disponibles, puedan ser consultados y preservados de alteraciones o eliminaciones. Por tanto, la finalidad de los servidores de archivo web preservados, es la preservación y difusión de recursos digitales para que puedan servir como herramienta de conocimiento para generaciones presentes y futuras.

La principal organización encargada de la preservación de patrimonio web es el Consorcio Internacional de Preservación de Internet [2] (IIPC por sus siglas en inglés), para lo cual han propuesto una arquitectura funcional para archivos web, que se basa en el modelo de referencia Open Archive Information System (OAIS). La IIPC es una organización que aglutina las iniciativas más importantes a nivel mundial en el campo del archivado web, en la que se integran bibliotecas nacionales de todo el mundo, así como instituciones patrimoniales (archivos y bibliotecas universitarias y de investigación). Basados en lo propuesto por IIPC se han venido implementando algunas arquitecturas a usar para la preservación de la web. En [3] se presenta una arquitectura basada en software libre, en la que se consideran herramientas existente para integrarlas en la preservación de la web.

A medida que avanza el tiempo se empieza a contar con un alto volumen de archivos web (patrimonio web preservado), motivo por lo cual se hace necesario tener herramientas de búsquedas efectivas. Se requiere que tales búsquedas no sólo se limiten a informar que recursos web se tienen preservados, sino que además se encarguen de recomendar en forma priorizada y personalizada cada uno de los recursos web preservados que se tengan, que se ajusten a los criterios de búsqueda de cierto usuario.

En este trabajo se presenta un prototipo de un sistema recomendador semántico personalizado de archivos web que está pensado para insertarse como una herramienta adicional en la arquitectura propuesta en [3]. Para ello, se diseñó e implementó un repositorio semántico de archivos web preservados, que contiene metadatos asociados con cada sitio web preservado y almacenado en el repositorio de archivos warc mostrado en [3]. El repositorio semántico es usado para que el sistema recomendador acuda al conocimiento de metadatos almacenado en él, y pueda dar recomendaciones priorizadas de los distintos documentos web preservados (o patrimonio web) que se ajustan a cierto criterio de búsqueda. El recomendador propuesto considera, además, asociaciones semánticas que no sólo recomienden los sitios web solicitados con el criterio de búsqueda, sino que recomienden sitios web relacionados semánticamente a él. Adicionalmente, la recomendación entregada es personalizada de acuerdo con el perfil del usuario que esté solicitando las recomendaciones.

Con este trabajo, se busca entregar una herramienta adicional a las propuestas en la arquitectura presentada en [3]. La inserción del sistema de recomendación en la arquitectura se ilustra en la Figura 1. Tal figura muestra el sistema recomendador implementado como un servicio web. El sistema recomendador se implementa como un sistema de recomendación basado en contenidos, que utiliza un sistema de inferencia con lógica descriptiva, para realizar recomendaciones de documentos web preservados. Cualquier máquina que lo requiera acude al sistema recomendador e

978-1-4673-9143-6/15/$31.00 ©2015 IEEE

invoca el servicio mediante una interacción por mensajes SOAP. La recomendación entregada la realiza basado en el conocimiento almacenado en el repositorio semántico de sitios web preservados.

Existen trabajos de máquinas y herramientas de búsqueda de recursos web preservados, como por ejemplo: solr [4], Memento Time Travel [5], NutchWAX (Nutch with Web Archive eXtensions) [6], WERA (WEb aRchive Access) [7], Wayback Machine [8] y Xinq (XML INQuire) [9]. Ahora bien, nuestro trabajo se diferencia de ellos en los siguientes aspectos:

- Cuenta con un perfil de usuario que permite entregar recomendaciones personalizadas, acorde con los requerimientos particulares de cada usuario.
- Es un sistema de recomendación semántico personalizado de archivos web. El sistema entrega recomendaciones no solo de los sitios web que se solicitan directamente, sino que además se recomiendan sitios web relacionados, a través de descripciones semánticas incorporadas como conocimiento en el recomendador. Sistemas de este tipo no se encontraron en la revisión de literatura realizada.
- Presenta un módulo de emparejamiento que realiza recomendaciones sugeridas basadas en los cálculos realizados por una función matemática, que define el grado de emparejamiento de un conjunto de sitios web con la solicitud realizada.

El artículo se compone de cinco secciones: Una primera sección de introducción, una segunda sección que describe el área de los sistemas recomendadores; en la tercera sección se describe la arquitectura del sistema recomendador, una cuarta sección está dedicada a la elaboración de pruebas y análisis de resultados, y una última sección presenta las conclusiones.

II. SISTEMA DE RECOMENDACION BASADO EN CONTENIDOS

Sintetizando lo planteado en [10],[11],[12] y [13], los sistemas recomendadores son sistemas que ayudan a emparejar a usuarios con productos. Se consideran como agentes software que contienen los intereses individuales y particulares de un consumidor, y hacen recomendaciones de acuerdo a ello. Ellos proveen sugerencias de calidad a un consumidor, en el instante en que requiera buscar y seleccionar algún tipo de producto online. De esta manera se obtiene un tipo de recomendación personalizada de acuerdo con las características y necesidades de los usuarios. Por tanto, un sistema recomendador debe contar con un modelo de usuario (en el que se pueden tener las características del usuario o estadísticas de uso), que describa las particularidades de cada usuario, y con ello poder realizar sugerencias personalizadas. Existen distintos tipos de sistemas de recomendaciones:

- Recomendación colaborativa: Lo que busca es recomendar a un usuario con base a sus anteriores escogencias, o las escogencias de otros usuarios cuyos perfiles sean muy similares. En este tipo, los usuarios asignan calificaciones a las recomendaciones ya usadas (así el sistema puede ir aprendiendo que tan bien recomendó), o puntuación a los productos usados (describiendo con ello la calidad del producto). Parte del supuesto de que usuarios que hayan tenido ciertas tendencias de uso, las pueden tener aquellos que tengan perfiles similares.
- Recomendación basada en contenidos: Consiste en emparejar los mejores productos con las preferencias de un usuario. En este caso, para recomendar se parte de una descripción de las características del producto y de los usuarios, las cuales son comparadas a través de una función de emparejamiento.
- Recomendación basada en conocimiento: En este tipo, el recomendador trata de enseñarle al usuario lo que puede adaptarse a sus necesidades.
- Recomendación híbrida: En este tipo de recomendación se combinan dos o más de las técnicas anteriores, buscando un tipo de recomendación lo más acertada posible.

