PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

METADATA RELEVANTE PARA EDUCACIÓN BÁSICA Y MEDIA EN CHILE

JORGE SALVADOR BOZO PARRAGUEZ

Tesis para optar al grado de
Doctor en Ciencias de la Ingeniería

Profesor Supervisor:
ROSA ALARCÓN

Santiago de Chile, Septiembre de 2016

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

METADATA RELEVANTE PARA EDUCACIÓN BÁSICA Y MEDIA EN CHILE

JORGE SALVADOR BOZO PARRAGUEZ

Tesis presentada a la comisión integrada por los profesores:
ROSA ALARCÓN
JAIME NAVÓN
IGNACIO CASAS
SERGIO OCHOA
JOSÉ LUIS SIERRA
CRISTIAN VIAL EDWARDS

Para completar las exigencias del grado de
Doctor en Ciencias de la Ingeniería

Santiago de Chile, Septiembre de 2016

A Paula Alejandra y mi familia, que
han sido un apoyo incondicional

AGRADECIMIENTOS

El trabajo de una tesis doctoral es fruto del esfuerzo, dedicación y perseverancia del candidato a doctor. A lo anterior se debe agregar el apoyo de las personas del entorno del candidato. Después de varios años en el programa de doctorado, debo expresar mis más sinceros agradecimientos a estas personas, acá algunas palabras para ellos.

A mi supervisora Rosa Alarcón Choque, quien dedicó tiempo y trabajo para que esta tesis concluyera exitosamente. Su apoyo fue constante, ella confió en mi hasta el final, y en los momentos críticos supo orientarme para seguir adelante.

A mis compañeros y amigos del programa, en especial a Jorge Pérez, quien desde que llegué al DCC hicimos juntas, compartimos experiencias, y estuvo en los buenos y malos momentos que viví. Destaco de Jorge la incondicionalidad y desinterés en su ayuda, es sin lugar a dudas un gran amigo. A Jesús Bellido, de quien recibí siempre palabras de apoyo, trabajamos juntos y mantenemos también una gran amistad. A Carla, Martín y Alejandro, todos compañeros de oficina también fueron soporte.

A las personas del DCC, Soledad, Jaime, Juan Pablo, y especialmente a Yadran, quien me ayudo contratándome como su asistente en los cursos del programa de magister del DCC, trabajamos y compartimos por mucho tiempo y fraguamos una gran amistad que valoro mucho.

A las personas de la Dirección de Postgrado de la Escuela de Ingeniería, muy especialmente a Debbie, Danisa y María de Lourdes, quienes siempre muy gentilmente atendieron todas mis inquietudes respecto del programa.

Al programa de becas para doctorado de CONICYT y Becas Chile, por el apoyo financiero de mis estudios y pasantía en el extranjero.

A Paula Alejandra, quien durante estos últimos años a estado a mi lado, dando su apoyo siempre incondicional.

INDICE GENERAL

INDICE DE FIGURAS

INDICE DE TABLAS

RESUMEN

Los *sistemas de recomendación* (SR) constituyen hoy una herramienta útil cuando se quiere buscar u ofrecer algún producto o servicio en Internet. *Amazon* y *Netflix* son buenos ejemplos de plataforma que utilizan fuertemente *sistemas de recomendación*, para ayudar a los usuarios en la búsqueda y selección de los productos y servicios que ofrecen.

En el ámbito de la educación, hay esfuerzos por poner al servicio de la comunidad plataformas con repositorios de material educativo, complementarios a los materiales tradicionales, como libros de texto. Esto con el objetivo que sean de apoyo para los estudiantes, en especial a nivel de primaria y secundaria. En Chile, el Ministerio de Educación, a través de los portales *Educar Chile* y *Catalogo Red*, ha puesto a disposición de la comunidad educativa una gran cantidad de *recursos digitales* tales como libros, presentaciones, fotos, vídeos, y material multimedia en general, para que los profesores y alumnos los utilicen en clases, o bien los usen como apoyo a las clases tradicionales.

A pesar de este esfuerzo, la búsqueda de material educativo apropiado es tediosa debido a la gran variedad y cantidad de recursos disponibles en Internet. Es difícil para un profesor navegar e inspeccionar decenas de recursos (algunos de ellos sitios Web, otros, links a otros repositorios, etc.), para encontrar el recurso más apropiado.

En esta tesis se busca diseñar un SR en un contexto educativo e identificar elementos claves para la organización de la información, es decir, *metadata*, sobre la cual opera el SR. Este trabajo busca contribuir en el diseño de software dedicado a profesores de educación básica y media. El público objetivo de este estudio se centra en Chile aunque creemos que los alcances de esta tesis se pueden aplicar a otros sistemas de educación primaria y secundaria en Latinoamérica.

Palabras Claves: sistema de recomendación, filtrado colaborativo, objetos de aprendizaje, clustering jerárquico, metadada

ABSTRACT

Recommender systems today are a useful tool when you want to search for or offer a product or service on the Internet. *Amazon* and *Netflix* are good examples of platforms that heavily use recommender systems to help users in the search and selection of the products and services they offer.

In the field of education, there are efforts to contribute to the community, platforms including educational repositories that are complementary to traditional materials, such as textbooks. The main objective is to provide support for students, especially at primary and high school. In Chile, the Ministry of Education, through the *Educar Chile* and *Catalogo Red* portals, has made available to the educational community various *learning resources*, such as digital books, presentations, photos, videos, and multimedia materials in general, that can be used by teachers and students in class or as a support for traditional classes.

Despite this effort, the search for appropriate educational material is tedious because of the variety and amount of resources available on the Internet. It is difficult for a teacher to browse and inspect dozens of resources (some websites, others, links to other repositories, etc.) and to find the most appropriate resource.

This thesis aims to design a recommender system (RS) in an educational context and identify key elements for the organization of information, i.e. *metadata*, on which the RS operates. This work aims to contribute to the design of software dedicated to teachers of primary and secondary education. The principal audience of this study is the Chilean educative system although we believe that the scope of this thesis can be applied to other systems of primary and secondary education in Latin America.

Keywords: Recommender Systems, collaborative filtering, learning objects, hierarchical clustering

1. INTRODUCCIÓN

Históricamente los materiales de estudios han sido: *libros de texto*, libros, enciclopedias, diarios y revistas. Hoy, esto ha cambiado pues la era digital ha permitido contar con una gran cantidad de nuevos materiales, en una infinidad de formatos. A los clásicos materiales mencionados, se suman nuevos formatos digitales como: presentaciones, hojas de cálculo, audio, video, multimedia, aplicaciones de software, etc.

Los *recursos de aprendizaje* son documentos digitales usados en *E-Learning* (Lehmann, Hildebrandt, Rensing, y Steinmetz, 2008); esta definición incluye recursos multimedia, recursos con hipertexto, cursos completos, o sitios Web. Los recursos de aprendizaje pueden estar referidos a recursos individuales o a muchos documentos digitales escritos en diferentes formatos. Por otra parte, los *Objetos de Aprendizajes* (o LOs, por su sigla en inglés), son pequeñas piezas o bloques de construcción, que pueden ser compartidas, reusadas, y combinadas en bloques de instrucciones mayores (Koper, 2003; Motelet, 2007). Los recursos de aprendizaje y LOs están descritos con metadatos que proveen información adicional para que puedan ser descubiertos desde diversas perspectivas. Los metadatos pueden ser estandarizados, siendo LOM uno de los estándares de mayor peso y el mas usado para LOs (Ochoa, Klerkx, Vandeputte, y Duval, 2011). También se puede seguir una estrategia de anotación más liviana y abierta, donde a los recursos se les asocian etiquetas informales, generando metadata no estandarizada, y taxonomías basadas en el usuario (Manouselis, Drachsler, Vuorikari, Hummel, y Koper, 2011).

Los recursos de aprendizaje se pueden proporcionar como colecciones organizadas administradas por comunidades de profesores, con o sin metadatos (Tiropanis, Davis, Millard, y Weal, 2009); o través de repositorios especializados que albergan varios ítems que van desde cientos a millones (Ochoa y Duval, 2009). Tales repositorios pueden contener recursos propios y sus metadatos asociados, o sólo los metadatos y una referencia a la ubicación real de los recursos, o a la localización del sitio Web que los aloja, lo que requiere una búsqueda adicional. La mayoría de los repositorios soportan búsquedas por palabras clave (sobre metadatos y contenido) principalmente, lo que resulta en una extensa lista de

recursos como respuesta a las búsquedas. De estos resultados los usuarios deben después buscar en estos resultados para encontrar el recurso que mejor satisfaga sus necesidades. La Web, esta plagada de estos materiales, y tanto profesores como alumnos recurren a Internet, para encontrar los mejores recursos, que permitan complementar y mejorar la enseñanza y aprendizaje de estudiantes. Sin embargo, ante la gran cantidad de oferta de material de estudio disponible, a los profesores se les hace difícil *buscar* material relevante para sus objetivos y *seleccionar* entre ellos cuáles tienen la mejor calidad, pierden tiempo en encontrar materiales apropiados, por ejemplo, usados y validados por otros profesores.

Algunos investigadores utilizan los SR con el fin de facilitar la búsqueda de recursos de aprendizaje. Sin embargo, generalmente se centran en los estudiantes como el principal consumidor de recursos (Manouselis, Drachsler, Verbert, y Duval, 2013); las necesidades y las prácticas seguidas por los profesores, en particular de enseñanza básica y media, a menudo se ignoran. Herramientas pedagógicas tales como los planes de estudio, las planificaciones, rúbricas, etc., corresponden a una terminología comúnmente utilizadas por profesores e instructores en estos niveles, pero están notablemente ausentes en ambos enfoques de metadata estandarizada y ligera (Tiropanis et al., 2009). Los SR en educación difieren de otros campos como películas, series o canciones (recomendadores de productos o servicios) debido a los objetivos del usuario (Verbert et al., 2012). Por ejemplo, los profesores podrían preparar material nuevo para una clase que es parte de la planificación de la asignatura, o seleccionar recursos que apoyen la búsqueda de información por parte de estudiantes, o motivar a la audiencia, o recordar conocimientos existentes, o elegir material simple para introducir un concepto, etc. (Manouselis et al., 2011). Los profesores pueden tener distintos intereses en función de su especialidad y las actividades administrativas exigidas por su trabajo.

Este trabajo de tesis, se centra en facilitar la búsqueda de materiales de estudio, por parte de profesores de enseñanza primaria y secundaria. Con el objetivo de acotar el alcance de esta investigación, este trabajo se limita a los recursos provistos por el portal *EducarChile*[1].

[1] http://www.educarchile.cl

Al respecto, la *red enlaces*[2] desde su creación a tenido varios hitos importantes relacionados con este trabajo de tesis, destacando dos. En el año 2001 se crea el portal *EducarChile*, respondiendo a la necesidad de contar con una plataforma Web educativa nacional, orientada a los actores involucrados en el mundo de la educación primaria y secundaria, esto es, docentes, estudiantes, familia e investigadores. En el año 2009, cuando se crea el *Catalogo en Red*[3], donde se pone a disposición de la comunidad educativa recursos digitales para apoyar la implementación de sus planes educativos (Valdebenito y Cruzat, 2012).

1.1. Motivación del problema

Es indudable que hay una proliferación del material educativo digital en la Web. Cuando un profesor busca materiales en Internet, de seguro requiere que su búsqueda sea precisa, pertinente, rápida, y que incluya recursos de calidad. Para esto existen varias alternativas para tratar de satisfacer esta tarea, por ejemplo:

a.- Realizar una búsqueda abierta en motores génericos tales como *Google* o *Yahoo*. En este caso, dependiendo de los términos de búsqueda, un profesor obtendrá los resultados más populares, los que no están necesariamente relacionados con las metas pedagógicas asociadas a su búsqueda. También puede ocurrir que los términos de su búsqueda den lugar a problemas de diferencias semánticas (ej. Jaguar, el auto, en lugar del animal) con lo que obtendrá resultados irrelevantes.

b.- También se podría realizar una búsqueda en un ámbito académico, es decir buscar usando motores tales como Google Scholar. Si bien éste es un buscador especializado en artículos de revistas científicas, filtrando una gran cantidad de material dedicado a entornos diferentes a los educativos, está enfocado al mundo académico relacionado con temas de investigación y alejado de materiales de estudio para la educación primaria y secundaria.

[2]http://www.enlaces.cl/
[3]http://www.catalogored.cl

c.- Es posible hacer búsquedas en repositorios especializados de educación primaria y secundaria, que almacenan recursos digitales abiertos (o no) tales como MERLOT[4]. En estos casos la facilidad con que un profesor encuentre los recursos que busca dependerá de la calidad de la interfaz del repositorio y su motor de búsqueda interno. En general, estos repositorios utilizan motores de búsqueda genéricos, tales como Google, o motores especializados que explotan alguna clasificación del material que contienen o permiten la búsqueda de material de acuerdo a algunos metadatos con los que se anotan a los recursos.

d.- Un profesor podría realizar búsquedas en sitios o portales Web cercanos a su contexto educativo, como el repositorio de contenido de *EducarChile* para profesores de enseñanza básica y media en Chile. La búsqueda en este portal, se inicia con la selección del *centro de recursos* en el menú del portal, continua con la selección de alguna opción en el centro de recursos tales como **Todos**, o **Interactivos**. Recién en un tercer paso podrá hacer una búsqueda basada en palabras, o bien seguir seleccionado por temas, hasta que en algún momento se encuentre un material de su interés. Este proceso de búsqueda es tedioso, poco amistoso y no es intuitivo.

e.- Por otra parte, cuando se busca un libro, u otro producto en tiendas online tales como *Amazon*[5], después de algunas interacciones con el sistema, por búsqueda o compra, *Amazon* sugiere productos/servicios similares, lo que puede facilitar el descubrimiento de material relevante para un profesor sin realizar esfuerzos adicionales. Es claro que un sistema que tiene una retroalimentación de preferencias propias y de usuarios con gustos similares facilita la búsqueda/selección de material relevante, sin embargo, *Amazon* está dedicado a recursos digitales de grano largo, tales como un libro que contiene a su vez muchos otros recursos que el profesor debe comprar para luego reusar partes de él.

[4]https://www.merlot.org/merlot/advSearchMaterials.htm
[5]http://www.amazon.com/

Este trabajo de tesis aborda los problemas asociados a la búsqueda en repositorios específicos de material educativo digital para educación primaria y secundaria, particularmente el portal *EducarChile*, aplicando las técnicas asociadas a sistemas tales como *Amazon*, es decir *sistemas de recomendación*. A diferencia de los recomendadores destinados a productos dirigidos a todo público (ej. libros, películas, etc.), se busca un recomendador que considere las características de los profesores que hacen la búsqueda (contexto).

1.2. Objetivos

Los objetivos planteados para esta tesis se dividen en un objetivo general y cuatro objetivos específicos, que se enumeran a continuación.

1.2.1. Objetivo General

Facilitar la búsqueda de recursos de aprendizaje para profesores de enseñanza básica y media en Chile.

1.2.2. Objetivos Específicos

- O1: Identificar un conjunto de metadatos para recursos de aprendizaje que sean relevantes en el proceso de descubrimiento de recursos (bśquedas).
- O2: Definir un conjunto de metadatos que caractericen a los profesores que realizan las búsquedas.
- O3: Definir una estrategia para identificar comunidades de interés (contexto de búsqueda) en base a la metadata definida.
- O4: Definir una estrategia para recomendar recursos a usuarios automáticamente que tome en cuenta la metadata de los recursos y el contexto de búsqueda.
- O5: Validar tanto la metadata identificada y la estrategia de recomendación en una tarea avanzada de los sistemas de recomendación: la recomendación de recursos nuevos a usuarios nuevos (*cold start problem*).

1.3. Hipótesis

Para este tesis se han plateado las siguientes hipótesis de trabajo:

H1: Los metadatos curriculares (sector/asignatura, nivel) influencian positivamente el descubrimiento de recursos de aprendizaje por parte de profesores.

H2: Existen metadatos que son importantes para los profesores de educación de media y básica del sistema escolar Chileno, que no están contemplados como metadata estándar.

