RECUPERACIÓN AUTOMÁTICA DE OA PARA APOYO A COMPETENCIAS EDUCATIVAS EN UN CURSO SOPORTADA EN MÉTRICAS DE SIMILITUD

Néstor Darío Duque M*, Silvia Baldiris†, Ramón Fabregat†, Sabine Graf‡ y Luis Eduardo Escobar*
*Universidad Nacional de Colombia, †Universidad de Girona, ‡Universidad de Athabasca
e-mail: ndduqueme*, leescobarg*{@unal.edu.co}, baldiris†, ramon.fabregat†{@udg.edu}, sabine‡{@athabascau.ca}

***Resumen*— La obtención relevante de objetos educativos en repositorios es un tema que cobra gran interés en la comunidad académica dada la necesidad de proveer al aprendiz con actividades que cubran sus requerimientos académicos y psicopedagógicos. En este artículo se presenta una propuesta validada para la ponderación de los Objetos de Aprendizaje recuperados en una búsqueda sobre repositorios, orientada a obtener recursos educativos que apoyen el cumplimiento de competencias definidas formalmente mediante IMS RDceo. A partir de todos los metadatos recuperados desde los repositorios de OA especificados en XML, se realiza un proceso que concluye con la selección de los OA cuyos metadatos presentan mayor similitud con lo especificado en la competencia esperada. Se usa un algoritmo de overlap y los algoritmos de Dice y Coseno para calificar los resultados.**

***Palabras claves*— Recuperación de OA, algoritmos de similitud en OA, recomendación de OA.**

I. INTRODUCCIÓN

Los Objetos de Aprendizaje, almacenados en repositorios, ofrecen ventajas de disponibilidad y reutilización de materiales de alto valor pedagógico y evita duplicar esfuerzos en la construcción de los mismos. Pero la utilización de material diseñado y desarrollados por otros requiere la recuperación de Objetos que apunten a cubrir las competencias que se proponen para un curso o área temática concreta y los interés determinados. La obtención relevante de objetos educativos en repositorios es un tema que cobra gran interés en la comunidad académica dada la necesidad de proveer al aprendiz con actividades que cubran sus requerimientos académicos y psicopedagógicos. Desde el punto de vista de la planificación instruccional la definición de los objetivos educativos o competencias es el punto de partida para el diseño y construcción de un curso que incluya las actividades y materiales que permitan lograr estas metas, que claramente reflejan la intencionalidad del proceso de enseñanza-aprendizaje [1]. Estas competencias pueden ser definidas de varias maneras, entre ellas la especificación ims_rdceo, que será la utilizada en este trabajo.

Los Sistemas de Recomendación surgieron a mediados de la década de los 90 con el fin de brindarles a los usuarios aquellos resultados de búsquedas de información cercanos a sus necesidades. En el caso de los materiales educativos, los OAs deben estar etiquetados con metadatos de modo que puedan ser identificados, localizados y utilizados para propósitos educacionales en ambientes basados en Web [2]. En lo anterior radica la importancia de los metadatos en los OA

En el trabajo realizado por Rodríguez y otros (2010) se concluyó que a pesar de existir estándares reconocidos e incluso adoptados, la mayoría de los repositorios no utiliza todas posibilidades, e incluso alguna información en los mismos es ambigua. Los metadatos mas utilizados tienen que ver con la información general que describe el OA como un todo, con el ciclo de vida donde se especifica el autor y la fecha de creación del Objeto de Aprendizaje y en la categoría de Educación, donde se describen las características educacionales o pedagógicas del OA.[3]

Este trabajo presenta una propuesta que permite la ponderación de los Objetos de Aprendizaje recuperados en una búsqueda sobre repositorios y el resultado obtenido. La búsqueda inicial se realiza a partir de las palabras claves incluidas en las competencias educativas (microcontexto de competencias), definidas formalmente mediante IMS RDceo.[4][5]

Los metadatos recibidos desde los repositorios, en formato XML, pasan por diferentes etapas, para lograr aplicar algoritmos de similitud [6][7] sobre las palabras importantes en el contexto de la actividad educativa. Al final bajo un creativo enfoque se utilizan las formulas para el coeficiente DICE y de la Distancia del Coseno, además de una medida genérica de overlap[8] .