Los sistemas recomendadores tienen el potencial para proveer sugerencias (recomendaciones) de calidad a un consumidor, en el instante en que requiera buscar y seleccionar algún tipo de producto. Lo que busca un sistema recomendador es que cada usuario obtenga un tipo de recomendación personalizada de acuerdo con sus características y necesidades. Por tanto, un sistema recomendador debe contar con un modelo de usuario (en el que se pueden tener características del usuario o estadísticas de uso), que le indique las particularidades de cada usuario, y con ello poder realizar sugerencias personalizadas.

El recomendador presentado en este trabajo es un recomendador basado en contenidos que busca emparejar los mejores productos (documentos web preservados) con las preferencias de un usuario. Para recomendar, el sistema de recomendación parte de una descripción de las características del producto (documentos web preservados), que se almacenan en un repositorio semántico.

Existen diversos trabajos asociados al uso de los sistemas recomendadores, como los propuestos desde hace ya casi una década en [16] y [17], y los sintetizados en revisiones recientes en [10], [11], [12] y [13]. Ahora bien, hasta el momento no se encuentra evidencia alguna en la literatura de que existan sistemas recomendadores semánticos personalizados de archivos web. Cabe resaltar que existen diversidad de tipos de sistemas recomendadores, pero ninguno dedicado a la recomendación semántica de archivos web.

III. ARQUITECTURA DEL SISTEMA RECOMENDADOR DE ARCHIVOS WEB PRESERVADOS

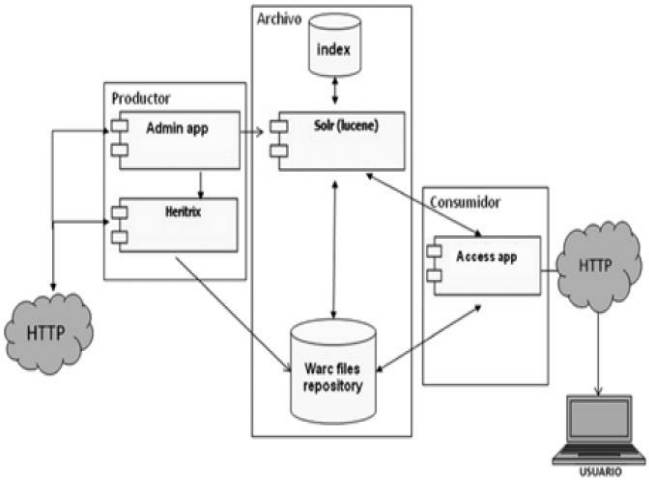
Este trabajo propone incorporar una herramienta de recomendación de recursos web preservados, dentro de la arquitectura presentada en [3]. Así, la arquitectura propuesta en

[3] es extendida con un sistema de recomendación de recursos web preservados. La arquitectura presentada en [3] se ilustra en

Fig. 1, la cual incorpora la herramienta apache solr [4], que consiste en un motor de búsqueda de código abierto, que proporciona una capa de abstracción sobre Apache Lucene [14], y que se encarga de realizar una indexación de los archivos warc provistos por el productor Hertrix, para posteriormente ofrecer un servicio de búsqueda clásica sobre los recursos web indexados.

En este trabajo se propone incorporar un sistema recomendador de documentos web preservados basado en Lógica Descriptiva (que posibilita desarrollar una herramienta de búsqueda adicional a solr) tal como se ilustra en Fig. 2. La Fig. 2 evidencia que el único componente adicional a lo plasmado en [3] es un sistema recomendador. El sistema recomendador propuesto genera recomendaciones para cierto criterio de búsqueda solicitado por un usuario en particular, priorizando y personalizando sus recomendaciones.

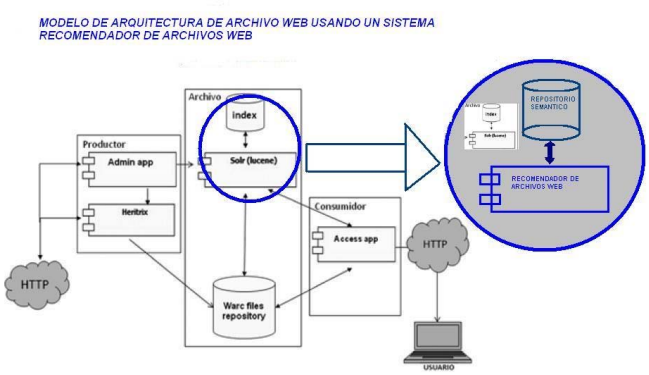
Fig. 1. Arquitectura de archivo web basada en software libre propuesta en [3].



El recomendador está implementado como un servicio web que acude a un repositorio semántico de archivos web para realizar los razonamientos necesarios, y con ello generar las recomendaciones. En este trabajo se plantea que un usuario experto administre e incorpore el conocimiento al repositorio semántico, en el que se describen metadatos de cada uno de los sitios web preservados, en forma complementaria al proceso de indexación clásico que se propone en [3] y que realiza solr.

El recomendador se implementó como un servicio web para posibilitar dos aspectos fundamentales. En primer lugar, la interoperabilidad entre diversas plataformas que requieran del uso del recomendador. Y en segundo lugar, porque la tecnología de Servicios Web permite la interacción transparente maquina maquina, posibilitando así que una aplicación cliente consulte el sistema recomendador directamente. Este último aspecto permite que el sistema recomendador propuesto en este trabajo pueda ser reutilizado por otras aplicaciones en el futuro.

Fig. 2. Modificación de la arquitectura de archivos web propuesta en [3].



Uno de los potenciales más significativos del sistema recomendador propuesto consisten en que no busca archivos web preservados únicamente por el concepto solicitado, sino que cuenta con una descripción semántica que le permite realizar recomendaciones de documentos que contienen conceptos asociados, teniendo en cuenta que deben recomendarse con menor prioridad que los documentos asociados directamente con el concepto solicitado por el cliente.