H3: Se puede prescindir de la historia de consumo de un usuario, para recomendar con alta precisión, recursos novedosos (no antes vistos) a un profesor.

1.4. Preguntas de investigación

De las hipótesis planteadas se desprenden las siguientes preguntas de investigación:

R1: ¿Qué información deben considerar los sistemas de recomendación en el sistema de educación de media y básica del sistema escolar Chileno?

R2: ¿Cuál es el impacto de considerar el contexto de la búsqueda y metadatos de los recursos digitales en la calidad de la recomendación?

R3: ¿Qué características del contexto del profesor son relevantes para la búsqueda de recursos educativos?

R4: ¿Cómo se debe representar esta información?

R5: ¿Cómo se puede obtener información (para la búsqueda) sin imponer sobrecarga de trabajo al profesor?

1.5. Organización del Documento de Tesis

Este documento de tesis tiene la siguiente estructura: el Capitulo 2 (**Marco teórico**), presenta el marco teórico que incluye los aspectos relevantes de la teoría de *sistemas de recomendación* (técnicas y clasificación), evaluación de sistemas, *estándares* de metadatos, y metadatos en educación; en el Capitulo 3 (**Encontrando categorías de metadata a**

través de recomendadores) se presentan las actividades de investigación en cuatro Fases que consisten en la identificación las comunidades de profesores que consumen recursos de aprendizaje (Fase 1), caracterización de profesores (Fase 2), validación de la metadata encontrada para solucionar un problema complejo de los sistemas de recomendación como es el arranque en frío (Fase 3), y un análisis de la metadata relevante (Fase 4). En el capitulo 4 (**Propuesta de Metadata para educación**) se analizan los principales estándares para metadatos y se propone como deben ser utilizados para incluir la metadata relevante descubierta en esta investigación. Finalmente, en el Capitulo 5 (**Conclusiones**) se presentan las principales conclusiones y se discute su relación con las preguntas de investigación.

2. MARCO TEÓRICO

El marco teórico de este trabajo esta dividido en dos secciones: *sistemas de recomendación*, *metadatos* y *metadatos en educación*. La sección de *sistemas de recomendación* muestras los aspectos teóricos centrales y notación de *sistemas de recomendación*, tanto en tipos, usos, y evaluación. La sección de *metadatos* muestra los aspectos teóricos y prácticos de *metadatos*, que termina con un análisis en metadatos para educación.

2.1. Sistemas de recomendación

Los sistemas de recomendación son hoy una herramienta útil, tanto para los usuarios que buscan o requieren productos/servicios, como para los oferentes de productos/servicios. En la Web, *Amazon* es un ejemplo de un sistema que desde sus inicios hace un uso eficiente de esta tecnología para ofrecer al público una variedad de productos/servicios en los que pueden estar interesados. Otros ejemplos de sistemas, que hacen uso de recomendadores son sistemas de música online, tales como *LastFM*[1] o *Spotify*[2], o sistemas para ver películas y series de TV como *Netflix*[3]. Los sistemas de recomendación que hay detrás de estas plataformas son aplicaciones que recogen información de usuarios, sus valoraciones sobre recursos, e información de los productos/servicios para hacer una recomendación.

Tradicionalmente los sistemas de recomendación se han catalogado de acuerdo al sistema de filtrado que utilizan en (Popescul, Ungar, Pennock, y Lawrence, 2001; Lika, Kolomvatsos, y Hadjiefthymiades, 2014a):

- Sistemas de recomendación por *contenido*
- Sistemas de recomendación *colaborativos*
- Sistemas de recomendación *híbridos*

Independiente del enfoque, los algoritmos comparten conceptos y notación que es usada en este tesis y se presenta a continuación:

[1] http://www.lastfm.es/
[2] http://www.spotify.com
[3] http://www.netflix.com

2.2. Conceptos fundamentales y notación

a) *Items*: Conjunto finito de recursos digitales de aprendizaje (*Learning Objects*), que denotaremos por $\mathscr{I}$, así:

$$\mathscr{I} = \{i_1, i_2, ..., i_m\}$$

En otros contextos un ítem puede ser cualquier tipo de recurso (o servicio) que se quiera ofrecer a una comunidad de posibles interesados.

b) *Usuarios*: Conjunto de personas profesores, directivos, o miembros de la comunidad educativa en general, que hacen uso y/o buscan recursos educativos en la Web. En esta investigación nos limitaremos a los profesores.

$$\mathscr{U} = \{u_1, u_2, \cdots, u_n\}$$

c) *Evaluación*: Es la calificación (o rating) que un usuario $u \in \mathscr{U}$ asigna a un ítem $i \in \mathscr{I}$, se denota por $r_{u,i}$ y se presenta en una escala, que puede ser expresada, por ejemplo, en valores discretos de 1 a 5, donde el valor 1 indique mínimo grado de satisfacción y 5 máximo grado de satisfacción. También se pueden usar rating implícitos que se derivan de otros datos, por ejemplo, se puede inferir la satisfacción de un usuario a partir del registro de su visita (o descarga) a algún ítem (J. L. Herlocker, Konstan, Terveen, y Riedl, 2004).

d) *Perfil de preferencias*: Para un usuario u, la expresión $\mathscr{I}_u$ denota el conjunto de ítems que ha calificado (o visitado). Dependiendo de la técnica de recomendación a utilizar, el *perfil de preferencias* de u, que denotaremos por PP_u, se asocia con el conjunto $\mathscr{I}_u$ o bien con el conjunto de pares *ítem-calificación* que define $\mathscr{I}_u$. Así, $PP_u = \mathscr{I}_u$ o bien $PP_u = \{\langle i, r_{u,i}\rangle \mid i \in \mathscr{I}_u\}$. Cuando un usuario u no a evaluado un ítem i, denotamos $r_{u,i} = \cdot$, al estilo de la notación usada por Ortega (Ortega, Bobadilla, Hernando, y Rodríguez, 2014).

e) *Matriz de preferencias*: Corresponde a la matriz *usuarios-ítems* con las preferencias (calificaciones o visitas) de los usuarios sobre los ítems $i \in \mathscr{I}$, tabla 2.1. Por lo general

la matriz de preferencias es escasa, con muchos valores para $r_{u,i}$ no determinados, esto es, con $r_{u,i} = \cdot$.

TABLA 2.1. La matriz de preferencias resume los valores de rating $r_{u,i}$ para $u \in \mathscr{U}$ e $i \in \mathscr{I}$.

	i_1	i_2	...	i_m
u_1	r_{u_1,i_1}	r_{u_1,i_2}	$\cdots$	r_{u_1,i_m}
u_2	r_{u_2,i_1}	r_{u_2,i_2}	$\cdots$	r_{u_2,i_m}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
u_n	r_{u_n,i_1}	r_{u_n,i_2}	$\cdots$	r_{u_n,i_m}

f) Similitud de usuarios (o ítems): Es la medida que establece que tan parecidos en preferencias son dos usuarios u y v. Esta medida se denota por $sim(u, v)$. El valor similitud entre usuarios dependerá de la métrica elegida y las calificaciones que han dado a ítems tanto u como v.

g) Matriz de similitud de usuarios: Corresponde a la matriz S con los valores de similitud entre usuarios, es decir, $S = [sim(u, v)]_{|\mathscr{U}|\times|\mathscr{U}|}$. Notar que S es una matriz simétrica, esto es, $S = S^T$, y su diagonal es la identidad. Además, los usuarios que no tienen evaluaciones tendrán similitud cero. La tabla 2.2 ilustra la estructura de la matriz de similitud S.

TABLA 2.2. La matriz de similitud S entre usuarios de $\mathscr{U}$, es una matriz simétrica.

	u_1	u_2	...	u_n
u_1	1	$sim(u_1, u_2)$	$\cdots$	$sim(u_1, u_n)$
u_2	$sim(u_2, u_1)$	1	$\cdots$	$sim(u_2, u_n)$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
u_n	$sim(u_n, u_1)$	$sim(u_n, u_2)$	$\cdots$	1

h) Vecindad de usuarios: Para un usuario u, su vecindad, que denotamos por V_u, corresponde al conjunto de usuarios de $\mathscr{U}$ que tienen similitud con él.

i) *Predicción*: Los sistemas de recomendación basados en filtrado colaborativo, predicen la calificación $r_{u,i}$ que un usuario u asignará a un ítem i, este valor de predicción se denota por $P_{u,i}$.

j) Para un ítem $i \in \mathscr{I}$, el conjunto $\mathscr{U}_i$ denotará el conjunto de usuarios que ha calificado el ítem i, esto es, $\mathscr{U}_i = \{u \in \mathscr{U} \mid r_{u,i} \neq \cdot\}$.

k) $\overline{r}_u$: Valor promedio de las calificaciones asignadas por un usuario $u \in \mathscr{U}$.

l) $\overline{r}_i$: Valor promedio de las calificaciones recibidas por un ítem $i \in \mathscr{I}$.

2.3. Métricas de similitud

Los métodos de recomendación requieren métricas que permitan comparar usuarios (o ítems dependiendo de la técnica), en términos de similitud de preferencias, o bien similitud de contenido para ítems. En la ecuación de predicción de rating $P_{u,i}$ (ecuación 2.6) un elemento central de la fórmula es el coeficiente de similitud $w_{u,v}$, que indica que tan parecidos son los perfiles de preferencias de los usuarios u y v. Esta componente de la ecuación , se determina en base a las evaluaciones en común de los usuarios u y v.

Sea C el conjunto de ítems que tienen evaluaciones de u y v, esto es, $C = \mathscr{I}_u \cap \mathscr{I}_v$, y sean r_u y r_v los vectores con las evaluaciones dadas por u y v a los ítems de C. A continuación se definen las métricas más utilizadas para similitud entre usuarios.

1. Coeficiente de correlación de *Pearson*:

$$w_{u,v} = \frac{\sum_{i \in C}[(r_{u,i} - \overline{r}_u)(r_{v,i} - \overline{r}_v)]}{\sqrt{\sum_{i \in C}(r_{u,i} - \overline{r}_u)^2}\sqrt{\sum_{i \in C}(r_{v,i} - \overline{r}_v)^2}} \tag{2.1}$$

2. Coeficiente de correlación de *Spearman*:

   El coeficiente de correlación de *Spearman* denotado por ρ, se define por en la ecuación 2.2.

$$w_{u,v} = 1 - \frac{6 \sum_{i \in C}[ord(r_{u,i}) - ord(r_{v,i})]^2}{|C| * (|C| - 1)^2} \tag{2.2}$$

   En la ecuación (2.2) la función ord hace referencia al ordinal asociado a los elementos de los vectores r_u y r_v.

3. Similitud por *Coseno*:

En este caso $w_{u,v}$ se define simplemente por el $\cos(u, v)$, es decir:

$$w_{u,v} = \cos(r_u, r_v) = \frac{r_u \cdot r_v}{||r_u|| \cdot ||r_v||} = \frac{\sum_{i \in C} [r_{u,i} \cdot r_{v,i}]}{\sqrt{\sum_{i \in C} r_{u,i}^2} \cdot \sqrt{\sum_{i \in C} r_{v,i}^2}} \tag{2.3}$$

4. Correlación por distancia de *Manhattan*:

$$w_{u,v} = \frac{1}{1 + \sum_{i \in C} |r_{u,i} - r_{v,i}|} \tag{2.4}$$

   En la ecuación 2.4, al igual que en otras métricas, el valor máximo de correlación entre u y v se alcanza cuando las evaluaciones de u y v son coincidentes.

5. Similitud por Distancia Euclidiana

   Esta es una medida relacionada con correlación de *Pearson* pero menos sensible, ya que es simplemente la raíz de la suma de las diferencias al cuadrado entre dos evaluaciones. En muchos casos, los análisis de distancias euclidianas y de correlación de *Pearson* arrojan resultados parecidos.

$$w_{u,v} = ||r_u, r_v|| = \sqrt{\sum_{i \in C} (r_{u,i} - r_{v,i})^2} \tag{2.5}$$

2.4. Sistemas de recomendación por *Contenido*

Las técnicas de recomendación basadas en contenido usan la información de los recursos, tales como la *descripción* o anotaciones afectivas (Canini, Benini, y Leonardi, 2013), para determinar ítems similares a los recursos usados anteriormente por un usuario u y generar recomendaciones (Jannach, Zanker, Felfernig, y Friedrich, 2003). No se consideran las evaluaciones asignadas por u a los ítems de $\mathscr{I}_u$, y tampoco las visitas de otros usuarios, por lo que el perfil de preferencias de u se reduce simplemente a $PP_u = \mathscr{I}_u$.

El contenido de los ítems de PP_u se utiliza para encontrar ítems de mayor similitud que aun no consume u (Nadolski et al., 2009). Las técnicas de *Recuperación de Información* ofrecen para esta tarea una amplia variedad de algoritmos (Drachsler et al., 2010; Nadolski

et al., 2009). Se asume, en general, que el contenido de los ítems es *texto* (al estilo de un libro o artículo). Para nuestro objeto de estudio los ítems son ecursos digitales en una variedad de formatos: fotos, hojas de cálculo, documentos pdf, archivos de música o multimedia, aplicaciones, etc.; muchos de los cuales no corresponden a formatos basados en texto. Para sortear la falta de texto de algunos ítems y además tener el mismo patrón de comparación entre los distintos elementos de $\mathscr{I}$, usamos como *texto* los metadatos asociados a los ítems.

2.5. Sistemas de recomendación *Colaborativos*

Un sistema de recomendación, busca sugerir algún producto o ítem a un usuario, para ello las técnicas de filtrado colaborativo son las más utilizadas (Jannach et al., 2003; Segaran, 2007). El términos simples, se entiende por filtrado colaborativo al método de hacer predicciones automáticas (filtrado) acerca de los intereses de un usuario, recolectando información sobre los gustos varios de otros usuarios con intereses similares (colaboración) (Drachsler et al., 2010; Jannach et al., 2003; Manouselis, Vuorikari, y Van Assche, 2010; Santos y Boticario, 2010; Segaran, 2007; Felfernig y Burke, 2008). Los sistemas colaborativos se pueden dividir en dos tipos, los basados en *memoria* y los basados en *modelos* (Adomavicius y Tuzhilin, 2005), según el algoritmo que utilicen. Los basados en memoria hacen predicciones basadas en la colección de recursos evaluados por el usuario, los basados en modelos usan colecciones de ratings, para que el modelo aprenda y luego haga predicciones.

De las técnicas de filtrado colaborativo hay variaciones, quizás las más importantes sean las de filtrado por *usuario* y por *ítem*. Una vista a estas variantes de filtrado colaborativo se presentan a continuación.

2.5.1. Filtrado colaborativo por *Usuarios*

El filtrado colaborativo también puede ser presentado como el problema de encontrar los valores de predicción para la matriz de preferencias *usuario-ítem* (Tabla 2.1) (J. Herlocker, Konstan, y Riedl, 2002; Segaran, 2007). Al ser una matriz de preferencias escasa,

esto porque pocos usuarios han evaluado una gran cantidad de ítems y además hay muchos ítems disponibles, se tiene una gran cantidad de valores de predicción $P_{u,i}$ por determinar.

Las técnicas de filtrado colaborativo para generar recomendaciones a un usuario u, están insertas como parte de un proceso, con una secuencia de pasos definidas que son: calcular similitudes entre usuarios, seleccionar vecinos cercanos, calcular los valores de predicción $P_{u,i}$ para ítems $i \in (\mathscr{I} \setminus \mathscr{I}_u)$. Finalmente se debe representar la recomendación. El esquema general de este proceso se ilustra en la figura 2.1, que se detalla a continuación.

Paso 1: Calculo de similitudes.

Elegida la métrica para similitud, sea correlación por *Pearson, Coseno, Manhattan, Spearman* u otra, el cálculo de similitud se reduce a determinar los valores de la matriz S, donde $s[u, v] = w_{u,v}$. Los valores para $w_{u,v}$ están normalizados (entre 0 y 1). Se considera que hay correlación (similitud) cuando $w_{u,v} > 0$. Correlaciones negativas (inversas) y cero se asumen en la matriz como 0.