El resto del documento está organizado así: La sección siguiente muestra los resultados obtenidos en búsquedas genéricas en diferentes repositorios y su nivel de similitud con lo esperado para un curso específico de Auditoría de Sistemas lo cual ilustra el problema encontrado. La sección 3 muestra la aplicación propuesta, definiendo y desarrollando los diferentes componentes y las herramientas utilizadas en el proceso. El numeral 4 está dedicado a la parte experimental utilizando la definición de competencias de granularidad fina. Se continúa con una sección a manera de conclusiones y se proponen trabajos futuros, algunos de ellos ya en marcha.

II. Búsquedas de OA en repositorios. Similitud de resultados vs. lo esperado

Con el fin de evaluar someramente la concordancia de los Objetos de Aprendizaje recuperados con lo esperado, se realizaron búsquedas genéricas a partir de palabras relevantes en la competencia. El alcance de esta parte del trabajo está limitado a encontrar la similitud de un conjunto de palabras en el marco del contexto en que se presentan, en concreto en la definición de los objetivos educativos y de la información asociada a los materiales educativos en los repositorios.

Para obtener los resultados de similitud para un caso de estudio en auditoría de sistemas se realizaron consultas en diferentes repositorios. La desambiguación de los sentidos de las palabras presenta gran interés y es una preocupación desde los primeros días de tratamiento informático de la lengua en la década de 1950[9], además en la comparación no se incluyeron ”palabras irrelevantes”[10] y se realizó desde cadenas en el microcontexto de la competencia versus el microcontexto disponible de los OA en los repositorios. Se operó sobre los metadatos los OA, y en el bloque de definición de competencias la sección asociada a competence_knowledge. El detalle de la aplicación se muestra en la sección 3. Se realizó una búsqueda simple para Auditoría de Sistemas y se tomaron los primeros resultados de los diferentes repositorios y se aplicaron las métricas de similitud para determinar que tanto los OA están relacionados con las palabras definidas en las competencias y que representan los logros esperados. Se realizaron las búsquedas en Merlot (www.merlot.org), Ariadne (www.ariadne-eu.org/) The Gateway (www.thegateway.org/) y Smete (www.smete.org) y Careo (http://theguide.ntic.org/search.php). La tabla 1 y la figura 1 recogen los resultados del ejercicio.

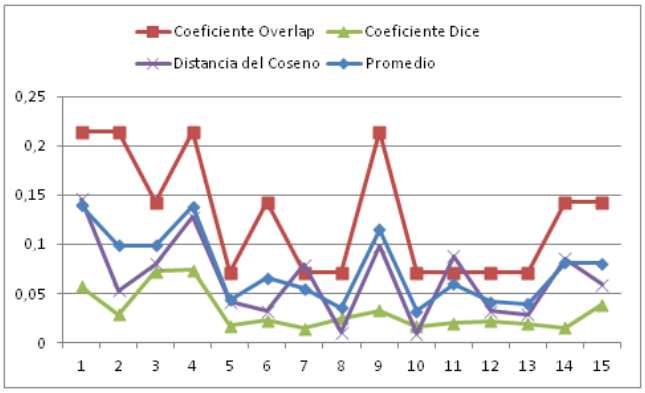

Fig. 1. Similitud en la recuperación genérica de OA

III. Recuperación de OA y métricas para ponderación de resultados. Medidas de similitud

El diagrama, en la figura 2 muestra las fases implementadas para obtener los resultados objeto de análisis en este artículo. Los pasos se describen brevemente a continuación:

1) **Obtener el archivo XML del contexto de los OA.**
En los repositorios locales propios se ha implementado la funcionalidad de entrega de resultados de búsquedas en archivos XML. Para los repositorios remotos, se obtiene el archivo XML del contexto por medio de web services.