En la Fig. 3 se muestra la arquitectura interna del sistema recomendador, planteado como herramienta adicional a la arquitectura planteada en [3]. Como se aprecia en la figura 3, el sistema recomendador se encuentra formado por cuatro componentes fundamentales: servicio web recomendador, repositorio semántico del patrimonio web, servicio web de perfil de usuarios, y descriptor de perfiles de usuario. Además, como lo muestra la Fig. 3, en el sistema de recomendación se llevan a cabo seis pasos:

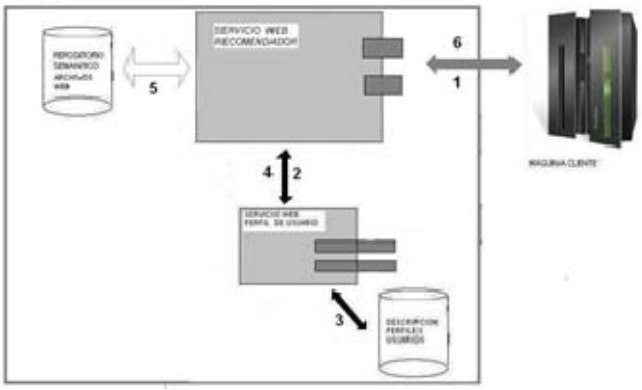
1. Una aplicación cliente solicita al servicio web recomendador una recomendación de que documentos web preservados (patrimonio web) sugiere que valen la pena recuperar del repositorio de archivos web preservados para cierto tema y usuario.
2. El servicio web recomendador solicita al servicio web de perfil de usuario que le entregue el perfil del usuario para el cual recibió una solicitud de recomendación.
3. El servicio web de perfil de usuario realiza una búsqueda en el descriptor de perfiles de usuario.
4. El servicio web de perfil de usuario entrega una descripción detallada del perfil del usuario para el cual se va a realizar la recomendación. La información entregada al servicio web recomendador contiene dos aspectos fundamentales, que son: preferencias explicitas dadas por el usuario, e historial de uso del usuario. El historial de uso estipula que recursos web archivados como patrimonio web han sido recuperados por el usuario, registrando

indicadores de uso, como por ejemplo: en que fechas y cuantas veces ha recuperado cada recurso web

5. El servicio web recomendador realiza una búsqueda en el repositorio semántico de patrimonio web, para obtener los metadatos de los sitios web preservados que formarán parte del conjunto de recomendación, aplicando a cada uno de los miembros del conjunto de recomendación una función de emparejamiento que indica la idoneidad de cada documento web preservado con lo solicitado y el usuario.
6. El servicio web recomendador retorna a la máquina cliente un conjunto de recomendación priorizado y personalizado (de acuerdo a las preferencias de usuario y a sus indicadores de uso), en el que sugiere no solo recursos web preservados asociados únicamente al concepto solicitado, sino también un patrimonio web asociado a conceptos vinculados mediante algún nivel de descripción semántica.

A continuación se describen cada uno de los componentes que integran el sistema de recomendación acá propuesto, que interactúan entre sí, tal como se explicó anteriormente.

Fig. 3. Arquitectura del Sistema Recomendador de Archivos Web



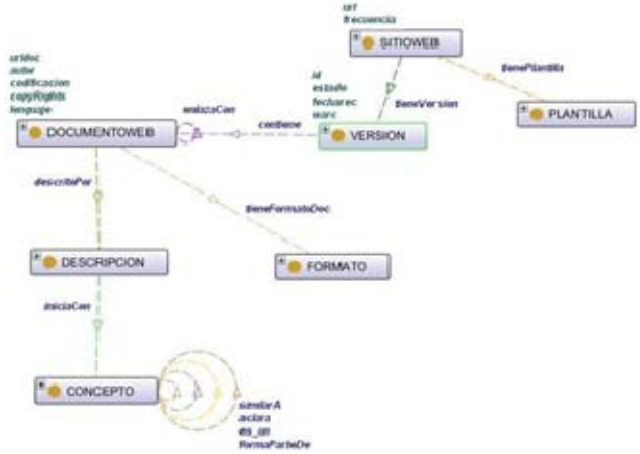
A. REPOSITORIO SEMANTICO DE PATRIMONIO WEB

Un repositorio de patrimonio web es un tipo especial de repositorio digital, es decir, es un depósito de sitios web previamente creados, donde se almacena y mantiene conocimiento de los documentos web preservados. Un repositorio de archivos web clásico contiene y almacena los distintos sitios web preservados, existiendo muchos formatos para almacenar los sitios web, uno de los más conocidos es el warc [15]. En el caso de la Arquitectura funcional IIPC basada en el modelo OAIS propuesta en [1], el repositorio planteado se ajusta completamente al repositorio de archivos warc que se contempla en [3], y su función es almacenar cada uno de los sitios web preservados (consistente en un archivo warc) que entrega el productor propuesto en la arquitectura que plantea [3].

Un *repositorio semántico de patrimonio web* es un tipo especial de repositorio de patrimonio web, que se propone en este trabajo, donde en lugar de tener los sitios web preservados (es decir, en lugar de almacenar los archivos warc), se tiene solo los metadatos de ellos. Para el almacenamiento de dichos metadatos se propone usar un modelo ontológico, en el que se especifique el conocimiento asociado con cada sitio web preservado, dicho modelo ontológico se muestra en la Fig. 4. La gestión del conocimiento almacenado en dicho modelo ontológico, debe ser realizada por un administrador del repositorio, mediante una interfaz que le permita describir semánticamente cada uno de los sitios y documentos web preservados en los archivos web.

En la ontología asociada con el repositorio semántico se almacena la información relacionada con los recursos web preservados. Como se muestra en la Fig. 4, en el modelo ontológico se almacena la información de cada uno de los documentos web preservados, vinculándolos al sitio web al que pertenecen (patrimonio web preservado). Cada documento web preservado en el modelo ontológico está relacionado con un concepto o tema principal que trata el documento, al cual se le conoce como concepto base. Esta asociación se realiza a través de una descripción semántica realizada para cada documento. con ello se pretende realizar búsquedas con mayor nivel de significado, de los conceptos incorporados en cada uno de los sitios web. Además de la descripción semántica, el modelo ontológico incorpora una serie de objectProperty y dataProperty que permiten proporcionar conocimiento de cada uno de los documentos preservados en el archivo web. En la ontología del repositorio semántico se almacena información de cada una de las versiones de los sitios web preservados, plasmando además conocimiento básico de cada uno de los documentos guardados en cada versión. Es importante resaltar que el modelo ontológico acá presentado está basado en lógica descriptiva, y fue especificado en el lenguaje OWL [18], el cual representa un modelo formal de representación de conocimiento en lógica descriptiva mucho más formal que, por ejemplo, las propuestas dadas en [19].

Fig. 4. Diseño del repositorio semántico de archivo web preservados



La descripción semántica consiste en definir la relación de los conceptos de acuerdo a cuatro niveles de descripción a considerar, que son: jerarquización, explicación, comparación y asociación. De esta manera, con la descripción semántica se pretende potencializar a la maquina con conocimiento que le permita no solo recomendar directamente el concepto solicitado por el cliente, sino además recomendar conceptos asociados con él. A continuación se explican cada uno de los cuatro niveles de descripción semántica incorporados en el modelo ontológico, tal como se ilustra en la Fig. 5.