Paso 2: Selección de vecinos cercanos.

Para determinar recomendaciones a un usuario u, se requiere encontrar usuarios de $\mathscr{U}$ que tienen gustos similares, estos usuarios constituyen una vecindad de u. Hay varias técnicas para definir vecindades, podemos destacar por *umbral* y la de los *mejores* $k-$*vecinos* kNN (J. Herlocker et al., 2002).

Los *mejores* $k-$*vecinos* de u se determinan, encontrando los k usuarios de $\mathscr{U}$ que tienen mayor similitud con u. En términos formales, si denotamos por V_u la vecindad de u, un usuario $v \in \mathscr{U}$ se agrega a V_u si y solo si $w_{u,v} > w_{u,\omega} \forall \omega \in \mathscr{U}$ y $|V_u \cup \{v\}| \leq k$

Paso 3: Calcular los valores de predicción $P_{u,i}$.

Para un usuario u, interesa determinar los valores $P_{u,i}$ para ítems $i \in (\mathscr{I} \setminus \mathscr{I}_u)$. La fórmula para $P_{u,i}$ depende de la técnica de filtrado utilizada. Por ejemplo, si la técnica fuese filtrar por las mejores votaciones, simplemente $P_{u,i} = \overline{r}_i$.

Una expresión que define la predicción $P_{u,i}$, utilizando votaciones de los ítems en PP_u y la de los usuarios de V_u, en filtrado colaborativo se muestra en la ecuación 2.6 (J. Herlocker et al., 2002). En esta ecuación un voto $r_{v,i}$ es relevante para $P_{u,i}$ cuando los usuarios u y v tienen mayor similitud. La similitud de usuarios se define a partir de PP_u y PP_v. Para calcular la similitud entre u y v se requiere que éstos tengan votaciones en común, esto es, $\mathscr{I}_u \cap \mathscr{I}_v \neq \emptyset$. De no darse esta condición la similud es simplemente $w_{u,v} = 0$.

$$P_{u,i} = \overline{r}_u + \frac{\sum\limits_{v \in \mathscr{U}_i \cap V_u} [r_{v,i} - \overline{r}_u] \cdot w_{u,v}}{\sum\limits_{v \in \mathscr{U}_i \cap V_u} w_{u,v}} \tag{2.6}$$

Paso 4: Presentar la recomendación.

En este paso, se seleccionan los ítem de $\mathscr{I}$ para los cuales se encontraron los mayores valores de predicción $P_{u,i}$. Otras consideraciones para el listado dependerán de interfaz, alternativas de presentación, información adicional a incluir en la recomentación, por mencionar algunos aspectos.

El esquema general de filtrado colaborativo ilustrado en la figura 2.1 es el que empleamos en la componete de filtrado colaborativo, de sistema de recomendación propuesto en ésta tesis.

2.5.2. Filtrado colaborativo por *Items*

A diferencia del filtrado basado en usuarios, la matriz de similitud se genera a partir de la similitud entre ítems. Para el cálculo de la similitud entre dos ítems $i, j \in \mathscr{I}$ se seleccionan los usuarios que han calificado ambos ítems, es decir, el conjunto $\mathscr{U}_i \cap \mathscr{U}_j$. Luego del cálculo de similitud, se procede a seleccionar los ítems más similares para entregar una recomendación.

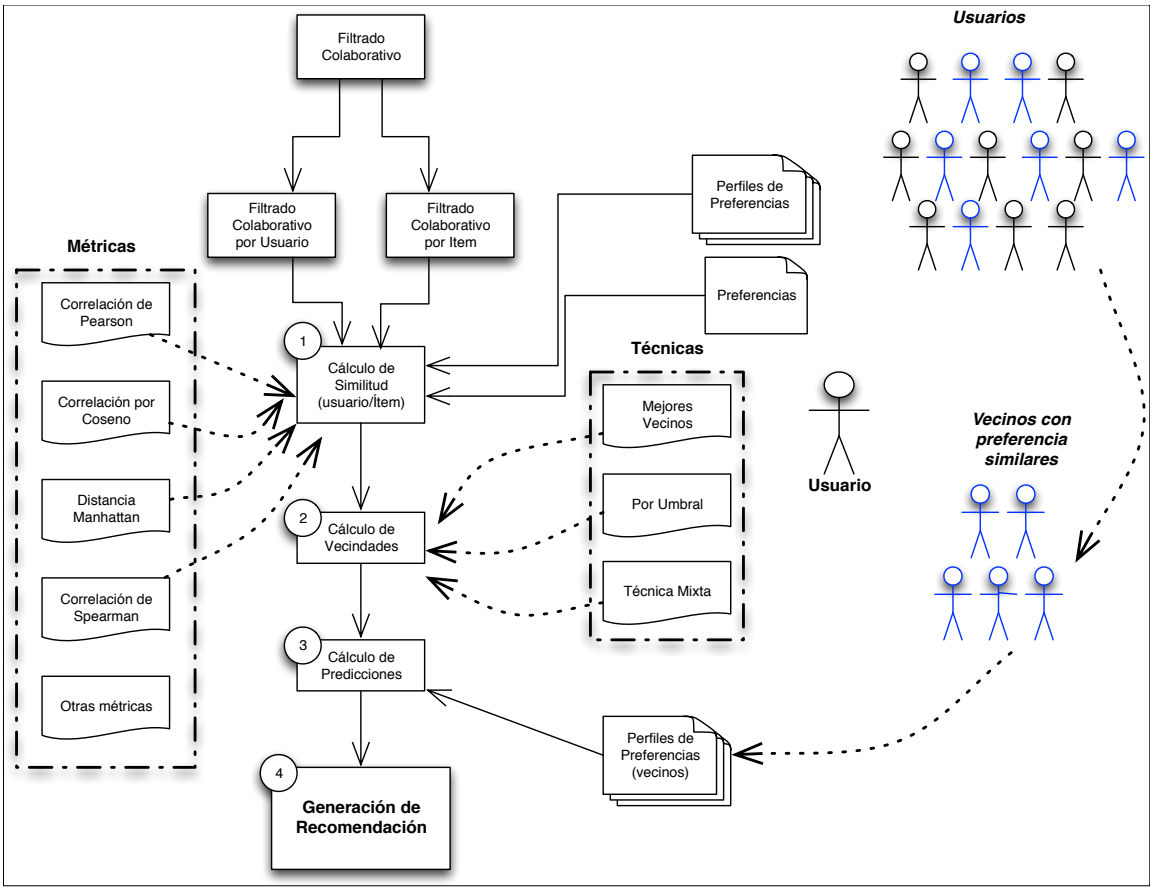


FIGURA 2.1. Componente de Filtrado Colaborativo, para el sistema de recomendación propuesto.

La ventaja de utilizar algoritmos de filtrado colaborativo basado en ítems es la eficiencia en *datasets* de gran tamaño (Segaran, 2007), esto debido a que la similitud entre ítems es más estable que la similitud entre usuarios. Esto permite pre-calcular la matriz de similitud haciendo que el proceso de generar recomendaciones sea mucho más eficiente. La similitud entre ítems se determina, al igual que la similitud entre usuarios, a partir de sus representaciones vectorial con los "pesos" de las palabras que componen el contenido de los ítems.

2.6. Sistemas de recomendación *híbridos*

Un sistema de recomendación hibrído, como su nombre lo indica, mezcla técnicas o enfoques distintos ya sea por que no se puede aplicar una técnica directamente o bien por que al mezclar técnicas se mejoran los resultados de la recomendación. Dependiendo de cómo se integren componentes de distintas técnicas, los sistemas de recomendación híbridos se pueden clasificar como (Burke, 2007):

1.- *Ponderado*: La puntuación de distintos componentes de recomendación se combinan numéricamente. Es simple, y se usa una fórmula lineal para combinar.

2.- *Conmutación*: El sistema elige entre componentes de recomendación y aplica el elegido. Según la situación elige uno u otro.

3.- *Mixto*: Recomendaciones de diferentes sistemas de recomendación son presentadas juntas. Se muestran una al lado de la otra en una lista combinada.

4.- *Combinación* de características: Características derivadas de distintas fuentes de conocimiento son combinadas y derivan en un solo algoritmo de recomendación. Toma prestada la lógica de recomendación de otra técnica, más que emplear un componente separado que lo implemente.

5.- *Incremento* de características: Es una técnica de recomendación usada para calcular una característica o conjunto de características, que luego son parte del input de la siguiente técnica. Se usa para añadir una fuente de conocimiento cuando ya hay un componente fuerte de recomendación.

6.- *Cascada*: Se da prioridades a los componentes de recomendación y los más débiles no pueden cambiar las recomendaciones de los más fuertes, sólo refinarlas.

7.- *Meta-nivel*: Una técnica de recomendación es aplicada y produce alguna clase de modelo, el cual es luego el input usado por la siguiente técnica.

En general, los sistemas híbridos de recomendación, teóricamente debiesen ser los mejores pues complementan las ventajas de los sistemas colaborativos y por contenido superando las desventajas que cada uno de ellos presenta por separado. Al ser el enfoque

mixto, una técnica que combina lo mejor de distintos enfoques, en general, obtiene mejores resultados.

2.7. Sistemas basados en conocimiento

Por otra parte, los sistemas basado en el conocimiento, usan un modelo de usuario que sirve como base para inferir las preferencias del usuario. Por ejemplo, en (Carrer-Neto, Hernández-Alcaraz, Valencia-García, y Sánchez, 2012) se plantea un sistema que considera la red social de los usuarios para determinar la similitud entre ellos. Una ontología de películas junto con información de contexto como el lugar, la hora y la multitud se utiliza en recomendar películas en (Mandl, Felfernig, Teppan, y Schubert, 2011). La información de contexto se utiliza para determinar la similitud de usuarios mientras que el conocimiento de películas es usado para determinar la similitud entre ítems en este sistema híbrido. En trabajo de Zhang (Zhang, Pablos, y Zhang, 2012), las preferencias de usuarios modeladas en base a una ontología, se tienen en cuenta para enfrentar el problema del arranque en frío y el comportamiento del usuario (por ejemplo, el tiempo gastado en la exploración) es usado como información implícita para calibrar las preferencias de los usuarios. También se considera la voluntad explícita de los usuarios por contribuir conocimiento (por ejemplo, recursos de aprendizaje, asesoramiento), y se encuentra que los incentivos sociales (es decir, el reconocimiento) influye positivamente en la contribución al conocimiento tácito, mientras que el incentivo monetario promueve la contribución explícita. Las preferencias de usuarios son analizadas en (Colombo-Mendoza, Valencia-García, González, Alor-Hernández, y Zapater, 2015), donde las preferencias son construidas dentro de una sesión de recomendación en lugar de estar predefinidas de forma estática. Los autores encontraron que aspectos tales como ítems de mala calidad, el orden de la recomendación (ranking ascendente o descendente), la formulación elegida por el usuario, y las las opciones por defecto pueden influir significativamente las preferencias de los usuarios. Las preferencias de los estudiantes, extraídas durante una actividad de ejercicios de preguntas y respuestas usando un modelo de cadenas de *Markov* en (Taraghi, Saranti, Ebner, Müller, y Großmann, 2015), sirven como base de un *perfil de aprendizaje* y una clasificación de

agrupación jerárquica. Otro sistema basado en el conocimiento se centra en las actividades más que en el usuario, por ejemplo, en (Rodrıguez, Gago, Rifón, y Rodrıguez, 2015), las actividades de aprendizaje contextualizadas conjuntamente con un enfoque multicriterio (factores de peso) son usadas para recomendar herramientas, personas y eventos a los profesores.

2.8. Evaluando predicciones

Los sistemas de recomendación se evalúan en base a la calidad de una recomendación, comparado los ratings reales ($r_{u,i}$) con los ratings de predicciones ($P_{u,i}$), a través de métricas tales como MAE (Error Medio Absoluto) ecuación 2.7, o NMAE (Error Medio Absoluto Normalizado) ecuación 2.8, (Goldberg, Roeder, Gupta, y Perkins, 2001; J. L. Herlocker et al., 2004; J. Herlocker et al., 2002).

$$\mathsf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|r_{u,i} - P_{u,i}| \tag{2.7}$$

$$\mathsf{NMAE} = \frac{\mathsf{MAE}}{\overline{r}_{max} - \overline{r}_{min}} \tag{2.8}$$

Otras medidas, relacionadas con la calidad de las recomendaciones son: *Precisión* denotada por P (ecuación 2.9), que es la relación entre los elementos que son relevantes para un usuario y el número total de artículos disponibles para recomendar; el *Recall* denotado por R (ecuación 2.10), que representa la probabilidad de seleccionar un elemento relevante; y finalmente, la *Cobertura* (o *Coverage*) denotado por C (ecuación 2.11), que es el porcentaje del conjunto de datos sobre el cual el sistema puede hacer predicciones (J. L. Herlocker et al., 2004).

$$\mathsf{P} = \frac{I_{rs}}{I_s} \tag{2.9}$$

TABLA 2.3. La calidad de una recomendación también depende de la capacidad del sistema de recomendar un sub-conjunto apropiado de elementos.

	Seleccionados	No seleccionados	Total
Relevantes	I_{rs}	I_{rn}	I_r
Irrelevantes	I_{is}	I_{in}	I_i
Total	I_s	I_n	I

$$\mathsf{R} = \frac{I_{rs}}{I_r} \tag{2.10}$$

$$\mathsf{C} = \frac{I_s}{I} \tag{2.11}$$

Los resultados obtenidos con las métricas acá mencionadas, se detallan el la sección 3.1 de este documento.

2.9. El problema del arranque en frío (*cold start problem*)

Uno de los problemas más difíciles de enfrentar en los SR es el problema del arranque en frío, que consiste en la recomendación a nuevos usuarios, o de nuevos recursos, o ambos (Lika, Kolomvatsos, y Hadjiefthymiades, 2014b), puesto que ni los nuevos usuarios ni los nuevos recursos registran algún tipo de preferencia o valoración (no tienen historia). La mayoría de los enfoques que enfrentan el problema del arranque en frío se centran en recomendar nuevos items (recursos en nuestro caso). En este trabajo nos centraremos en recomendaciones de nuevos items para nuevos usuarios de una comunidad educativa, particularmente profesores; ya que esta comunidad exhibe un comportamiento específico debido a las regulaciones que estructuran sus contenidos y actividades.

La caracterización de usuarios en el área de recomendación educativa, se ha propuesto para fines diferentes a enfrentar el problema de arranque en frío. Por ejemplo, (Drachsler et al., 2010) propone el uso de estereotipos de estudiantes, que son categorías basadas en datos demográficos y en objetivos de aprendizaje, sin embargo, los autores no describen cómo este enfoque se pondría en práctica. (Tang, Winoto, y McCalla, 2014) proponen una

recomendación multidimensional y el reconocimiento del contexto alumno con el fin de diferenciar los grupos de alumnos. Encontraron que las recomendaciones dentro del mismo grupo de aprendizaje son más eficaces. (Song y Gao, 2014) proponen etiquetas para identificar contenido de aprendizaje recursos (texto) junto con filtrado colaborativo con el fin de enfrentar el problema de arranque en frío, con resultados prometedores. Indicadores como la diversidad y la similitud de interés de los usuarios en un grupo e intra-grupos se utilizan para determinar grupos de aprendizaje en (Dascalu, Bodea, Lytras, Pablos, y Burlacu, 2014). Los estudiantes se agrupan automáticamente, por elección de los profesores, o de acuerdo a un algoritmo especializado basado en dichos indicadores. La diversidad y la similitud se determinan a partir de un perfil de aprendizaje (intereses explícitos de los estudiantes) y de su historia durante un proceso de aprendizaje.

En esta tesis modelamos las preferencias de usuarios por medio de un perfil demográfico del profesor. También usamos los metadatos de los recursos de aprendizaje con el fin de definir un método híbrido que combina la recomendación basada en el contenido y el filtrado colaborativo para enfrentar el problema de arranque en frío para nuevos usuarios y nuevos ítems.