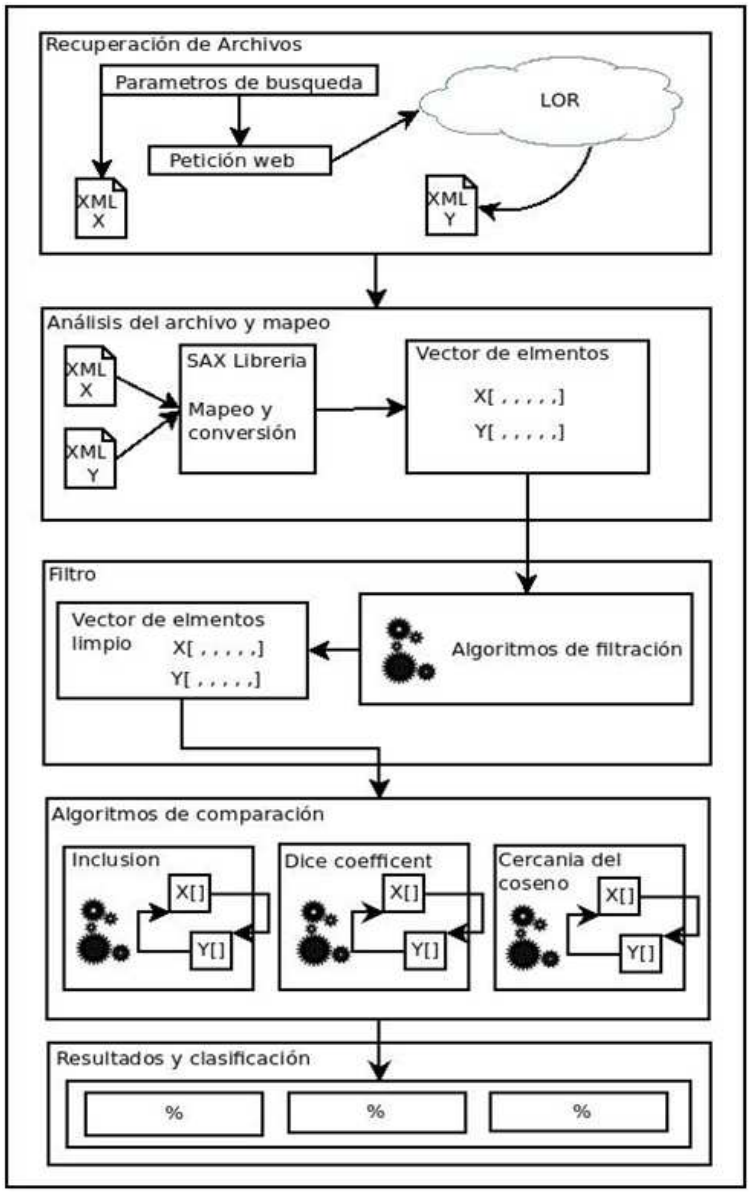

Fig. 2. Fases en el proceso de ponderación de OA por similitud

2) **Análisis y conversión del archivo XML.**
Para comparar estos resultados con las competencias los términos especificados en el archivo de competencias se realiza un pre-proceso para llevarlo a una estructura vectorial sobre la cual se aplicaran los algoritmos. Por medio de la librería SAX, se hace el mapeo de la estructura del archivo XML y se realizan las manipulaciones requeridas.

3) **Filtro del vector.**
Uno de los problemas encontrados es la existencia de muchos espacios en blanco, palabras irrelevantes y

OA	Coeficiente Overlap	No. características Coeficiente Dice	Distancia del coseno	Promedio
1	0, 21428572	0, 05714286	0, 14569113	0, 1390399
2	0, 21428572	0, 02857143	0, 0533114	0, 09872285
3	0, 14285715	0, 07272727	0, 0798723	0, 09848557
4	0, 21428572	0, 07407408	0, 1276668	0, 13867553
5	0, 07142858	0, 01709402	0, 04186446	0, 04346235
6	0, 14285715	0, 02259887	0, 03223861	0, 065898210
7	0, 07142858	0, 01492537	0, 07912107	0, 055158345
8	0, 07142858	0, 02439024	0, 01074338	0, 03552073
9	0, 21428572	0, 0326087	0, 0978232	0, 11490587
10	0, 07142858	0, 01639344	0, 00914014	0, 03232072
11	0, 07142858	0, 02020202	0, 08812103	0, 05991721
12	0, 07142858	0, 02222222	0, 0319167	0, 04185583
13	0, 07142858	0, 01904762	0, 02867697	0, 03971772
14	0, 14285715	0, 01532567	0, 08512566	0, 08110283
15	0, 14285715	0, 03883495	0, 05923489	0, 080309

Tabla I
MÉTRICAS DE SIMILITUD EN LA RECUPERACIÓN GENÉRICA DE OA

signos de puntuación que no aportan a los resultados de la comparación y la hacen poco efectiva. Para resolver esto se implementó un método de limpieza que retira estas partes del vector de palabras. Para las palabras irrelevantes se creó un vector que las incluye, por lo cual se pueden definir según el caso particular. El resultado es un vector "limpio" listo para ser utilizado con los algoritmos de comparación.

4) **Algoritmos de comparación.**
Se definió la aplicación de tres algoritmos de comparación.