1) JERARQUIZACIÓN:

Mediante esta relación de conceptos se plasma que el concepto A es un concepto incluido dentro de la jerarquía del concepto B. Con ello la maquina infiere que el concepto B es un concepto más genérico que el buscado, pero que puede ser recomendado, ya que este concepto permite conocer los aspectos fundamentales y más genéricos del concepto A solicitado. La jerarquización se estipula mediante el objectProperty *es_un*.

2) EXPLICACION:

Esta relación busca expresar que un concepto C aclara o profundiza sobre cierto aspecto del concepto A. Mediante este tipo de relación la maquina infiere que los conceptos relacionados mediante la relación *aclara* lo que buscan es profundizar o describir aspectos particulares del concepto A (concepto base). Para indicar que un concepto es explicado por otro se recurre al objectProperty *aclara*.

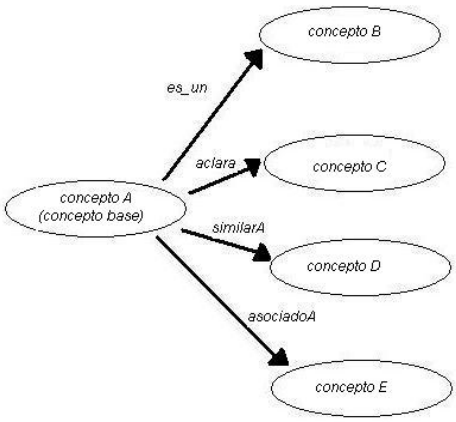
3) COMPARACION:

Este tipo de relación es expresada mediante el objectProperty similarA. Es una relación usada cuando dos conceptos tienden a ser similares en cuanto a funciones, composición o definición. En otras palabras, si el concepto A está relacionado con el concepto D mediante comparación, lo que se indica es que los conceptos A y D son muy parecidos ya sea en cuanto a su uso, composición o definición.

4) ASOCIACION

Este nivel de descripción semántica muestra con que conceptos interactúa o se puede asociar un concepto base. La asociación se realiza mediante el objectProperty *asociadoA*. Cuando un concepto A se asocia con un concepto E mediante esté relación, lo que se expresa es que dichos conceptos tienen algún grado de interrelación entre sí.

Fig. 5. Niveles de descripción semántica usados en el modelo ontológico.



B. SERVICIO WEB DE PERFIL DE USUARIO Y DESCRIPTOR DE PERFILES DE USUARIO

Es un servicio web invocado por el servicio recomendador para obtener conocimiento de las preferencias del usuario y de su historial del uso de documentos web. El servicio recibe como parámetro de entrada un identificador de usuario, con el cual acude al sistema de razonamiento basado en lógicas descriptivas para consultar un descriptor de perfiles de usuario, cuyo modelo ontológico se muestra en la Fig. 6. Una vez consultado el descriptor de perfiles de usuario, el servicio web retorna las preferencias del usuario almacenadas ,y un arreglo de historias de usuario en las que se describe cada uno de los documentos usados por el usuario, junto con conocimiento adicional que permite calcular INDICADORES DE USO de cada uno de los documentos, tal como se definen en el numeral 3 de esta sección.

Fig. 6. Modelo ontológico del descriptor de perfil de usuarios.

El modelo ontológico del descriptor de perfil de usuarios consiste fundamentalmente en lo mostrado en la Fig 6. Es importante resaltar que en este modelo ontológico, además de los criterios de preferencias básicos (max_antiguedad, formato_pref), se cuenta con un historial de uso de cada uno de los documentos. En cada historia almacenada en la ontología se registra conocimiento de qué documento se asocia a la historia, que concepto describe dicho documento, cuantas veces ha sido visitado por el usuario, y cuál es la última fecha de visita del usuario al documento.

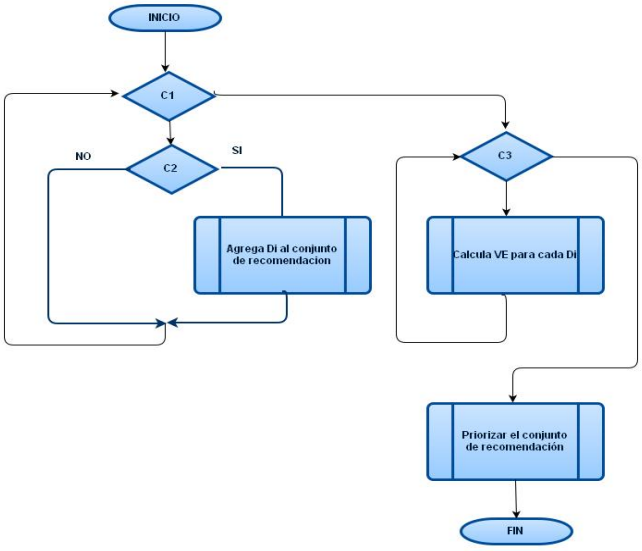
C. SERVICIO WEB RECOMENDADOR

El núcleo fundamental del servicio web recomendador es el proceso de emparejamiento, el cual consiste en emparejar (comparar) las características solicitadas por el usuario (con su histórico comportamental) con las características de cada uno de los documentos almacenados en el archivo web. Para ello se propone una función de emparejamiento (f), o función de coincidencias, que consiste en una relación matemática que calcula un valor, llamado *VALOR DE EMPAREJAMIENTO (VE)* asociado a cada uno de los documentos candidatos a recomendar, que se agrupan en un conjunto llamado conjunto de recomendación.

La Fig 7 ilustra el proceso llevado a cabo para realizar el emparejamiento. Para ello, el servicio recomendador acude al repositorio semántico de archivos web preservados. Una primera fase del proceso se encarga de la construcción del conjunto de recomendación, para lo cual el razonador basado en lógica descriptiva le proporciona al servicio recomendador el conocimiento referente al grupo de documentos cuyo concepto se asocia con el concepto solicitado en la recomendación, a través de alguno de los cinco niveles de descripción semántica planteados en la sección III.A. Como se ilustra en la Fig. 7, cada vez que se cumpla la condición C1 (mientras exista un documento D_i sin analizar registrado en el repositorio semántico) se verifica si el concepto asociado con el documento analizado D_i cumple con la condición C2 (si el concepto asociado con D_i está vinculado con el concepto solicitado en la recomendación, mediante alguno de los cinco niveles de descripción semántica existente). Una vez construido el conjunto de recomendación (conjunto de documentos a recomendar), el servicio recomendador procede a aplicar a cada uno de los documentos pertenecientes al conjunto de recomendación la función de emparejamiento (calcula VE), valor que sirve como criterio de priorización en el momento de recomendar. Dicha labor la realiza mientras existan documentos pendientes de procesar en el conjunto de recomendación (condición C3 de la Fig 7). Para cada uno de los documentos del conjunto de recomendación se usa la ecuación (1), que permite calcular el valor de emparejamiento VE mediante la función f, para indicar que tan recomendable es ese candidato para la solicitud realizada. Cuando ya se termina el proceso de cálculo del VE de cada documento perteneciente al conjunto de recomendación, el recomendador organiza (prioriza) los elementos del conjunto en forma descendente de acuerdo a su valor VE. El valor de VE, como ya se mencionó, indica que tanto empareja ese documento con la recomendación solicitada por cierto usuario.