2.10. Sistemas de Recomendación Educativos

Los sistemas de recomendación en educación se han estudiado como un medio de ayuda a usuarios en la búsqueda de recursos adecuados, entre otras tareas (Duval, Vuorikari, y Manouselis, 2009). Como indicamos, las técnicas usadas en SR están clasificadas como *basadas en contenido*, *filtrado colaborativo*, *basada en conocimiento*, o *técnicas híbridas*.

Los sistemas de recomendación en el contexto de la educación son muy diferentes de los sistemas de recomendación para productos o servicios, ya que deben tener en cuenta no sólo las preferencias de alumnos o profesores por ciertos materiales o tópicos, sino también cómo este material puede ayudarles a lograr sus objetivos. Por ejemplo, el proyecto SMART (Duval et al., 2009) recomienda ítems teniendo en cuenta características pedagógicas, tales como el conocimiento previo de los alumnos junto con sus intereses.

Nadolski (Nadolski et al., 2009) evalúa una estrategia basada en ontologías para modelar el perfil de los alumnos frente a un enfoque ligero (evaluación de pares) y encuentra que una técnica basada en ontologías es costosa pero más precisa. Manouselis (Manouselis y Costopoulou, 2007) propone un algoritmo clásico de filtrado colaborativo basado en vecindades para recomendar LOs considerando evaluaciones multidimensionales sobre LOs, proporcionadas por los profesores (evaluación de pares). Una buena revisión del área se puede encontrar en el trabajo de Duval (Duval et al., 2009). Shepitsen (Shepitsen, Gemmell, Mobasher, y Burke, 2008) sigue un enfoque diferente, aunque no en el campo de la educación. En este trabajo se usan *tags* (etiquetas) asignados libremente por los usuarios para definir indirectamente un perfil de los recursos preferidos, para cada usuario. Los tags son agrupados siguiendo un método de agrupamiento jerárquico; los grupos (*clusters*) se utilizan como base para un algoritmo personalizado que a partir de una sola consulta a una etiqueta, encuentra los ítems más similares. Los resultados se ponderan y se clasifican en base a los intereses del usuario, entendidos como la suma de los productos de las anotaciones del tag para los usuarios de cada grupo y la proporción de los recursos anotados con una etiqueta en el cluster. El trabajo de Chatti (Chatti, Dakova, Thus, y Schroeder, 2013), dieciséis algoritmos de filtrado colaborativo basados en etiquetas se ponen a prueba en precisión y exhaustividad (*recall*). Se encontró que la agrupación jerárquica basada en ítems y el agrupamiento de k-vecinos obtiene los mejores resultados mientras que las técnicas de agrupamiento basada en el usuario muestra los resultados más pobres debido a la poca densidad.

En los experimentos de Kurilovas (Kurilovas, Serikoviene, y Vuorikari, 2014), a un grupo de alumnos se le asigna un conjunto de tags a recursos de aprendizaje con el fin de crear en forma bottom-up un modelo del contexto del alumno, mientras que para usuarios expertos (por ejemplo, profesores para el caso de las propiedades pedagógicas) se obtienen modelos top-down de su propio ranking de tags, para determinar la calidad de un recurso de aprendizaje. En (Sieg, Mobasher, y Burke, 2010) una ontología de dominio (en lugar de tags) se utiliza como la base para el perfil de usuario, en un experimento de recomendación colaborativa no-educacional. Los conceptos en la ontología se organizan jerárquicamente

y dicha relación se considera al determinar el interés del usuario en un concepto a través de votaciones. La similitud de usuarios es establecida a partir de los ratings contenidos en los perfiles de usuario y el algoritmo de recomendación considera la similitud sobre la base de la distancia entre los intereses de usuarios. El etiquetado colaborativo asistido por una ontología de conceptos también se utiliza en la búsqueda multimedia en (Gayo, Pablos, y Lovelle, 2010) con el fin de mejorar las respuestas de los algoritmos de búsqueda. Las consultas de usuarios son enriquecidas con información de la ontología (expresada en la consulta) de modo que se obtienen los videos más similares para el interés del los usuarios.

2.11. Metadatos en Educación

Se pueden encontrar recursos de aprendizaje sin control de calidad mediante una búsqueda de contenido realizada a través de motores de búsqueda en la Web como Google o Yahoo, o incluso en sitios públicos como Wikipedia, YouTube, iTunes o el MIT OpenCourseWare (Tiropanis et al., 2009). El problema con este enfoque no es la falta de contenido, sino la abundancia excesiva de material y la falta de reconocimiento de la cultura y la práctica educativa en el diseño de los algoritmos de búsqueda (Tiropanis et al., 2009). Profesores, particularmente de enseñanza básica y media, forman una comunidad de práctica que comparte una terminología, herramientas y habilidades comunes. Esta propiedad se refleja en los repositorios especializados, tales como MERLOT [4], que contiene más de 45.000 recursos y proporciona una clasificación de estos sobre la base de una lista limitada de 23 disciplinas académicas (ej. Agricultura y Ciencias del Ambiente), o de apoyo académico a comunidades (por ejemplo, desarrollo de Facultades), entre otros.

La concepción actual y tradicional de *metadatos* es de ***datos sobre datos***. Antes del desarrollo de Internet, el concepto *metadato* era utilizado en el ámbito de bibliotecas, en donde los recursos (libros o revistas) se utilizaban catalogados y organizados por datos asociados al recurso tales como: *autor, título, editorial*, etc. Estos datos, se conocen como *metadatos*; un *metadato* es un dato estructurado sobre un recurso. En un sentido estricto, los

[4]Multimedia Educational Resources for Learning and Online Teaching (`www.merlot.org`)

metadatos sólo serían posibles en un contexto digital y en red (Steinacker, Ghavam, y Steinmetz, 2001), ya que en este contexto se pueden utilizar los metadatos con la función que les caracteriza que es la de la *localización, identificación* y *descripción de recursos* legibles e interpretables por una máquina. Esta concepción estricta, es la que define *Berners-Lee* y la *World Wide Web Consortium* (W3C) en *Metadata Architecture* (Berners-Lee, 1997), "Los *metadatos* son información inteligible para el computador sobre recursos Web u otras cosas". Dadas estas características, los metadatos pueden constituirse en un eje central de los mecanismos de búsqueda y recuperación de recursos (Ding et al., 2004).

Existen distintos modelos de metadatos que presentan diferentes esquemas de descripción para los objetos que se quieren describir, muchos modelos comparten metadatos referidos a: el contenido o concepto al que se refiere el recurso, el copyright del recurso, características de formato (tipo, tamaño, fecha, lenguaje, etc.), etc. Existen también iniciativas de estandarización de metadatos, entre ellos destacan dos como los estándares más usados, esto los describimos a continuación:

a. **Dublin CORE:**

Modelo elaborado y auspiciado por la DCMI (Dublin Core Metadata Initiative) (D. C. M. Initiative et al., 2000), organización que fomenta la adopción de estándares, y promueve el desarrollo de vocabularios especializados para describir recursos y permitir con ello el descubrimiento del recursos a través de sistemas inteligentes. La iniciativa DCMI comprende un conjunto de 13 elementos que a su vez definen 15 descriptores que son:

1. DC. *Title* (Título): Título dado a un recurso.
2. DC. *Creator* (Autor): Entidad principalmente responsable de la creación del contenido intelectual del recurso.
3. DC. *Subject* (Materias y palabras clave): El tema del contenido del recurso. Un tema es expresado como palabras claves, frases claves o códigos de clasificación que describan el tema de un recurso.
4. DC. *Description* (Descripción): La descripción del contenido del recurso.

5. DC. *Publisher* (Editor): La entidad responsable de hacer que el recurso se encuentre disponible.
6. DC. *Contributor* (Colaborador): Entidad responsable de hacer colaboraciones al contenido del recurso.
7. DC. *Date* (Fecha): Fecha asociada con un evento en el ciclo de vida del recurso.
8. DC. *Type* (Tipo): Naturaleza o categoría del contenido del recurso.
9. DC. *Format* (Formato): El formato puede incluir el tipo de media o dimensiones del recurso.
10. DC. *Identifier* (Identificación): Referencia no ambigua para el recurso dentro de un contexto dado.
11. DC. *Source* (Fuente): Es una referencia a un recurso del cual se deriva el recurso actual.
12. DC. *Language* (Lenguaje): Lengua del contenido intelectual del recurso.
13. DC. *Relation* (Relación): Referencia a un recurso relacionado.
14. DC. *Coverage* (Cobertura): La extensión o ámbito del contenido del recurso.
15. DC. *Rights* (Derechos): Información sobre los derechos de propiedad

b. **IEEE-LOM (*Learning Object Metadata*)**

Es un modelo de datos usado para describir un objeto de aprendizaje y otros recursos digitales similares usados para el apoyo al aprendizaje (Committee, 2002). Este estándar es generalmente usado en sistemas de gestión de aprendizajes LMS (*Learning Management Systems*) (*IMS Global Learning Consortium*, 1995).

La metadata de IEEE-LOM se divide en 9 categorías y estas a su vez se dividen en sub-categorías (o sub-niveles). La Figura 2.2 presenta una vista general de esta estructura. El detalle de cada categoría de metadatos IEEE-LOM y sus respectivas sub-categorías se presenta a continuación:

1. Categoría *general*. Representa información general sobre el material educativo que describe el mismo como un todo. Esta categoría de divide en los siguientes sub-niveles:

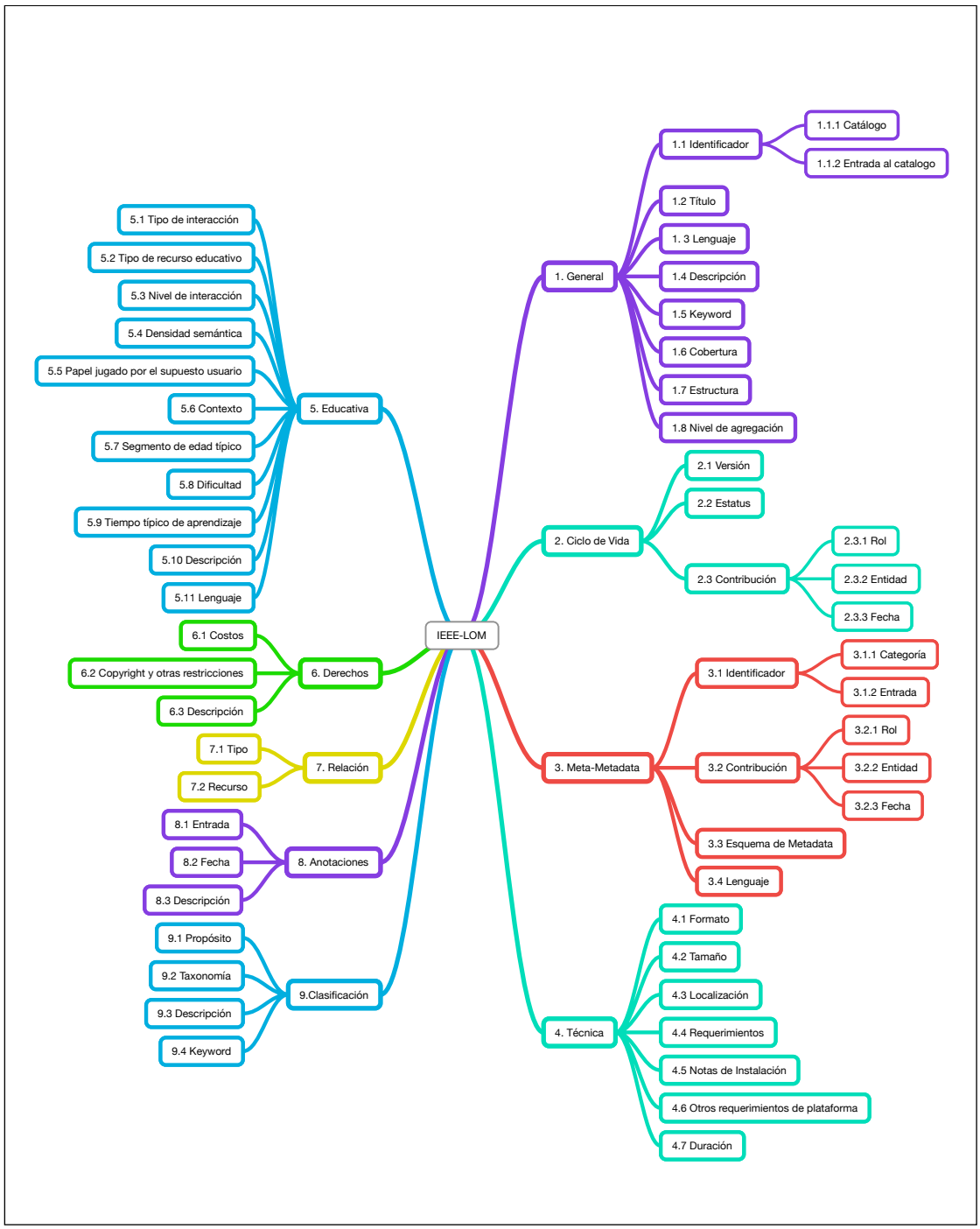


FIGURA 2.2. Esquema general de metadatos IEEE-LOM.

1.1. Identificador (*identifier*). Corresponde a un identificador descriptivo del material educativo, que debe tener un valor que permita identificar unívocamente el material en su contexto educativo. Esta compuesto por un par formado por un nombre de *catálogo*, y el nombre de la *entrada* en dicho catálogo. El uso que se da a este metadato es la selección de recursos cuando estos se encuentran indexados en catálogos.

1.2. Título (*title*): Nombre descriptivo del material educativo.

1.3. Lenguaje o idioma (*language*). Es el idioma primario utilizado en el material para comunicarse con los potenciales consumidores.

1.4. Descripción (*description*). Texto que describe el contenido del material.

1.5. Palabra clave (*keyword*). Una colección de frases que representan palabras claves sobre el material.

1.6. Cobertura (*coverage*). Son eventos temporales, culturales o geográficos asociados con el material.

1.7. Estructura (*structure*). Es la estructura interna del material. LOM propone un vocabulario controlado para describir la estructura de un recurso pero los autores pueden utilizar sus propios vocabularios, adaptados a sus necesidades pedagógicas particulares.

1.8. Nivel de agregación (*aggregation level*). Define la granularidad del material. También se define un vocabulario controlado para definir granularidad.

2. Categoría *lifecycle* (ciclo de vida). Agrupa metadatos referidos a la historia y estado actual del proceso de producción y mantenimiento del material educativo por parte de los autores. Esta categoría de divide en los siguientes sub-niveles:

2.1. Versión (*version*). Es la edición o versión del material.

2.2. Estatus (*status*). El estado de producción del material. LOM al igual que en otras categorías o subcategorías propone un vocabulario para estos metadatos que en este caso son: draft (borrador), final, revised (revisado), unavailable (no disponible).

2.3. Contribución (*contribute*). Acá se introduce información acerca de contribuyentes a la producción del material. La contribución puede incluir información de: el papel o rol del que contribuye, su identidad y fecha de la contribución.

3. Categoría *metametadata* (meta-metadatos). Agrupa información relativa a los metadatos en sí. Esta categoría de divide en los siguientes sub-niveles:
   3.1. Identificador (*identifier*). El identificador del conjunto de metatados para el recurso. Este identificador puede utilizarse para seleccionar el conjunto de metadatos, cuando éste se encuentra almacenado externamente.
   3.2. Contribución (*contribute*). Contribuyente a la elaboración de estos metadatos. Para cada contribuyente es posible especificar, al igual que en la categoría de ciclo de vida el rol, la identidad y la fecha.
   3.3. Esquema de metadatos (*metadata scheme*). El esquema de metadatos utilizado (scheme XML), que puede ser, por ejemplo, LOMv1.0
   3.4. Lenguaje (*language*). El idioma/lenguaje por defecto utilizado para proporcionar los metadatos.
4. Categoría *technical* (*técnica*). Agrupa metadatos relativos a las características y requisitos técnicos de los recursos. Esta categoría de divide en los siguientes subniveles:
   4.1. Formato (*format*). Formato del material. Dado que el material no tiene porque ser atómico, es posible que integre múltiples formatos, por ejemplo, una página web puede integrar un documento HTML con un conjunto de imágenes JPG.
   4.2. Tamaño (*size*). Tamaño en bytes del material.
   4.3. Localización (*location*). Forma de localizar al material, una URL por ejemplo.
   4.4. Requerimiento (*requirement*). Plataforma necesaria para utilizar este material. Dicha plataforma puede describirse en términos de las siguientes características como *tipo* y *nombre* de la plataforma.