A. Coeficiente Overlap

el cual itera sobre el vector del contexto del repositorio buscando las palabras y calcula el porcentaje así:

$$Coeficiente\_Overlap = \left| \frac{(X \cap Y)}{\min(X, Y)} \right| \qquad (1)$$

Donde el numerador es el número de palabras comunes obtenidas desde los dos XML que representan la competencia definida y el microcontexto de los metadatos de los OA respectivamente, sin incluir espacios ni palabras no relevantes. El denominador es el mínimo número de palabras entre los dos arreglos.

B. Coeficiente de Dice

La aplicación original de este algoritmo toma el vector como una cadena de caracteres y forma pares con los mismos desde la posición x y su adyacente inmediato, como se aprecia en siguiente ejemplo, donde para comparar las cadenas 'objetos con 'objatos', se crean los siguientes vectores de caracteres:

X=[ob,bj,je,et,to,os], Y=[ob,bj,ja,at,to,os]

Para la finalidad de este trabajo esto no es práctico, por lo cual se modificó su funcionamiento, aplicándolo sobre palabras completas.
Finalmente se calcula la cercanía así:

$$Coeficiente\_Dice = \frac{2(X \cap Y)}{|X| + |Y|} \qquad (2)$$

Donde (XnY) se obtiene del número de palabras comunes en los dos vectores. y $|X|$ e $|Y|$ son los números de palabras respectivamente en cada vector, luego de quitar los espacios y las palabras no relevantes.

C. Distancia del Coseno

Realmente más que una distancia es una medida de similitud y usualmente este cálculo se ha utilizado con vectores cuyos elementos son valores numéricos, pues se requieren operaciones matemáticas sobre tales elementos. En este proyecto extendimos esta aplicación a datos categóricos y el cálculo se realiza así:

$$Distancia\_Coseno = \sum_{1}^{n} (P_i * Q_i) \Big/ \sqrt{\left(\sum_{1}^{n} P_i^2 * \sum_{1}^{n} Q_i^2 \right)} \qquad (3)$$

Donde:

Pi: frecuencia Termino i en Vector1
Qi: frecuencia Termino i en Vector2

Para hacer el conteo de las ocurrencias de cada palabra fue necesario, previamente, dejar una única ocurrencia por cada palabra eliminando las repetidas y recorrer los dos arreglos.

IV. PRUEBAS PARA SIMILITUD ENTRE MICROCONTEXTOS.

La propuesta se enmarca en un ambiente de microcontextos controlados, aprovechando que se cuenta con repositorios propios y con políticas de cosechado que permiten almacenar la información que apoye el proceso de selección. En concreto se partió de las competencias definidas para un curso de Auditoría de Sistemas, en el tema Tecnicas de Auditoría Asistidas por Computador TAAC, especificadas en formato IMS RDceo y con granularidad fina lo que se orienta a logros

Competencia	OA	Datos de prueba	Minicompañía	Comparación de códigos	Simulación paralela	Software general
Datos de Prueba	Promedio	0, 6312209	0, 363394073	0, 313579387	0, 35217921	0, 35285242
Minicompañia	Promedio	0, 42786259	0, 504325767	0, 33202123	0, 46956542	0, 4445411
Comparación de Códigos	Promedio	0, 39912005	0, 36136341	0, 44813319	0, 35781707	0, 35846035
Simulación Paralela	Promedio	0, 51260958	0, 36676381	0, 31645384	0, 5506436	0, 34977447
Software General	Promedio	0, 39141298	0, 378682503	0, 314201173	0, 36141377	0, 58201545

Tabla II
MÉTRICAS DE SIMILITUD ENTRE LOS MICROCONTEXTOS DE COMPETENCIAS Y DE AO.

específicos. Los OA disponibles son recursos atómicos de diferente tipo: Ejercicios, simulaciones, diagramas, figuras, html, slides, tablas, textos, experimentos, problemas, cuestionarios, etc. Cada OA cuenta con metadatos LOM Las pruebas buscaron demostrar que la propuesta presentada en este artículo es válida en la localización de OAs acordes con los requerimientos de cada competencia, aprovechando las medidas de similitud entre el vector que representa la competencia y el vector de metadatos del OA. Aplicando los pasos mostrados en la sección anterior se obtuvieron los siguientes resultados. La tabla 2 y la figura 3 recogen los cálculos para las diferentes métricas aplicadas a cada microcontexto de la competencias vs los microcontextos de los metadatos de los diferentes OA. El promedio obtenido es un calculo simple de la suma de las métricas dividido el numero de las mismas.