Antes de describir la función de emparejamiento se explican las variables usadas por la función.

Fig. 7. Proceso de emparejamiento de documentos web



1) NIVEL DE EXPRESIVIDAD SEMANTICA (NE)

Es un valor que indica que tanto peso se asigna en la función de emparejamiento al NIVEL DE DESCRIPCION SEMANTICA en que se encuentra el concepto base asociado al documento con el concepto buscado por el usuario (relación semántica entre los dos), para lo cual se usa la Tabla 1.

TABLE I. CALCULO DEL NIVEL DE EXPRESIVIDAD SEMÁNTICA

NIVEL	VAL	DESCRIPCION
CONCEPTO BASE	25	El concepto es el concepto base
JERARQUIZACION	20	El concepto está en el nivel de descripción de jerarquización
EXPLICACION	15	El concepto está en el nivel de descripción de explicación
COMPARACION	10	El concepto está en el nivel de descripción de comparación
ASOCIACION	5	El concepto está en el nivel de descripción de asociación

2) PREFERENCIAS DE USUARIO (VP)

Son los criterios de preferencia usados para emparejar. En el caso del sistema recomendador se usaron solo dos, que son: formato y antigüedad de la versión más reciente del documento. Ello no implica que un sistema recomendador no pueda usar más preferencias de usuario, para ello se debería extender los criterios a considerar. Su forma de cálculo (el VALOR DE EMPAREJAMIENTO) se muestra en el segundo término de la ecuación (1).

3) INDICADOR DE USO (IU)

Es un criterio almacenado en cada una de las historias de usuario en el repositorio de perfil de usuario. Para el sistema recomendador presentado se utilizan tres indicadores de uso, que son: IU_1 (número de veces en que el usuario recuperó el documento para el que se está calculando VE), IU_2 (última fecha en que el usuario recuperó el documento) e IU_3 (número de veces en que el usuario recuperó documentos cuyo concepto es el mismo que el concepto para el cual se va a recomendar).

4) RANGO PERCENTIL ASOCIADO A UN INDICADOR DE USO (RP(IU))

Es un valor estadístico que indica el nivel de influencia del valor asociado al indicador de uso para el documento a recomendar. El rango percentil es una forma de comparar el valor asociado al indicador de uso de ese documento respecto a los valores asociados para el mismo indicador de uso en el resto de los documentos almacenados en el historial del usuario que solicita la recomendación. El rango percentil para cada indicador de uso es un número entre 0 y 1, que indica el porcentaje de documentos cuyo indicador de uso son igual o menores que el indicador de uso del documento analizado. En otras palabras, dice cuál es el nivel de influencia del valor registrado para el indicador de uso de ese documento respecto del total. Por ejemplo, si un indicador de uso tiene registrado que ha sido recuperado 10 veces por ese usuario, es necesario determinar que tanto representa esta información respecto del

historial de uso de este usuario. En la medida que el rango percentil se acerque más a 1 se concluirá que el valor de 10 visitas para un documento en el historial de ese usuario es muy significativo.

Como se mencionó al inicio de la sección, la función de emparejamiento es una función cuyo dominio es el producto cartesiano de documento D_i y el usuario U_k. La función de emparejamiento se usa para calcular el valor VE($f(D_i, U_k)$) de cada uno de los documentos que pertenezcan al conjunto de recomendación, de acuerdo con la ecuación (1). El costo computacional de la función de emparejamiento es de complejidad n^2. Se resalta este hecho, debido a que se trata del algoritmo fundamental usado por el sistema recomendador.

$$f(D_i, U_k) = \mathrm{NE}(X_{ij}) + \sum_{k=1}^{np} VP_{i,k} + \sum_{m=1}^{N_{IU}} RP(IU_{m,k}) \qquad (1)$$

donde:

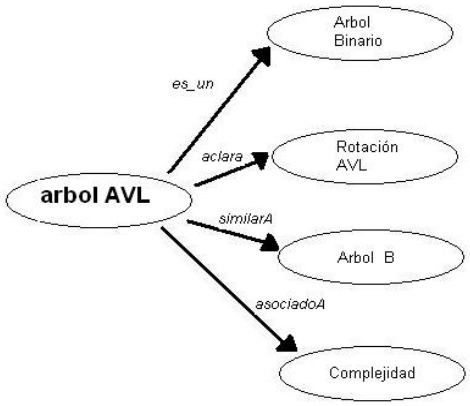
- **NE(C_i)** es el nivel de expresividad semántica que recibe C_i (concepto asociado con el documento D_i), y retorna cierto valor estipulado en la Tabla 1. El valor asignado depende del nivel de descripción semántica en que se encuentre al comparar C_i con el concepto solicitado en la recomendación.
- **np** número de preferencias del usuario estipuladas en el sistema
- **$VP_{i,k}$** valor de preferencia asociado al documento i en la preferencia de usuario k, calculado así: 1 si empareja la preferencia j del usuario con la característica asociada en el documento i, 0 en caso contrario.
- **N_{IU}** es el número de indicadores de uso
- **RP($IU_{m,k}$)** es el rango percentil calculado para el indicador de uso m del usuario k.

IV. EXPERMIENTACION

Para verificar el funcionamiento del sistema recomendador es necesario la realización de dos pruebas fundamentales, que son la prueba de descripción semántica de conceptos y la prueba de personalización de la recomendación según el usuario. Para la realización de estas dos pruebas se implementó una aplicación cliente que invoca al servicio web recomendador de documentos preservados. Las dos pruebas realizadas, y su análisis de resultados, se describen a continuación.

A. PRUEBA DE DESCRIPCION SEMANTICA DE CONCEPTOS.

Fig. 8. Descripción semántica del concepto avl.