4.5. Otros requerimientos de plataforma (*other platform requeriments*). Acá se especifican otros requerimientos de hardware y software.

4.6. Duración (*duration*). La *duración*, pensada para el material que tiene sentido una duración en su reproducción, como por ejemplo, un video o una presentación Flash.

5. Categoría *educational* (educativa). Agrupa metadatos relativos a los usos educativos del material. Esta categoría de divide en los siguientes sub-niveles:

   5.1. Tipo de interacción (*interactivity type*). Tipo de interacción soportado por el material. Se propone como vocabulario para caracterizar este tipo de interacción: activos, expositivos, mixtos, no definidos.

   5.2. Tipo de recurso educativo (*learning resource type*). Especifica el tipo de material (por ejemplo, ejercicio, figura, etc.). Un mismo material puede tener distintos tipos asociados. Parte del vocabulario sugerido para caracterizar el tipo de materia es: ejercicio, simulación, cuestionario, diagrama, figura, gráfico, etc.

   5.3. Nivel de interacción (*interactivity level*). Especifica el nivel de interacción del material. Se propone el siguiente vocabulario controlado para especificar dicho nivel: very low (muy bajo), low (bajo), medium (medio), high (alto), very high (muy alto)

   5.4. Densidad semántica (*semantic density*). Es una medida subjetiva de la utilidad educativa del material en comparación con su tamaño y/o duración. Se propone usar para expresar este nivel el mismo vocabulario controlado que para *interactivity level*.

   5.5. Papel jugado por el supuesto usuario (*intended user role*). Determina el papel del usuario final del material. Se propone como vocabulario para describir dicho papel: profesor/educador, autor, aprendiz, gestor.

   5.6. Contexto (*context*). El entorno educativo típico en el que se usará el material. Se propone el siguiente vocabulario: educación primaria, educación secundaria, educación superior, primer ciclo universitario, segundo ciclo

universitario, postgrado, primer ciclo de escuela técnica, segundo ciclo de escuela técnica, formación profesional, formación continua, formación vocacional.

5.7. Segmento de edad típico (*typical age range*). Rango de edades típico de los usuarios a los que va dirigido el material.

5.8. Dificultad (*difficulty*). Grado de dificultad del material. Se propone el siguiente vocabulario para caracterizar dicho grado: muy fácil, fácil, medio, difícil), muy difícil.

5.9. Tiempo típico de aprendizaje (*typical learning time*). Tiempo de aprendizaje típico asociado con el material.

5.10. Descripción (*description*). Comentarios sobre el uso del material desde un punto de vista pedagógico.

5.11. Idioma (*language*) Idioma del usuario final.

6. Categoría *rights* (derechos). Agrupa metadatos relativos a los derechos de propiedad intelectuales del material. Esta categoría de divide en los siguientes subniveles:

   6.1. Costos (*cost*). Establece si el recurso es o no de pago.

   6.2. Copyright y otras restricciones (*copy right and other restrictions*). Establece si el recurso está o no sujeto a derechos de copia y otras restricciones.

   6.3. Descripción (*description*). Comentarios sobre las condiciones y derechos de uso del recurso.

7. Categoría *relation* (relación). Metadatos utilizados para establecer relaciones entre distintos recursos. Un mismo material puede mantener múltiples relaciones con otros materiales. Cada una de estas relaciones exhibe las siguientes características:

   7.1. Tipo de la relación. Como es parte de (isPartOf), el materia tiene otro como integrante (hasPart), es una versión, etc.

   7.2. Recurso.

8. Categoría *annotation* (anotaciones). Anotaciones y comentarios sobre el material educativo. Estas anotaciones pueden caracterizarse por:

8.1. El anotador que realiza la anotación.

8.2. La fecha de la anotación.

8.3. El texto de la anotación

9. Categoría *classification* (clasificación). Metadatos para la clasificación del material en taxonomías. Cada clasificación puede tener asociada la siguiente información:

9.1. El propósito de la clasificación. Se propone un vocabulario controlado con términos de propósitos como: disciplina, idea, pre-requisitos, objetivo educativo, restricciones de acceso, nivel educativo, etc.

9.2. Taxonomía (*taxonomy*). Una serie de rutas en distintas taxonomías.

9.3. Descripción (*description*). Una descripción textual del material relativa al propósito de clasificación establecido.

9.4. Palabras claves (*keywords*). Un conjunto de palabras clave relativas al propósito de clasificación establecido.

Tanto IEEE-LOM como *Dublin Core* (DC), se pueden ver como equivalentes, es más los mismos estándares así lo indican. La tabla 2.4 muestra el mapeo del estándar DC a IEEE-LOM que hace la misma IEEE (Learning Technology Standards Committee of the IEEE, 2002).

Otras iniciativas orientadas a la estandarización de metadatos, para compartir recursos digitales de aprendizaje, es SCORM (*Sharable Content Object Reference Model*), modelo de referencia de objetos de contenido compartible. No es un modelo distinto a IEEE-LOM, sino que establece reglas para poder compartir recursos entre distintas plataformas. Los principales requerimientos que el modelo trata de cumplir son: accesibilidad, adaptabilidad, durabilidad, inter-operabilidad, y reusabilidad (A. D. L. Initiative, 2001).

TABLA 2.4. Correspondencia entre elementos *Dublin Core* y IEEE-LOM

DC	IEEE-LOMM
DC.Identifier	1.1.2:General.Identifier.Entry
DC.Title	1.2:General.Title
DC.Language	1.3:General.Language
DC.Description	1.4:General.Description
DC.Subject	1.5: General.Keyword or 9: Classification with 9.1: Classification.Purpose equals “Discipline” or “Idea”
DC.Coverage	1.6:General.Coverage
DC.Type	5.2:Educational.LearningResourceType
DC.Date	2.3.3:LifeCycle.Contribute.Date when 2.3.1:LifeCycle.Contribute.Role has a value of “Publisher”
DC.Creator	2.3.2:LifeCycle.Contribute.Entity when 2.3.1:LifeCycle.Contribute.Role has a value of “Author”
DC.OtherContributor	2.3.2:LifeCycle.Contribute.Entity with the type of contribution specified in 2.3.1:LifeCycle.Contribute.Role
DC.Publisher	2.3.2:LifeCycle.Contribute.Entity when 2.3.1:LifeCycle.Contribute.Role has a value of “Publishe”
DC.Format	4.1:Technical.Format
DC.Rights	6.3:Rights.Description
DC.Relation	7.2.2:Relation.Resource.Description
DC.Source	7.2:Relation.Resource when the value of 7.1:Relation.Kind is “IsBasedOn”

3. ENCONTRANDO CATEGORÍAS DE METADATA A TRAVÉS DE RECOMENDADORES

En este capítulo se presenta nuestra estrategia para alcanzar los objetivos de esta tesis, es decir, identificar metadatos relevantes para el proceso de búsqueda (O1), caracterizar a los profesores mediante metadatos (O2), definir comunidades de interés (O3), definir una estrategia de recomendación en base a la metadata definida en O1 y O2 (O4), y validar ésta estrategia mediante la recomendación de nuevos recursos a nuevos usuarios (O5). Nuestra estrategia sigue cuatro Fases:

Fase 1 : Para alcanzar el objetivo O3, se utilizó la información de una base de datos entregada por el portal *EducarChile*, que registra las visitas de profesores a recursos LOs durante 5 años. Con la base de datos se creó un dataset, es decir, un subconjunto de los datos, sobre el cual se basa esta investigación. Para identificar las comunidades relevantes se utilizó un sistema de recomendación de filtrado colaborativo y un algoritmo de agrupación (*clustering*) jerárquico que permitió identificar *comunidades de interés común* (O3).

Fase 2: En una segunda etapa, se caracterizó a los profesores de enseñanza básica y media (O2) y se definió un estrategia de recomendación basada en la metadata definida (O4).

Fase 3: Se definió un experimento para validar la calidad de la recomendación, basada en esta metada y estrategia, con una tarea avanzada en los sistemas de recomendación: la recomendación de recursos nuevos a usuarios nuevos (*cold start*). Para ello, se realizó un experimento con profesores en práctica y se analizó la calidad de la recomendación basada en metadata versus una recomendación clásica de filtrado colaborativo (O5).

Fase 4: Finalmente, con la validación del experimento se procedió a analizar el conjunto de metadatos que permitieron una mejor calidad de recomendación (O1).

3.1. Fase 1: Objetivos O3

La naturaleza educativa de nuestra propuesta de sistema de recomendación de material educativo para profesores, hacen inviable el uso de un *datasets* al estilo de *MoviLens* (Miller, Albert, Lam, Konstan, y Riedl, 2003) para evaluar las recomendaciones generadas, esto porque los profesores tienen un comportamiento de consumo de objetos sujeto a sus objetivos pedagógicos en el marco de una especialización y una planificación de enseñanza. Propuestas como la de Tang (Tang et al., 2014) sugieren una recomendación multidimensional para estudiantes, que reconoce el contexto del alumno con el fin de diferenciar grupos de alumnos, y en esta investigación se propone lo mismo para el grupo de profesores.

El sistema educacional chileno (sin considerar la enseñanza universitaria o terciaria) incluye tres niveles: pre-escolar (al margen de este estudio), educación primaria (o básica) y educación secundaria (o media). El nivel primario incluye a estudiantes de 6 a 13 años de edad, inscritos en desde 1^{ero} a 8^{vo} básico. El nivel secundario (o enseñanza media), comprende 4 años (1^{ero} a 4^{to} medio), con alumnos en edades de 14 a 17 años de edad. El plan de estudios de nivel primario es uniforme para todo tipo de colegio, pero el nivel de secundaria se divide en enseñaza *científico-humanista* y *técnica-profesional*. Estos últimos dos sistemas, comparten un mismo plan de estudios para los dos primeros años (1^{ero} y 2^{do} medio), con algunas diferencias en cursos electivos, talleres y horas asignadas a las asignaturas. A partir de 3^{ero} medio, hay una profundización en las áreas de ciencias exactas o humanidades para el caso de educación *científico-humanista* como pueden ser biología, matemáticas o letras. En el caso de la educación *técnico-profesional*, se reduce el número de asignaturas obligatorias de ciencias y humanidades, dejando espacio para asignaturas donde los alumnos adquieren conocimientos técnicos profesionales, en distintos ámbitos del quehacer económico como mecánica-automotriz, contabilidad, agricultura, pesca, minería, forestal, etc. Además, en el sistema de educación *técnica-profesional*, los alumnos al terminar sus estudios reciben un título técnico profesional, de su área de especialización como pueden ser *Contabilidad, Turismo, Mecánica Automotriz*, por mencionar algunos ejemplos.

Con el fin de acotar el alcance de este trabajo, la información a ser considerada se centra en aquella generada por la plataforma Web *EducarChile*, al ser ésta la plataforma más usada por profesores de enseñanza básica y media (60 % de los profesores de Chile están inscritos en el portal). *EducarChile* es una iniciativa de la Fundación Chile[1] en conjunto con el Ministerio de Educación[2]. A continuación se presentan detalles del origen, manejo y resultado final de la información usada en esta tesis.

3.1.1. El *dataset*

EducarChile entregó una copia de su base de datos, que contenía información relacionada con las búsquedas de recursos digitales entre los años 2007 a 2011 (5 años). A partir de esta base de datos, se generó un *dataset* que contiene información de usuarios registrados como alumnos, profesores, personal administrativo, investigadores, estudiantes de pedagogía y público en general (usuarios no registrados). Además, contiene información sobre intereses manifestados por los usuarios, es decir, las visitas y calificaciones dadas a recursos de aprendizaje. Los recursos contienen también metadatos que han sido revisados por un conjunto de expertos en educación de *Educarchile*. La información que consideramos fue completamente revisada por estos expertos.

Para los experimentos sólo considerado usuarios registrados en *categoría de profesor*, y que han interactuado con la plataforma. La información recogida de los usuarios pertenece a dos categorías, la de datos personales (nombre, genero, nacionalidad, documento de DNI) y la datos de profesor (disciplina: por ejemplo, matemáticas, lenguaje, ciencias, biología, etc. Nivel: por ejemplo, 1^{ero} básico, 1^{ero} medio, etc.). Esta información no está disponible al público, pero se puso a disposición para esta investigación sujeta a su manejo confidencial.

La base de datos contenía originalmente 3TB de información, incluyendo datos que no eran relevantes para nuestro estudio. Se filtraron los datos, tanto en recursos de aprendizaje, metadatos, y usuarios registrados (profesores), como en número de visitas y calificaciones

[1] ww.fundacionchile.com
[2] www.mineduc.cl

dadas por los usuarios. Las características del *dataset* resultante se pueden apreciar en la tabla 3.1.

TABLA 3.1. Características de *dataset*.

Característica	Número	Porcentaje
Profesores registrados (usuarios)	11.260	100 %
Recursos de aprendizaje (ítems)	43.092	100 %
Total *visitas*	60.651	100 %
Total *ítems visitados*	5.621	13 %
30-top *ítems visitados*	2.736	21 %
Usuarios con al menos 10 visitas a items (U_{10})	1.376	12 %
Items visitados por U_{10}	4.538	7 %
Número de visitas generadas por U_{10}	25.544	42 %

El *dataset* contiene 60,651 visitas de profesores (usuarios registrados); sin embargo, tales usuarios visitaron sólo el 13 % *de los ítems del dataset*. Por otra parte, los 30 ítems más visitados representan representan el 21 % de estas visitas. Con el fin de reducir la escasez de visitas realizar pruebas utilizando un enfoque de validación cruzada (Devyver y Kittler, 1982), hemos limitado los usuarios establecidos a los que visitaron al menos 10 recursos diferentes (U_{10}), resultando en 1.376 usuarios, 4.538 recursos de aprendizaje y 25.544 visitas. La tabla 3.2 resume la distribución del total de visitas recibidas por los ítems asociados a U_{10}.

TABLA 3.2. Distribución de las visitas de U_{10}.

Número de visitas	Items visitados	Porcentaje
1	16.933	66.3 %
2	6.113	23.9 %
3	1.164	4.6 %
4	732	2.9 %
5	602	2.4 %

El dataset contiene ratings o evaluaciones hechas por los usuarios a los recursos, sin embargo estas evaluaciones no afectan ni al 1 % de los recursos de la base de datos, por lo que fueron descartados y en su lugar se consideraron las visitas a los mismos (*implicit rating*). Notesé además que pocos usuarios visitaron más de 5 veces un recurso (sólo 3

ocurrencias) por lo que se limitó el número máximo de vistas a 5. El número de recursos que fueron visitados por los profesores 1 o 2 veces, representa el 90,2 % del conjunto de datos. Lo que refleja que pocos recurso son visitados muchas veces, menos del 10 %.

3.1.2. Metadatos

El dataset contiene información sobre los intereses de los usuarios (es decir, visitas y evaluaciones a los recursos), los recursos de aprendizaje y sus metadatos, etc. La plataforma permite también a los profesores subir materiales cuya metadata es más tarde revisada, completada, discutida, anotada o modificada por el grupo de expertos. En este último caso, la metadata se revisó y normalizó. Todos los metadatos del dataset fueron revisados y modificados como parte de un proceso de curación estándar. EducarChile no realizó ningún proceso de curación adicional en la versión del dataset utilizado en este trabajo. Los expertos de Educarchile definen un conjunto específico de categorías de metadatos, sin embargo, cuando analizamos los metadatos asociados a los recursos como parte del proceso de agrupación jerárquica, la mayoría de las categorías de metadatos no es utilizada en la anotación de recursos, por lo que creemos que tal proceso de categorización no afecta a los resultados de nuestro enfoque. También consideramos todos los elementos descriptivos que pueden asociarse al recurso (por ejemplo, el titulo), como se describirá en detalle más adelante.