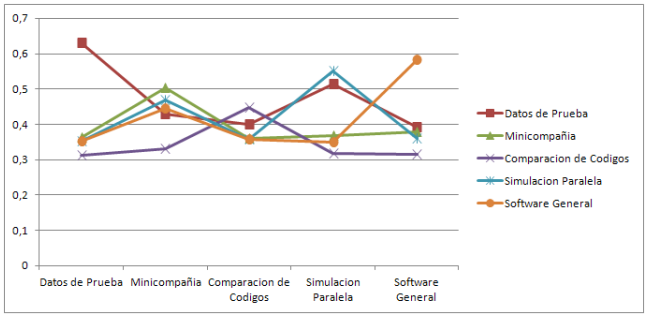

Fig. 3. MÉTRICAS DE SIMILITUD ENTRE LOS MICROCONTEXTOS DE COMPETENCIAS Y DE AO

V. CONCLUSIONES Y TRABAJO FUTURO.

Este trabajo presenta una buena alternativa para la selección o recomendación ponderada de Objetos de Aprendizaje obtenidos en búsquedas simples, mediante la aplicación de algoritmos de similitud entre los metadatos de los OA y los términos en la competencia.

La experiencia demostró que la precisión en la definición de los metadatos y en los componentes de las competencias determina las posibilidades de obtener OA relevantes para el contexto de las competencias que se espera que el estudiante logre. Se puede concluir que se requiere refinar los microcontextos de los metadatos de los objetos de aprendizaje en los repositorios, para que puedan ser aprovechados directamente en el apoyo a cursos. También queda expuesto que las competencias deben ser definidas con mayor precisión para evitar tantas palabras irrelevantes en el contexto de estudio que puedan inducir a recuperación también irrelevante. Actualmente se hacen pruebas con otros algoritmos de similitud y se mejoran los procesos de filtrado, además se ha logrado mejoras iniciales utilizando morfema de base (radical o raíz de la palabra). Con el fin de generalizar la propuesta se harán pruebas con otros repositorios, que permitan la recuperación del micro-contexto en formato XML.

Aunque el análisis de reducción de dimensión mostró reaultados adecuados se debe tener cuidado en la escogencia de los niveles de significancia al momento de realizar la selección de características: Este es el caso de FTC en el cual se observa que los resultados de desempeño obtenidos en la clasificación disminuyen apreaciablemente según el analisis realizado.

REFERENCIAS

[1] Duque, Néstor D., Jiménez, C., Guzmán, J. IA Planning for Automatic Generation of Customized Virtual Courses. 16th European Conference on Artificial Intelligence (ECAI'04). Planning and Scheduling of the series Frontiers in Artificial Intelligence, IOS Press. España. 2004.

[2] Li, J. Z. "Quality , Evaluation and Recommendation for Learning Object," International Conference on Educational and Information Technology, no. Iceit, pp. 533-537, 2010

[3] Rodríguez, Paula; Isaza, Gustavo, Duque, Néstor D. Búsqueda personalizada en Repositorios de Objetos de Aprendizaje a partir del perfil del estudiante. Cava 2011. Bogotá.

[4] Baldiris, S., Graf, S., Fabregat, R., Duque, N. D. Looking for Contextualized Learning Objects to support Semi-automatic Learning Design Generation. Workshop on Technology-Enhanced Learning and Living (TELL 2012). Edmonton, AB, Canadá, February 14-16, 2012.

[5] Baldiris, S., Graf, S., Fabregat, S., Duque Méndez, ND. Searching and Positioning of Contextualized Learning Objects. International Review of Research in Open and Distance Learning. In press. Canadá. 2012.

[6] Chapman, Sam. String Similarity Metrics for Information Integration. Natural Language Processing Group. 2008

[7] Amón, Iván; Jiménez, Claudia. Funciones de Similitud sobre Cadenas de Texto: Una Comparación Basada en la Naturaleza de los Datos. Medellín

[8] UK Sheffield University. String Similarity Metrics for Information Integration. United Kingdom. Disponible en: http://sourceforge.net/projects/simmetrics/

[9] Ide, N. y Véronis, J. Word Sense Disambiguation : The State of the Art. New York, 1997

[10] Castillo, JL., Fernández, J., González, L. Agrupamiento de documentos con sistemas evolutivos. Conferência IADIS Ibero-Americana WWW/Internet 2008