Con esta prueba se busca demostrar como el recomendador usa la descripción semántica de conceptos que tiene almacenada como conocimiento dentro del repositorio semántico. Para la elaboración de esta prueba se utilizarán una serie de documentos web en los que el principal concepto descrito (tema o concepto base) son los arboles avl. Como se explicó en la sección 4.1, una descripción semántica consiste en la elaboración de una relación de conceptos de acuerdo a cuatro niveles de descripción a considerar, que son: jerarquización, explicación, comparación y asociación. La Fig. 8 ilustra como para esta prueba se realiza la descripción semántica de conceptos, usando como base el concepto avl.

En este caso se incorporó dentro del repositorio una descripción semántica para el concepto avl, en ella se contempla que es un tipo de árbol binario (descripción en el nivel de jerarquización), que para explicarlo mejor se puede recurrir al concepto de rotaciones avl (nivel de explicación), que en cuanto a su función es similar al concepto árbol B (nivel de comparación), y que el árbol avl está directamente asociado con el concepto de complejidad computacional (nivel de asociación).

En esta prueba se realiza una solicitud de recomendación para documentos web preservados asociados con el concepto avl. El servicio recomendador además de recomendar los documentos web preservados como patrimonio asociados con el concepto avl, recomienda también los documentos preservados que cuentan con conceptos asociados a él, mediante la descripción semántica. Para ello compara el concepto solicitado en la recomendación (avl) con el concepto asociado a cada uno de los documentos, y en base en ello calcula su valor NE. Por ejemplo, en la Tabla II se muestra que para el documento con id rota2 el concepto asociado es rotación avl. Como el concepto rotación avl se encuentra vinculado mediante el nivel de explicación con avl (de acuerdo a la descripción semántica registrada en la ontología para el

concepto avl), el valor de NE es de 15, tal como se estipula en la Tabla I. La prueba evidencia además la forma en que la función de emparejamiento prioriza los niveles de descripción semántica (jerarquización, explicación, comparación y asociación), tal como se describió en la sección 4.3.

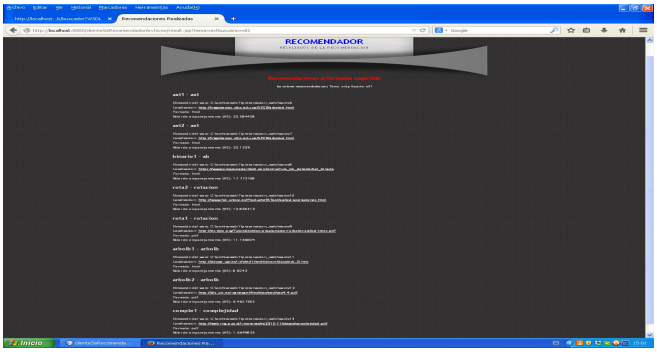
TABLE I. VALORES OBTENIDOS EN LA PRUEBA DE DESCRIPCIÓN SEMÁNTICA DEL RECOMENDADOR

Niv. Descripción.	Concepto	Id docum.	VE	NE	VP+RP
Concepto base		avl1	27,58	25	2,58
		avl2	27,12	25	2,12
Jerarquización	Á. binario	binario1	22,77	20	2,77
Explicación	Rotación AVL	rota2	17,85	15	2,85
		rota1	16,16	15	1,16
Comparación	Árbol B	arbolb1	13	10	3
		arbolb2	11,46	10	1,46
Asociación	complejidad	comple1	6,66	5	1,66

Como se ilustra en la Tabla II, el conjunto de recomendación se encuentra compuesta no solo por documentos preservados cuya temática sea solo la del concepto solicitado en la recomendación, sino que además se compone de documentos preservados cuyas temáticas están asociadas a conceptos que se encuentran en alguno de los niveles de la descripción semántica del concepto base (avl). Los cálculos de VE van cambiando de acuerdo al nivel de expresividad semántica asociado al tema del documento recomendado. Por ejemplo, para el documento preservado cuyo identificador es binario1 se obtiene un VE de 22,77, y el componente de nivel de expresividad aporta 20 debido a que el concepto árbol binario está asociado por jerarquización al concepto solicitado en la recomendación (avl). Por el contrario, el documento preservado con identificador arbolb2 solo tiene un VE=11,46, debido a que el concepto asociado árbol b se encuentra relacionado por nivel de comparación, con un aporte de 10 en su nivel de expresividad NE.

La recomendación sugerida por el servicio web recomendador es mostrada en la Fig. 9. en ella se aprecia que además de recomendar dos documentos asociados directamente con el concepto avl (con VE de 27,5844 y 27,1229, respectivamente), la recomendación involucra además una serie de documentos asociados con conceptos pertenecientes a la descripción semántica del concepto avl. Los documentos recomendados están priorizados de acuerdo al valor de emparejamiento calculado por la función de emparejamiento para cada uno de los documentos web pertenecientes al conjunto de recomendación, tal como se aprecia en la Fig. 9.

Fig. 9. Recomendación realizada al cliente cli1 para documentos preservados con tema avl



B. *PRUEBA DE PERSONALIZACION DE USUARIO.*

Con la realización de esta prueba se demuestra como la priorización de un conjunto de recomendación se realiza acorde al valor de emparejamiento de cada documento, el cual cambia de acuerdo con las preferencias del usuario y su historial asociado. Para ello se acude a probar cuando dos usuarios distintos solicitan recomendación de documentos web preservados asociados a un mismo concepto solicitado (colonia de hormigas). Los resultados obtenidos cambian su priorización para cada cliente. Para el cliente cli1 los resultados arrojados se muestran en la Fig. 10. y para el cliente cli2 en la Fig. 11.

Fig. 10. Recomendaciones realizadas para el concepto coloniahormigas cliente cli1

En la Fig. 11 se muestran las recomendaciones sugeridas por el sistema recomendador para documentos web preservados asociados con el concepto coloniahormigas y el cliente cli2. Como es claro en las Figuras 10 y 11, el conjunto de recomendación está conformado por los mismo cuatro documentos web, pero difiere su forma de priorizar y sus valores de emparejamiento, tal como se muestra en la Tabla II

Los valores de emparejamiento VE mostrados en la Tabla II son los valores calculados por la función de emparejamiento f para cada uno de los miembros del conjunto, tal como se muestra en la ecuación (1). Cada uno de los documentos web

preservados que se agreguen al conjunto de recomendación poseen su propio valor de VE (un valor entre 0 y 30 calculado mediante la función de emparejamiento f). A más alto valor en VE más adecuado es el documento web para el cliente que lo solicite.

TABLE I. SÍNTESIS DE PRUEBAS REALIZADAS PARA COMPROBAR LA PERSONALIZACIÓN DE USUARIO.