La información considerada en el data set es: ID del recurso, nombre, descripción, palabras claves, título, identificador Web de recursos (o `URL`), ID del metadato del recurso (código), ID del valor del metadato (código) y descripción del valor del metadato.

Los recursos de aprendizaje han sido anotados con 81 tipos de metadatos diferentes (ver la tabla 3.3), cada uno con una lista fija de valores posibles, con un total de 6.957 valores. Clasificamos los metadatos en tres categorías: De administración del sitio Web (incluye 19 metadatos), metadatos curriculares (nivel y asignatura), y los metadatos sociopedagógicos (62 metadatos curriculares). También incluimos *texto plano* (*libre*) escrito en lenguaje natural, que se utiliza para describir los recursos en la base de datos como *nombre, título y descripción*.

TABLA 3.3. Clasificación de la Metadata.

Categoría	Entrada de Metadata	Valores	Descripción
Metadata de administración del sitio web	19	774	Referencias a otros sitios web, palabras clave de sitios web, clasificaciones de materiales utilizados por el administrador del sitio web etc.
Metadata curricular	2	2.994	Incluye niveles(ejemplo 1^{ero} básico, 3^{ero} medio, etc.) y asignatura (ej. lenguaje y comunicación, etc.)
Metadata Socio Pedagógica	62	6.183	Incluye metadata curricular metadata (2.994 valores) y agrega metadata socio-pedagógica (3.189 valores) tales como *uso educativo, taxonomías, formación de los padres, etc.*
Contenido	3	-	Nombre, título, descripción.

La tabla 3.4 presenta un ejemplo de valores de metadatos para dos objetos clasificada de acuerdo a las categorías de la Tabla 3.3. La Figura 3.1 presenta una vista de un recurso del portal *EducarChile*, se pueden ver algunos de los metadatos, correspondientes a un recurso titulado "*Propiedades Métricas*", que es uno de los objetos sugeridos de la consulta de búsqueda del término *Geometría*

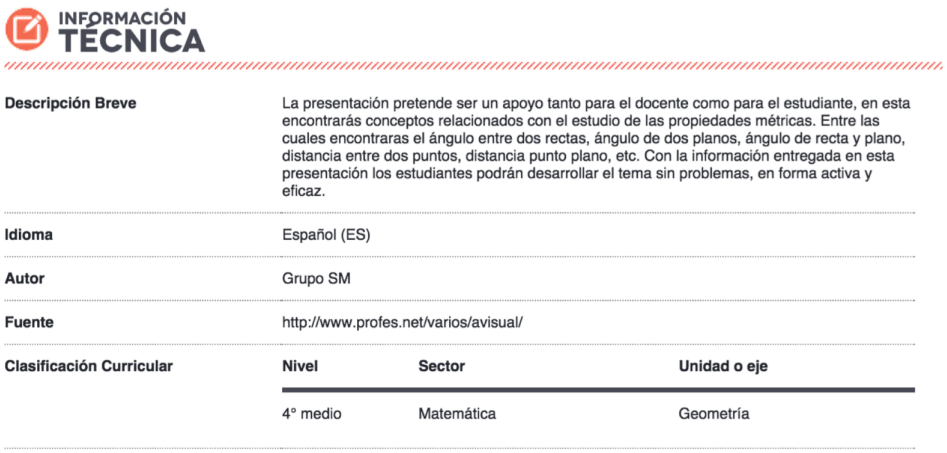


FIGURA 3.1. Vista de los metadatos para el recurso "Propiedades Métricas".

TABLA 3.4. Ejemplo de valores de metadatos por categorías para dos ítems

Nombre del recurso	Metadata de administración de sitio Web	Metadata Curricular	Metadata Socio Pedagogica
José de Espronceda, biografía	sitio; educación; biografías; literatura; estudiante; educación; c-h (científico humanista); t-p (técnico profesional); internacional	1^{ero} medio, comunicación; oral; leyendo	sitio; José; Espronceda; biografía; página; gastar; poeta; Español; característica; trabajo; autor; escritor; romántico; romanticismo; Español; poesía; poemas;
Crónica del siglo XX	software; Educación; estudiante; 2^{do} medio; c-h (científico humanista); t-p (técnico profesional)	8^{vo} básico; historia; geografía; ciencia; social; 4^{to} medio; America; Latina; contemporáneo; mundo; actualidad	SW (software); crónica; siglo; XX; equipo; referencia; formar; archivo; periodismo; noticias; evento; descubrimiento; pertinente; tipo; mundo; contemporáneo; guerra; historia; revolución; Ruso; Vietnam; ciencia; social; biografía; comunismo; EducarChile; calendario; America;

Los profesores, por su parte, están especializados según los tres niveles de educación, es decir, preescolar (profesores de párvulos), primaria (profesores de educación básica) y profesores de secundaria (profesores de educación media). Los profesores de educación básica se dividen en dos grupos, los generalistas que son responsables de un grupo de estudiantes de 1^{ero} a 4^{to} básico, donde enseñan todas las materias requeridas para esos niveles, y los que están especializados en una determinada asignatura (por ejemplo, las matemáticas) y enseñan en diversos cursos de 5^{to} a 8^{vo} básico. Los profesores de educación media, por otra parte, están especializados en una sola asignatura, la que enseñan en los diversos niveles, desde 1^{ero} a 4^{to} medio.

3.1.3. Tratamiento de la información

Con el fin de hacer un análisis del texto y metadatos que describen los recursos de aprendizaje, pre-procesamos el texto, y eliminamos las palabras como artículos, conjunciones, etc. También se redujeron las variaciones verbales (por ejemplo, *comer, comimos, comiendo* es reducida simplemente a ***comer***). Esta tarea de pre-procesamiento se realizó sobre las palabras de metadatos, así como en todo el "texto" para el caso de la categoría *Contenido* de la Tabla 3.3. El pre-procesamiento se realizó usando los algoritmos de *Stemming* y *Lemmatization* (Jurafsky y Martin, 2008), que en esta tesis fueron implementados por la librería *Freeling*[3], capaz de procesar texto en español. Finalmente aplicamos un algoritmo de mapeo para reducir el conjunto de valores posibles de metadatos curriculares, debido a las diferentes convenciones de nombres, por ejemplo, NB1, nivel básico 1, primero básico, 1^{ero} básico, NB-1, etc., todos estos nombres fueron mapeados a *primero básico*.

3.1.4. Aplicación de filtrado colaborativo en el *dataset*

De los enfoques de recomendación planteados en el capitulo 2, se escogió un enfoque híbrido, esto es, incluir un sistema de filtrado colaborativo clásico y un sistema basado en contenido. En el filtrado colaborativo esquematizado en la figura 2.1, la recomendación a un usuario u, está basada en su perfil de preferencias (PP_u) y las preferencias de usuarios similares.

Al aplicar el proceso de filtrado colaborativo definido en la sección 2.5, obtuvimos recomendaciones utilizando las métricas de similitud definidas en la sección 2.3, con diferentes tamaños de vecindades. Para validar nuestros resultados, se utilizó un enfoque de validación cruzada por diez (o *ten-fold cross-validation*) (Devyver y Kittler, 1982). Es decir, hemos dividido el conjunto de datos en 10 segmentos equivalentes al azar, luego hacemos 10 iteraciones y en cada iteración tomamos un segmento (diferente cada vez) para servir como conjunto de validación al contener las respuestas correctas, es decir, lo que el usuario realmente visitó (*gold standard*). Llamamos a este segmento *conjunto de prueba* (Test set, TS) y llamamos a los restantes 9 segmentos *conjunto de entrenamiento* (Training

[3]http://nlp.lsi.upc.edu/freeling/

set, TrS). Promediamos la calidad de las métricas en cada iteración (por ejemplo el MAE: la diferencia entre la recomendación generada en base al TrS y lo que el usuario realmente visitó registrado en el TS), y luego hacemos un promedio de las 10 iteraciones. Seguimos este enfoque con el fin de usar el mismo set de datos como conjunto de validación (puesto que contiene visitas reales de usuarios), y hacemos la validación cruzada para minimizar el ruido de los datos a través de la randomización.

Los conjuntos TrS y TS cumplen las siguientes condiciones:

1.- $|\mathscr{I}_u| \geq 10, \forall u \in \mathscr{U}$, esto, para garantizar que en cada iteración al menos una predicción debe ser realizada para cada usuario, y

2.- $S = (\mathsf{TrS} \bigcup \mathsf{TS})$ and $(\mathsf{TrS} \bigcap \mathsf{TS}) = \emptyset$.

Se analizaron las recomendaciones usando las métricas MAE, NMAE, *Recall*, *Precisión* y *Coverage*. La Tabla 3.5 presenta un resumen de los 10 mejores resultados, teniendo en cuenta las métricas de similitud y tamaños de vecindades (kNN) previamente definidas, ordenadas por el MAE.

TABLA 3.5. Los diez mejores algoritmos en términos de MAE, *Recall* y *Coverage*.

Ranking	M. Similaridad	kNN	MAE	NMAE	*Recall*	*Precision*	*Coverage*
1	Distancia	10	0.643978	0.12880	0.04570	0.18830	0,86239
2	Distancia	9	0.644287	0.12886	0.04570	0.18659	0.86239
3	Manhattan	10	0.644506	0.12890	0.04437	0.18416	0.86239
4	Distancia	8	0.644702	0.12894	0.04836	0.19326	0.86239
5	Distancia	7	0.644773	0.12895	0.04925	0.19072	0.86239
5	Manhattan	9	0.644774	0.12895	0.04392	0.18066	0.86239
7	Manhattan	7	0.644969	0.12899	0.04703	0.18435	0.86239
8	Manhattan	8	0.645114	0.12902	0.04614	0.18705	0.86239
9	Distancia	6	0.645302	0.12906	0.05235	0.19799	0.86239
10	Manhattan	6	0.645558	0.12911	0.05013	0.19120	0.86239

Como se aprecia en la tabla 3.5, los mejores resultados corresponden a *distancia euclidiana*, con 10 vecinos, un MAE de 0.643978, *Coverage* 0.86239, mientras que el *Recall* es 0.04570. Esta tabla no incluye los resultados de las métricas de *Coseno*, *Spearman* y *Pearson* ya que tienen un peor desempeño que los diez primeros. Dado que los mejores

resultados en términos de MAE se produjeron por Similitud por *Distancia Euclidiana* con un valor de $k = 10$ para kNN, elegimos esta configuración para el resto del proceso.

En nuestro caso las diferencias entre las diversas estrategias son mínimas al considerar cualquiera de las métricas aplicadas, pero el *recall* (probabilidad de que se seleccione un elemento relevante) es muy bajo; esto puede explicarse por la escasez de visitas (Ghazarian, Shabib, y Nematbakhsh, 2014; J. L. Herlocker et al., 2004) (90,2 % de los ítem reciben 1 o 2 visitas, ver Tabla 3.2). Los resultados, sin embargo, son comparables con experimentos similares en educación (Manouselis y Costopoulou, 2007, 2008). Por ejemplo, Manouselis (Manouselis y Costopoulou, 2008) hace un experimento basado en una evaluación de objetos de aprendizaje con atributos múltiples, por parte de profesores. El dataset fue creado para ese experimento específicamente y el problema de tener el dataset disperso (*sparsity*), es decir, con pocas visitas a los recursos, se evita pues todos los profesores evalúan un conjunto limitado de recursos. En este caso, la mejor métrica es MAE $= 0,57$ mientras que el *coverage* se mantiene en 69,08 % para la métrica de similitud por coseno y $k = 4$ en kNN. Hay una evaluación multiatributo en lugar de una discreta (0 o 1), se propone para mejorar el error (MAE). Las métricas de *precisión* y *recall* no son reportadas. En el trabajo de Verbert (Verbert et al., 2011), la métrica de MAE para un dataset en educación es muy similar para la similitud por *coseno* y *Pearson*, acá tampoco se reporta resultados de las métricas de *precisión* y *recall*. Zhao utiliza un algoritmo de filtrado colaborativo basado de ítems (Zhao, Liu, y Zhang, 2015), para recomendar recursos de aprendizaje (vídeo) en un entorno de aprendizaje a distancia, el dataset se caracteriza por una dispersión (escasez de evaluaciones) incremental (40 millones de usuarios, 9 mil ítems). En este caso, el mejor valor de MAE es de 0,8336 mientras que el *coverage* se mantiene en 99,8 % para un nivel de dispersión de 94,86 %. En nuestro caso, considerando únicamente el subconjuntos donde los usuarios han visitado al menos 10 ítems (diferentes), tenemos un nivel de dispersión de 99,59 %. Las métricas de *precisión* y *recall* no son reportadas.

Los resultado de esta sección, principalmente las vecindades generadas de la mejor combinación de métrica de similitud y tamaño k para los mejores vecinos con el algoritmo kNN serán utilizadas en la siguiente sub-sección para determinar *comunidades de interés*.

3.1.5. Identificando comunidades de interés a través de *clustering* jerárquico de profesores

Para descubrir las comunidades con intereses comunes, utilizamos *clustering* jerárquico (Hastie et al., 2009). Esta técnica agrupa los elementos (o los separa) en grupos basados en ciertos criterios. El método de aglomeración ascendente consiste en formar grupos progresivamente, agregando miembros hasta que se cree un conglomerado. Para aglomerar miembros es necesario definir una métrica para determinar los elementos que no pertenecen todavía a un cluster y que se encuentran a una distancia menor. Las métricas como *Pearson*, o *Distancia Euclidiana* se utilizan normalmente para estos efectos.

Recordemos que en la sección 3.1.4 se aplicó la distancia euclidiana y el algoritmo kNN (con $k = 10$) para recomendar recursos de aprendizaje basados en las visitas de usuarios. Ahora se agrupan las vecindades (de usuarios) que se encontraron en la sección 3.1.4 sobre la base de un umbral de p usuarios compartidos. Para ello se definió una métrica de similitud de vecindades a partir de una relación $\sim$. Esto es, dos vecindades N_i y N_j son similares, lo que denotamos por por $N_i \sim N_j$, cuando N_i y N_j comparten al menos el porcentaje p de sus vecinos (profesores). La ecuación 3.1 describe formalmente la relación de similitud $\sim$.

$$N_i \sim N_j \iff |N_i \cap N_j| \geq p \cdot |N_i| \text{ y } |N_i \cap N_j| \geq p \cdot |N_j| \tag{3.1}$$

Basados en la ecuación 3.1, definimos un cluster jerárquico $\mathscr{C}$, como la unión de todas las vecindades que son similares, la ecuación 3.2 así lo formaliza. El número total de clusters (comunidades de interés) encontrados es 1.376.

$$\mathscr{C} = \bigcup_{l \neq m} N_i \text{ donde } N_i \sim N_j \tag{3.2}$$

La tabla 3.6[4] presenta la distribución de grupos (*cluster*) de acuerdo al porcentaje p de vecinos que comparten. Si consideramos valores altos de p, como se muestra en la Tabla

[4]En la tabla 3.6, la notación $\#\mathscr{C}$ indica número (o cantidad) de vecindades que agrupa el cluters $\mathscr{C}$.

3.6, por ejemplo, un valor de $p = 0,85$ obtenemos 7 cluster que agrupan a 7 vecindades, pero no hay clusters que agrupen a 5 vecindades. Un valor más alto, $p = 0,9$, produce más agrupaciones de una sola vecindad (1.338), y ningún cluster que agrupe a 5, 6, o 7 vecindades. Como podemos ver, un umbral de similitud más restrictivo, genera menos posibilidades de agrupaciones. Por lo tanto, elegimos el valor de $p = 0,85$, ya que no solo garantiza un elevado número de usuarios comunes, sino también a diversos cluster agrupados.

TABLA 3.6. Cantidad de vecindades aglomeradas en cada cluster para distintos niveles de p.

p	$\#\mathscr{C} = 1$	$\#\mathscr{C} = 2$	$\#\mathscr{C} = 3$	$\#\mathscr{C} = 4$	$\#\mathscr{C} = 5$	$\#\mathscr{C} = 6$	$\#\mathscr{C} \geq 7$
$p = 0{,}75$	1282	38	13	5	5	1	32
$p = 0{,}80$	1312	36	2	5	2	2	17
$p = 0{,}85$	1338	24	3	3	0	1	7
$p = 0{,}90$	1354	17	3	2	0	0	0

3.1.6. Metadatos asociados a comunidades

En esta sección se busca descubrir los metadatos asociados a las comunidades de interés definidas en la sección 3.1.5, tomando como base los conceptos detallados en la sección 2.11.

Para efectos de notación, el conjunto de usuarios de un cluster lo denotaremos por $\mathscr{U}_{\mathscr{C}}$ y será el conjunto de todos los usuarios de las distintas vecindades de un cluster, tal como se define en la ecuación 3.3. De forma análoga, los ítems de un cluster denotados por $\mathscr{I}_{\mathscr{C}}$, corresponden al conjunto de ítems que han consultado cualquiera de los usuarios de $\mathscr{U}_{\mathscr{C}}$, tal cual define la ecuación 3.4.

$$\mathscr{U}_{\mathscr{C}} = \{u \in \mathscr{U} \mid \exists N \in \mathscr{C} \wedge u \in N\} \tag{3.3}$$

$$\mathscr{I}_{\mathscr{C}} = \{i \in \mathscr{I} \mid \exists u \in \mathscr{U}_{\mathscr{C}} \wedge i \in \mathscr{I}_u\} \tag{3.4}$$

Los ítems (recursos) disponibles en *EducarChile*, están descritos por metadata en alguna de las tres categorías descritas en la tabla 3.3: Metadata *curricular*, contenido en *texto plano*, y metadatos *socio pedagógicos*. Los valores de los metadatos podrían ser pequeñas sentencias de hasta 100 términos (o palabras) o textos libres cortos en el caso del *nombre* y *título del recurso*, o textos libres mucho más largos en el caso de la *descripción* (contenido).

Las palabras asociadas a un recurso en cualquiera de esas categorías las denominaremos "término". El conjunto de términos para un cluster lo denotaremos por $\mathscr{T}_{\mathscr{C}}$ y estará definido por cualquier término asociado a los ítems del cluster, como se indica en la ecuación 3.5.

$$\mathscr{T}_{\mathscr{C}} = \{t \in i \mid i \in \mathscr{I}_{\mathscr{C}}\} \tag{3.5}$$

La relevancia o "peso", de un termino t, en un cluster $\mathscr{C}$, que denotaremos por $w_c(t)$, estará definida con base en:

a.- El número de veces que este término es parte de la descripción de un ítem i, denotado por $\eta(i,t)$ y

b.- El número de usuarios que ha visitado ese ítem i, que será denotado por $\kappa(\mathscr{U}_{\mathscr{C}}, i)$

La ecuación 3.6 especifica en términos algebraicos la forma de calcular $w_c(t)$. El subíndice c en la expresión que define a $w_c(t)$ denota una categoría de metadatos para el término (t) (por ejemplo curricular, socio pedagógico etc.).

$$w_{\mathscr{C}}(t) = \sum_{i \in \mathscr{I}_{\mathscr{C}}} \eta(i,t) \cdot \kappa(\mathscr{U}_{\mathscr{C}}, i) \tag{3.6}$$

Un vector de términos ponderados $w_c(t)$ está asociado con cada cluster, por cada categoría. Así la representación vectorial de un cluster $\mathscr{C}$ será como lo indica la la ecuación 3.7. La tabla 3.7 presenta un fragmento de los vectores de términos asociados a dos cluster $\mathscr{C}_1$ y $\mathscr{C}_2$, cada uno de estos agrupa 3 o más vecindades que comparten al menos $p = 85\,\%$ de vecinos. Para facilitar el análisis hemos considerado sólo los 40 términos más relevantes por cluster y separamos estos términos en tres sub-categorías adicionales: asignatura,

nivel curricular, y otros términos. La separación de los términos en estas sub-categorías se realizó de forma manual por expertos en educación. Presentamos solo los cinco primeros términos por cada sub-categoría.

$$\overrightarrow{\mathscr{C}_c} = \langle w_c(t_1), w_c(t_2), \ldots, w_c(t_n) \rangle \tag{3.7}$$

De la tabla 3.7, las palabras con mayor peso son para metadatos curriculares, que corresponden a las sub-categorías de *asignatura* y *nivel*, mientras que para la categoría socio-pedagógica, la sub-categoría más pesada es *otros términos*, que describen aspectos más técnicos de una asignatura. El peso, de los términos en la categoría de *contenido* es sub-categoría *otros términos*, sin embargo, para el caso de las asignaturas términos genéricos como: *secundaria*, *científico-humanista*, y *técnico-vocacional*, sobrepasan las sub categorías de asignatura.

3.2. Fase 2: Objetivo O2, O4

En esta fase se busca caracterizar a los profesores, para ello, en conjunto con expertos de la Facultad de Educación de la PUC (Pontificia Universidad Católica de Chile), definimos las categorías de atributos para el perfil de profesor (Tabla 3.8) que incluye entre otros, datos demográficos para caracterizar la práctica del profesor.

Caracterizaciones de este tipo, expresadas durante la aplicación de un test de preguntas y respuestas, han sido utilizadas en modelos de cadenas de *Markov* (Taraghi et al., 2015) sirviendo como de *perfil de aprendizaje* para la *clasificación* jerárquica de aprendices. Otras propuestas basadas en el conocimiento y actividades de usuarios (Rodrıguez et al., 2015) contextualizadas conjuntamente con un enfoque multicriterio (factores de peso) se usan para recomendar contenido a profesores. Con base en estos enfoques se diseñó un cuestionario para recopilar información del perfil del profesor y utilizar este perfil posteriormente para la clasificación jerárquica de recursos y profesores. El cuestionario puede verse en las Figuras 3.2 y 3.3. Posteriormente, se diseñó un experimento con 39 profesores en etapa de iniciación profesional que en adelante consideraremos *usuarios nuevos*, los cuales se

Encuesta de Perfiles y Preferencias en Recursos Digitales (Forma C)

Información de Perfil del encuestado.

Estimado profesor, en esta sección, se le solicitan algunos datos de carácter general asociados a su perfil como profesor.

1. Asignatura (subsector) principal en que se desempeña Usted como profesor.

De ejercer docencia en más de un sector puede seleccionar más de una alternativa.

- [ ] Lenguaje y Comunicación
- [ ] Educación Matemática
- [ ] Educación Artística
- [ ] Educación Tecnológica
- [ ] Estudio y Comprensión de la Naturaleza
- [ ] Estudio y Comprensión de la Sociedad
- [ ] Comprensión de Medio Natural Social y Cultural
- [ ] Educación Física
- [ ] Idiomas (Inglés/Francés)
- [ ] Otro

2. Nivel educacional donde ejerce principalmente su docencia.

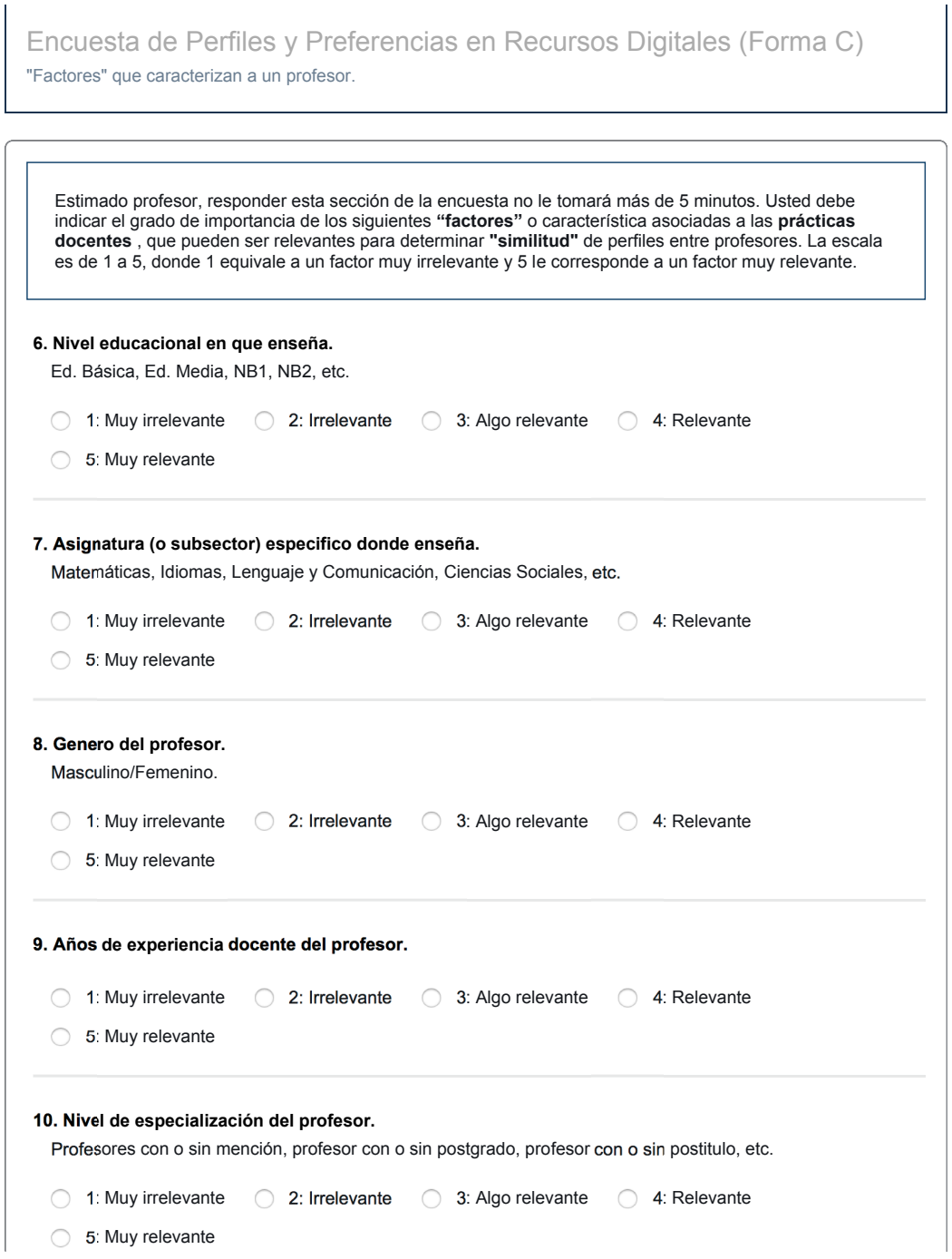
De ejercer docencia en más de un nivel puede marcar más de una alternativa.

- [ ] NB1 (1er y 2do Básico)
- [ ] NB2 (3ro y 4to Básico)
- [ ] NB3 (5to Básico)
- [ ] NB4 (6to Básico)
- [ ] NB5 (7mo Básico)
- [ ] NB6 (8vo Básico)

3. Tipo de administración del establecimiento donde ejerce su docencia.

Si trabaja en más de un colegio y estos tienen distinto tipo de administración puede marcar más de una alternativa.

FIGURA 3.2. Encuesta de *perfil curricular* preguntas 1, 2 y 3



Encuesta de Perfiles y Preferencias en Recursos Digitales (Forma C)

"Factores" que caracterizan a un profesor.

Estimado profesor, responder esta sección de la encuesta no le tomará más de 5 minutos. Usted debe indicar el grado de importancia de los siguientes **"factores"** o característica asociadas a las **prácticas docentes** , que pueden ser relevantes para determinar **"similitud"** de perfiles entre profesores. La escala es de 1 a 5, donde 1 equivale a un factor muy irrelevante y 5 le corresponde a un factor muy relevante.

6. Nivel educacional en que enseña.

Ed. Básica, Ed. Media, NB1, NB2, etc.

1: Muy irrelevante 2: Irrelevante 3: Algo relevante 4: Relevante 5: Muy relevante

7. Asignatura (o subsector) especifico donde enseña.

Matemáticas, Idiomas, Lenguaje y Comunicación, Ciencias Sociales, etc.

1: Muy irrelevante 2: Irrelevante 3: Algo relevante 4: Relevante 5: Muy relevante

8. Genero del profesor.

Masculino/Femenino.

1: Muy irrelevante 2: Irrelevante 3: Algo relevante 4: Relevante 5: Muy relevante

9. Años de experiencia docente del profesor.

1: Muy irrelevante 2: Irrelevante 3: Algo relevante 4: Relevante 5: Muy relevante

10. Nivel de especialización del profesor.

Profesores con o sin mención, profesor con o sin postgrado, profesor con o sin postitulo, etc.

1: Muy irrelevante 2: Irrelevante 3: Algo relevante 4: Relevante 5: Muy relevante

FIGURA 3.3. Encuesta de *perfil curricular* preguntas de 6, 7, 8, 9 y 10.

TABLA 3.7. Un ejemplo de los primeros 5 términos de los clusters $\mathscr{C}_1$ y $\mathscr{C}_2$. Los términos están acompañados de su peso y se clasifican en las sub-categorías de , *Asignatura*, *Nivel Curricular* y *Otros términos*.

Metadata Curricular			
Cluster	Asignatura	Nivel Curricular	Otros términos
$\mathscr{C}_1$	comunicación: 12.8 historia: 4.5 ciencia: 4.5 social: 4.2 lenguaje: 4	5^{to} básico: 12.4 1^{ero} medio: 9 8^{vo} básico: 7.9 6^{to} básico: 5.9 3^{ero} básico: 2	lectura: 2.7 ambiente: 2.1 verbal: 2 interacción : 1.7 organismo: 1.5
$\mathscr{C}_2$	ciencia: 7.3 ambiente: 6.1 interacción: 4.8 función: 3.6 biología: 2.4	1^{ero} medio: 11.9 5^{to} básico: 6.8 2^{do} medio: 2.7 8^{vo} básico: 2.6 1^{ero} básico: 2.2	naturales: 8.5 química: 4.6 organismo: 4.4 organismo-vivo: 4.2 estructura: 4.2
Metadada Socio-Pedagógica			
$\mathscr{C}_1$	languaje: 0.23 matemáticas: 0.19 historia: 0.18 comunicación: 0.05 musica: 0.04	básico: 0.02	tipo: 53.55 sitio: 21.22 site_type: 8.47 educarchile: 3.69 artículo: 3.66
$\mathscr{C}_2$	humano: 0.35 química: 0.34 física: 0.34 artes: 0.16 biología: 0.1		tipo: 54.55 sitio: 17.17 site_type: 8.24 artículo: 4.16 educarchile: 2.93
Contenido			
$\mathscr{C}_1$	educación media: 19.94 científico-humanista: 10.90 técnico-profesional: 7.55	2^{do} básico: 6.99 1^{ero} básico: 2.23	estududiante: 36.35 educación: 13.07 texto: 1.68 actividad: 0.49 artículo: 0.34
$\mathscr{C}_2$	educación media: 22.04 científico-humanista: 13.23 técnico-profesional: 9.60	2^{do} básico: 3.88 1^{ero} básico: 1.76	educación: 17.65 estudiante: 29.34 texto: 1.89 profesor: 0.19 actividad: 0.28

desempeñaban en educación básica. Los profesores se especializan en matemáticas (7 profesores); lenguajes y comunicación (7 profesores); lenguaje y matemáticas (8 profesores) y de más de dos áreas de desempeño (18 profesores).