RESULTADOS PARA LA PRUEBA DE PERSONALIZACION DE USUARIOS											
hormi1-cli1 VE=29,9397			**hormi2-cli1 VE=29,3396**			**hormi3-cli1 VE=27,7396**			**hormi4-cli1 VE=27,3397**		
NE=25	VP=2	RP=2,93	NE=25	VP=2	RP=2,33	NE=25	VP=1	R=1,73	NE=25	VP=1	RP=1,33
hormi4-cli2 VE=29,9495			**hormi3-cli2 VE=29,5495**			**hormi2-cli2 VE=26,9495**			**hormi1-cli2 VE=26,3495**		
NE=25	VP=2	RP=2,94	NE=25	VP=2	RP=2,54	NE=25	VP=1	R=0,94	NE=25	VP=1	RP=0,34

Como lo muestra la Tabla III, el conjunto de recomendación entregado tanto a cli1 como a cli2 es el mismo (conformado por los documentos con identificador hormi1, hormi2, hormi3 y hormi4), ya que los dos clientes solicitaron recomendación de documentos web preservados cuyo tema o concepto principal sea colonia de hormigas. A este hecho se debe que el componente de nivel de expresividad sea NE=25 para todos los casos, pues el concepto (tema) colonia de hormigas es el concepto principal de los tres documentos encontrados, y no se encontraron documentos web cuyos conceptos asociados tuvieran relacionado otro nivel de expresividad (lo cual ya se evidenció en la prueba anterior).

Ahora bien, el orden es diferente debido al uso de los perfiles de usuarios que influyen en el cálculo de *f*. En particular, el valor aportado por VP en el cálculo del valor de emparejamiento depende de la cantidad de coincidencias que se encuentren en cada una de las preferencias estipuladas por el usuario y los valores registrados en el repositorio para cada documento preservado. Por ejemplo, para el cliente cli2 el documento con identificador hormi4 presenta un valor de VP=2, el cual se obtiene del hecho de que cli2 prefiere el formato pdf (el mismo formato de hormi4, ver tablas IV y V) y el documento hormi4 fue creado hace menos de 1500 días, que es la máxima antigüedad de preservación que prefiere cli2 (ver tabla IV).

Otro ejemplo es el documento web preservado cuyo identificador es hormi1, para el cliente cli1 se tiene un VE=29,9397. Tal valor se obtiene de sumar el nivel de expresividad semántica (NE=25) con el valor aportado por el emparejamiento de preferencias del usuario (VP=2), que indica que las dos preferencias de usuario emparejan, pues el cliente uno prefiere un formato html y una antigüedad máxima de 9000 días (ver Tabla III). Tales valores emparejan con los valores de formato y antigüedad de la última versión del documento preservado y los rangos percentiles (como muestra la tabla III un valor de RP=2,93) asociados a los indicadores de uso, como se ilustra en la Tabla II.

El cálculo de RP se explica a continuación: Como se explicó en la sección III, apartado C, inciso 3, se tienen tres indicadores de uso (IU), para cada uno de los cuales se calcula un valor numérico entre 0 y 1, que indica el nivel de influencia del indicador asignado al documento que se está emparejando respecto del total de documentos recuperados registrados en el historial de usuario. En adelante se ilustra el cálculo de cada uno de los tres IU considerados por el recomendador:

- Calculo de IU_1: como se explicó, IU_1 está asociado con el número de veces que el usuario recuperó el documento para el cual se está emparejando. El valor entregado es el percentil que dicho número de veces representa respecto del total del historial de dicho usuario. La Tabla VI evidencia que en el documento hormi1 (documento que se está emparejando), el usuario recuperó 90 veces, lo cual representa un valor de 1 ya que es el mayor número de veces en todo el historial.
- Calculo de IU_2: hace referencia al número de días transcurridos desde la última recuperación del documento hasta la fecha actual. Como se aprecia en la Tabla VI, desde la fecha actual hasta la fecha de la última recuperación (10/12/09) han transcurrido 1980 días (valor calculado por el recomendador comparando la fecha de la última recuperación y la fecha actual). Como se trata de la fecha más reciente respecto de las demás fechas de recuperación de la historia del usuario cli1 (ver Tabla VI), el recomendador asigna un 1 para este indicador de uso.
- Calculo para IU_3: se refiere al número de veces en que el usuario recuperó documentos cuyo concepto es el mismo que el concepto para el cual se va a recomendar. Como se aprecia en la Tabla V, el único documento preservado que se asocia con el historial del usuario cli1 que no está asociado con el concepto coloniahormigas (concepto solicitado para la recomendación) es el documento con id avl5. Para calcular el RP del indicador IU_3 se suman todos las veces que el usuario recuperó documentos con el tema solicitado (coloniahormigas) y se divide entre el número total de recuperaciones realizadas por el usuario. La Tabla VI muestra que para el cli1 el valor de RP=0,93=187/199, donde 187 es la cantidad de veces que el usuario recuperó documentos con el mismo concepto para el que se está haciendo la recomendación (coloniahormigas), y 199 es el número total de veces que el usuario recupero cualquier tipo de documento. El valor de RP=0,93 para el IU_3 plasma que el 93% de las ocasiones en que el usuario recuperó documentos lo hizo para documentos que tienen asociado el concepto colonia de hormigas.

El cálculo total del tercer sumando en la ecuación (1) se obtiene de la sumatoria de los valores calculados para cada uno de los RP de los tres indicadores de usuario para el cliente cli1 en el documento hormi1, el valor de RP=2,93 (ver Tabla III) se obtiene de sumar 1 (valor asociado a IU_1) con 1 (valor asociado a IU_2) con 0,93 (valor asociado a IU_3)

TABLE I. REFERENCIAS ESTIPULADAS POR EL USUARIO

PREFERENCIAS	CLI1	CLI2
formato	html	pdf
antigüedad (días) ult. Vers.	9000	1500

TABLE II. INDICADORES DE USO DE LOS USUARIOS RESPECTO DEL DOCUMENTO WEB.

Valores en los documentos	formato	fecha creación	Concepto asociado
hormi1	html	10/11/08	coloniahomigas
hormi2	html	10/11/08	coloniahomigas
hormi3	pdf	11/11/11	coloniahomigas
hormi4	pdf	11/11/11	coloniahomigas
avl5	pdf	12/05/04	avl

Es importante notar como los valores proporcionados por los indicadores de uso influencian en el cálculo del valor de emparejamiento. Por ejemplo, si para el cliente cli2, se considera el documento con identificador hormi4 con RP=2,94 (ver Tabla III) y el documento con identificador hormi1 con RP=0,34, se observa que el historial de usuario influencia mucho más en el documento hormi4 que en el documento hormi1. Esta diferencia en el valor de RP se explica al analizar la Tabla VI. En ella se observa que el cliente cli2 recuperó muy pocas veces el documento preservado hormi1 (2 veces), comparado con el número de veces que recuperó los demás documentos web que están registrados en su historial de uso (aspecto que se refleja en el RP=0,34). Mientras que recuperó 100 veces el documento identificado con hormi4, valor bastante influyente dentro de su historial de uso (ver Tabla III), reflejándose en un valor de RP=2.94.

TABLE III. INDICADORES DE USO DE LOS USUARIOS RESPECTO DEL DOCUMENTO WEB.

Documento preservado	Veces recuperado		Ultima recuperación	
	cli1	**cli2**	**cli1**	**cli2**
hormi1	90	2	10/12/09	01/01/11
hormi2	85	20	01/01/07	01/01/12
hormi3	9	85	01/01/06	01/01/13
hormi4	3	100	01/01/05	01/01/14
avl5	12	11	01/01/07	01/02/12

C. ANALISIS DE RESULTADOS.

La prueba de descripción semántica de conceptos muestra como el servicio recomendador, además de recomendar los documentos web preservados como patrimonio asociados con el concepto solicitado (avl), recomienda también los documentos preservados que cuentan con conceptos asociados a él, mediante la descripción semántica. Como se ilustra en la Tabla II, el conjunto de recomendación se encuentra compuesto no solo por documentos preservados cuya temática sea solo la del concepto solicitado en la recomendación, sino que además se compone de documentos preservados cuyas temáticas están asociadas a conceptos que se encuentran en alguno de los niveles de la descripción semántica del concepto base (avl).

La prueba de personalización de usuarios demuestra como la priorización de un conjunto de recomendación se realiza acorde al valor de emparejamiento de cada documento, el cual cambia de acuerdo con las preferencias del usuario y su historial asociado. Para ello se acude a probar cuando dos usuarios distintos solicitan recomendación de documentos web preservados asociados a un mismo concepto. La Tabla III ilustra los valores de VP calculados. El orden es diferente debido al uso de los perfiles de usuarios que influyen en el cálculo de VP.

V. CONCLUSIONES

En la literatura no se encuentra evidencia alguna del concepto de sistema recomendador de patrimonio web basado en un repositorio semántico de patrimonio web. Esto permite concluir que se trata de un uso novedoso de los sistemas recomendadores, que es de gran utilidad al usar lógica descriptiva y su poder de inferencia, para generar recomendaciones personalizadas de patrimonio web preservado.

Para ello, el sistema recomendador presentado introduce el concepto de descripción semántica de conceptos, lo cual representa un aspecto innovador que permite al sistema recomendador mucha más flexibilidad y cobertura en el momento de recomendar. Además, el módulo de emparejamiento involucra tratamiento estadístico de los indicadores de uso. Este aspecto permite evidenciar como se pueden usarse diversas herramientas para lograr un mayor nivel de personalización de recomendaciones sugeridas.

El sistema recomendador presentado ofrece dos ventajas fundamentales sobre herramientas como solr. La primera, su nivel de expresividad semántica, pues permite no solo buscar y recomendar documentos asociados con un tema base, sino que recomienda documentos asociados con otros niveles de descripción semántica. Una segunda ventaja que ofrece el recomendador frente a solr es su nivel de personalización, ya que tal herramienta solo ofrece búsquedas convencionales, sin considerar a qué tipo de usuario va dirigido.

El sistema recomendador desarrollado complementa la arquitectura presentada en [3], porque solr es una herramienta solo para buscar sitios web preservados en servidores que guardan el patrimonio web. El sistema recomendador permite la búsqueda y recomendación de documentos web almacenados en las distintas versiones de cada sitio web preservado.

Las pruebas presentadas ilustran como el sistema recomendador logra cumplir las expectativas en cuanto a personalización de las recomendaciones dependiendo del tipo de usuario, y en cuanto a la potencia para recomendar no solo

documentos cuyo concepto sea el solicitado, sino documentos preservados que se asocian a conceptos que se encuentran en alguno de los niveles de expresividad semántica.

El valor de emparejamiento VE (calculado mediante la función f) es un indicador que permite medir el grado de efectividad del recomendador. A través de él se puede notar como este valor aumenta a medida que los documentos web preservados recomendados emparejan más con las necesidades del usuario. El rango que toma el valor VE va desde 0 hasta 30. Donde un 0 indica que el documento preservado es lo menos recomendado para el usuario que solicita la recomendación, y un 30 indica que es lo más altamente recomendado.

RECONOCIMIENTO

Dr. Aguilar ha sido parcialmente financiado por el Proyecto Prometeo del Ministerio de Educación Superior, Ciencia, Tecnología e Innovación de la República del Ecuador.

REFERENCIAS

[1] Masanés, J. (2006). Web Archive. New York, USA: Springer-Verlag.

[2] IIPC. (2011). International Internet Preservation Consorcium. Recuperado de http://netpreserve. org/

[3] *Ospina Torres, M.H. y León Luna, C.P. (2013). Una arquitectura basada en software libre para archivos web. Enl@ce Revista Venezolana de Información, Tecnología y Conocimiento*, 10 (1), 53-72

[4] http://lucene.apache.org/solr/

[5] https://chrome.google.com/webstore/detail/memento-time-travel/jgbfpjledahoajcppakbgilmojkaghgm?hl=en&gl=US

[6] http://archive-access.sourceforge.net/projects/nutch/

[7] http://archive-access.sourceforge.net/projects/wera/

[8] http://netpreserve.org/netpreserve.org/tools/openwayback

[9] http://sourceforge.net/projects/xinq/

[10] P. Pan, C. Wang, G. Horng, and S. Cheng. The development of an Ontology-Based Adaptive Personalized Recommender System. in Electronics and Information Engineering (ICEIE), 2010 International Conference On. 2010.

[11] Adomavicius, G. and A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE

[12] Burke, R., Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, 2002. 12(4): p. 331-370.

[13] Bobadilla, J. Recommender systems survey, 2013. España

[14] http://lucene.apache.org

[15] ISO. (2009). ISO. 28500 Information and documentation-WARC file format. Nueva Zelanda.

[16] Punam, B; Harmeet, K. and Sudeep, M. Trust based Recommender System for the Semantic Web, 2007, India.

[17] Ziegler, C. Semantic Web Recommender Systems, 2004, Germany.

[18] http://www.w3.org/2001/sw/wiki/OWL

[19] http://www.w3.org/TR/skos-reference/