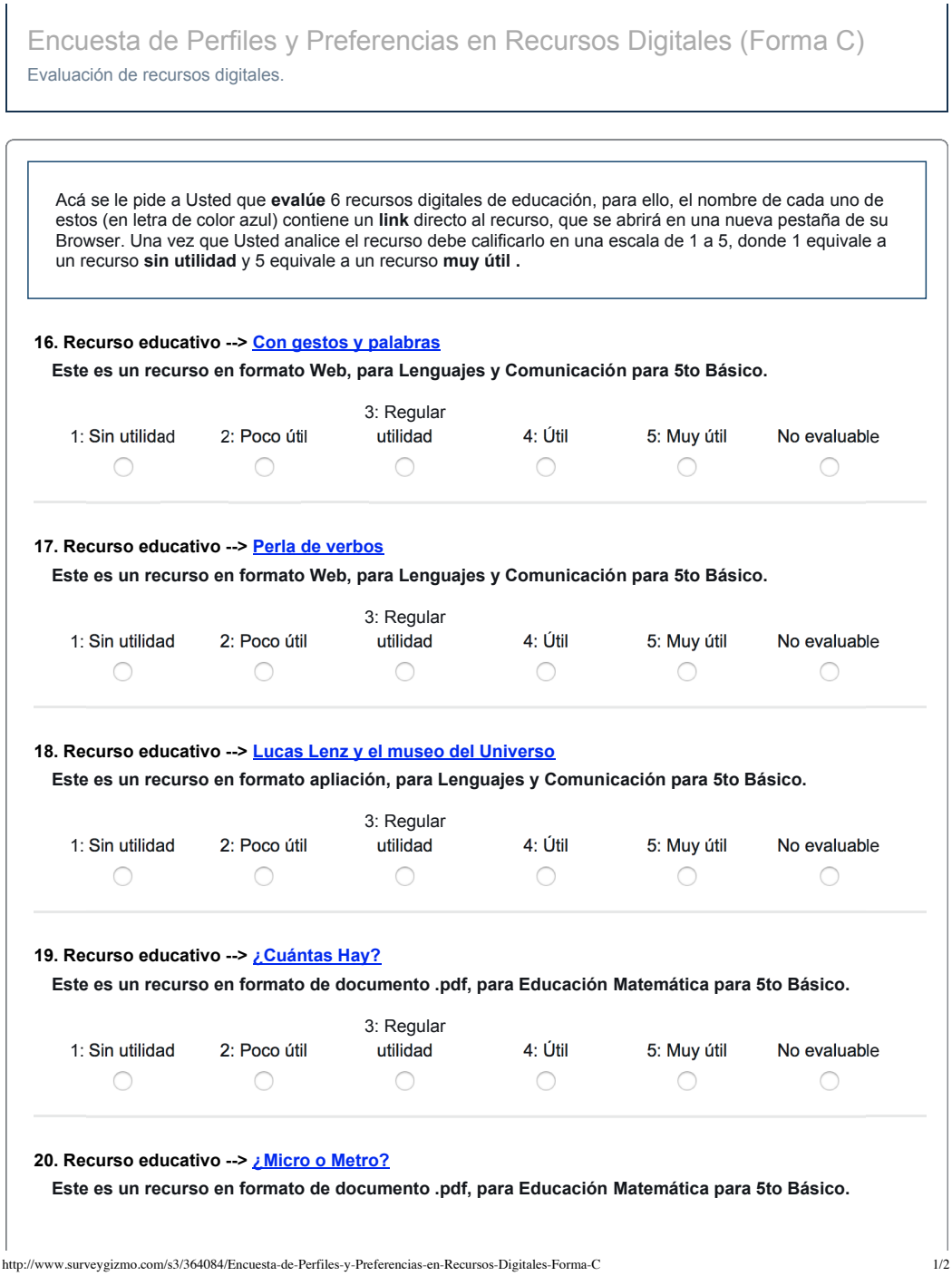
TABLA 3.8. Características del Perfil de Profesor

Categoría	Ejemplo
Asignatura/Disciplina	Matematicas, Languaje, Ciencias ...
Grado/Nivel	1^{ero} básico, 1^{ero} medio, ...
Tipo de Colegio	Público, particular o subvencionado
Ubicación geográfica del establecimiento	Urbana o rural

Siguiendo una metodología similar al trabajo de Manouselis (Manouselis et al., 2010), se pidió a los 39 profesores evaluar un subconjunto de 6 ítems, de un conjunto de 22 LOs disponibles en el portal *EducarChile*, que no fueron previamente visitado por cualquiera de estos usuarios (recursos nuevos). Los ítems correspondían a las asignaturas de *Matemáticas* (11) y *Lenguaje y Comunicación*. 4 ítems eran obligatorios evaluar por todos los participantes (se asignaron a los 39 nuevos usuarios), mientras que los elementos restantes (18) fueron asignados al azar. Se asignaron los LOs de esta manera ya que el número de participantes era pequeño y queríamos garantizar la existencia de más de una evaluación de al menos el 18 % del set de datos de este este experimento. Los recursos de aprendizaje asignados y los ítems evaluados se muestran en la Tabla 3.10.

Se pre-validó el diseño del experimento con expertos en educación y se encontró que para los profesores era más natural evaluar los recursos en una escala más amplia que la escala de 0 a 1 (es decir, me gustaría visitar el recurso Si/No), ya que una escala con más valores permite evaluar mejor recursos donde los profesores tienen maś experiencia. Hay recursos (ítems) que se vuelven interesantes, y otros que podrían ser considerados como malos recursos. Por esta razón definimos una escala de 5 puntos, es esto, 1 para ítem *inútil* hasta 5 para un ítem *muy útil* (ver la Tabla 3.9), con el fin de capturar la percepción de los profesores. Después definimos una escala de transformación para homologar las diferentes escalas (1 a 5 de este experimento, con 0 a 1 del dataset), lo que se explica en la Sección 3.3.2. La evaluación se hizo siguiendo el formato de cuestionario de la Figura 3.4.

Finalmente, se diseño e implementó un sistema de recomendación de *filtrado colaborativo* considerando los datos de la encuesta realizada. Dividimos el dataset obtenido en este experimento en dos subconjuntos, uno de 76 evaluaciones sobre dos objetos comunes



Encuesta de Perfiles y Preferencias en Recursos Digitales (Forma C)

Evaluación de recursos digitales.

Acá se le pide a Usted que **evalúe** 6 recursos digitales de educación, para ello, el nombre de cada uno de estos (en letra de color azul) contiene un **link** directo al recurso, que se abrirá en una nueva pestaña de su Browser. Una vez que Usted analice el recurso debe calificarlo en una escala de 1 a 5, donde 1 equivale a un recurso **sin utilidad** y 5 equivale a un recurso **muy útil .**

16. Recurso educativo --> Con gestos y palabras

Este es un recurso en formato Web, para Lenguajes y Comunicación para 5to Básico.

1: Sin utilidad	2: Poco útil	3: Regular utilidad	4: Útil	5: Muy útil	No evaluable
○	○	○	○	○	○

17. Recurso educativo --> Perla de verbos

Este es un recurso en formato Web, para Lenguajes y Comunicación para 5to Básico.

1: Sin utilidad	2: Poco útil	3: Regular utilidad	4: Útil	5: Muy útil	No evaluable
○	○	○	○	○	○

18. Recurso educativo --> Lucas Lenz y el museo del Universo

Este es un recurso en formato apliación, para Lenguajes y Comunicación para 5to Básico.

1: Sin utilidad	2: Poco útil	3: Regular utilidad	4: Útil	5: Muy útil	No evaluable
○	○	○	○	○	○

19. Recurso educativo --> ¿Cuántas Hay?

Este es un recurso en formato de documento .pdf, para Educación Matemática para 5to Básico.

1: Sin utilidad	2: Poco útil	3: Regular utilidad	4: Útil	5: Muy útil	No evaluable
○	○	○	○	○	○

20. Recurso educativo --> ¿Micro o Metro?

Este es un recurso en formato de documento .pdf, para Educación Matemática para 5to Básico.

FIGURA 3.4. Encuesta de valoración de ítems, 6 recursos a valorar.

TABLA 3.9. Escala de evaluación usuarios que evalúan nuevos recursos.

Ranking	Significado
1	Inútil
2	Poco útil
3	De regular utilidad
4	Util
5	Muy útil

TABLA 3.10. Número de evaluaciones por recursos de aprendizaje.

Recurso	Sector	Número de asignaciones	Número de evaluaciones	Recurso	Sector	Número de asignaciones	Número de evaluaciones
1	Lenguaje	39	39	12	Matemáticas	39	37
2	Lenguaje	39	38	13	Matemáticas	39	39
3	Lenguaje	4	4	14	Matemáticas	4	4
4	Lenguaje	5	5	15	Matemáticas	5	5
5	Lenguaje	4	3	16	Matemáticas	4	4
6	Lenguaje	3	3	17	Matemáticas	3	3
7	Lenguaje	5	5	18	Matemáticas	6	6
8	Lenguaje	4	2	19	Matemáticas	5	5
9	Lenguaje	2	2	20	Matemáticas	2	2
10	Lenguaje	2	2	21	Matemáticas	2	1
11	Lenguaje	2	2	22	Matemáticas	8	8

TABLA 3.11. Resultados MAE fase 2, para similitud con *Pearson, Manhattan* y *Spearman*. Por la cantidad de estadísticas no fue necesario usar otras métricas.

Ranking	Versión Similitud	MAE
1^{ero}	Pearson	0.970598657
2^{do}	Spearman	0.997978555
3^{ero}	Manhattan	1.006772909

(recordar que los usuarios evaluaron 4 objetos en común) y el otro conteniendo las 149 evaluaciones restantes. Usamos el primer grupo para hacer predicciones y el segundo como conjunto de validación (*gold standard*), con el fin de determinar el impacto del error de los ítems a recomendar (MAE (J. L. Herlocker et al., 2004)).

Los resultados del experimento no son alentadoras, en contraste con el experimento de *Manouselis* (Manouselis y Costopoulou, 2008; Manouselis et al., 2010). La Tabla 3.11 así lo ilustra. Ello se debe principalmente a que las evaluaciones son pocas. El tipo de encuestados permitió establecer claramente tres grupos de usuarios con gustos similares, y no todos evaluaron la misma cantidad de ítems.

3.3. Fase 3: Objetivo O5

Para validar la relevancia de la metadata de recursos, y la identificación de comunidades de interés se utilizará una tarea avanzada en los sistemas de recomendación: resolver el problema del *arranque en frío* o *cold start*. Con el fin de enfrentar el problema de arranque en frío, se considerarán los clusters jerárquicos identificados en la Tabla 3.6 y los vectores de términos (palabras) calculados con la ecuación 3.6.

Los términos del vector que representa a cada cluster son extraídos de los metadatos asociados a los recursos de aprendizaje visitados por los usuarios del cluster. Los *nuevos profesores* se describen también a través de un vector de términos, que se obtienen de su perfil. Con ambos vectores de términos (del *nuevo profesor* y del cluster) se determia el cluster asociable al *nuevo profesor*. También se genera un vector de términos para un *nuevo recurso* a ser recomendado a este *nuevo profesor* y se compara este vector con el vector de términos asociado a los ítems del cluster al cual pertenecería el *nuevo profesor*. Los ítems más similares del cluster al que el *nuevo profesor* pertenece, se utilizan para predecir la evaluación del *nuevo recurso*.

A continuación se presenta el detalle y la formalización de este enfoque, así como un experimento que considera los profesores de la Fase 2. Comparamos el resultado de este enfoque con el resultado de la técnica de filtrado colaborativo para el grupo de *nuevos profesores* y encontramos que este enfoque mejora notablemente la recomendación.

3.3.1. Clasificando nuevos Usuarios

Para clasificar, nuevos usuarios usaremos la siguiente notación:

a.- $\mathscr{U}_N$: Conjunto de usuarios nuevos, esto es, usuarios de los que se conoce solo su *perfil curricular*, pero no tienen ninguna preferencia sobre ítems.

b.- $\mathscr{U}_{DS}$: Conjunto de todos los usuarios que contiene el *dataset*.

c.- $\mathscr{I}_N$: conjunto de nuevos ítems

d.- $\mathscr{I}_{DS}$: conjunto de ítems del *dataset* que tiene alguna evaluación realizada por los usuarios de $\mathscr{U}_{DS}$

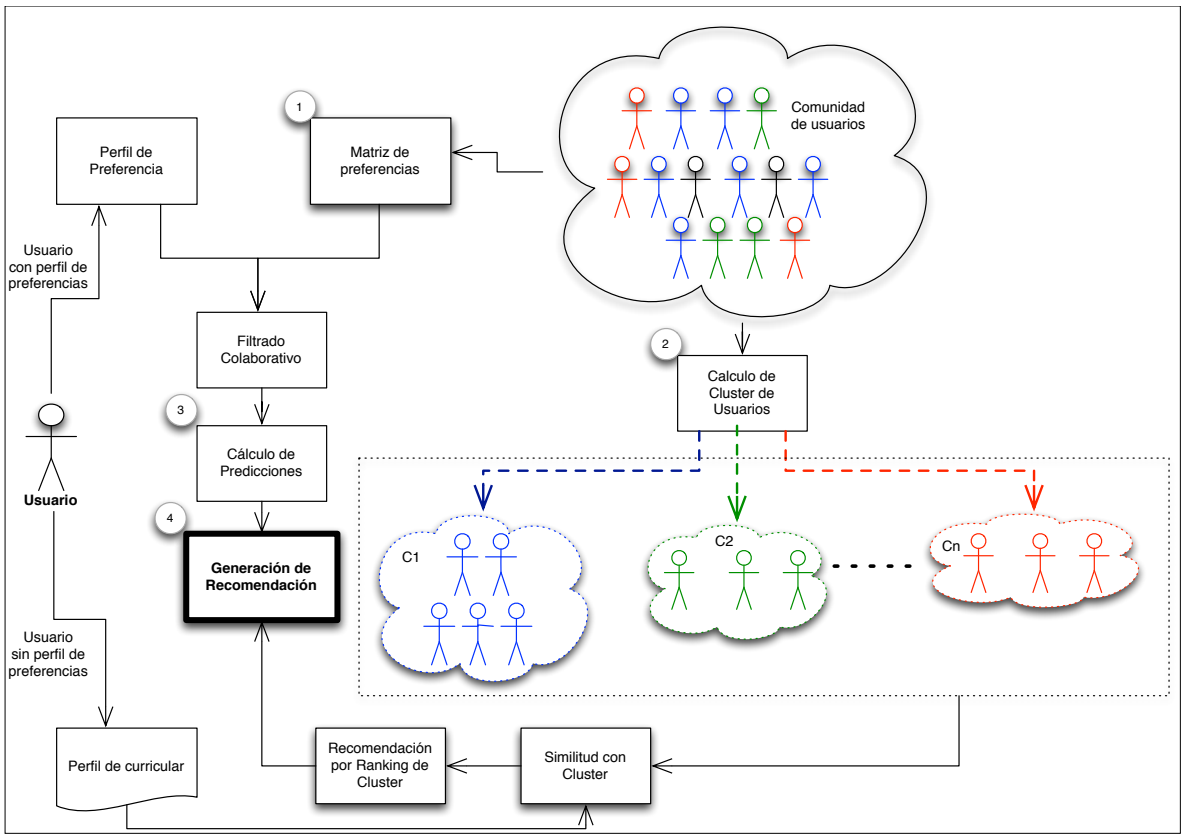


FIGURA 3.5. Propuesta de sistema de recomendación extendido, que contempla una solución al problema de arranque en frío.

e.- $\mathscr{T}$: El conjunto de términos contenidos en el perfil de profesor y los ítems del dataset diferenciados por categoría.

Cada cluster $\mathscr{C}$, definido en la sección 3.1.5, se representa como un vector de términos ponderados $\overrightarrow{\mathscr{C}} = \langle w(t_1), w(t_2), \ldots, w(t_n) \rangle$, donde $w(t_i)$ se calcula por la ecuación 3.6. Para cada nuevo usuario $nu \in \mathscr{U}_N$, un vector de términos ponderados, según el *perfil curricular* del profesor, se define como $\overrightarrow{nu} = \langle w_u(t_1), w_u(t_2), \ldots, w_u(t_n) \rangle$, donde $w_u(t_i)$ se define también por la ecuación 3.6, para el contenido del *perfil curricular* del usuario. La caracterización de los usuarios se ha propuesto para otros fines que enfrenta el problema de arranque en frío. Por ejemplo, Drachsler propone el uso de estereotipos para los estudiantes (Drachsler et al., 2010), que son categorías basadas en los datos demográficos y de LO.

Indicadores como la diversidad y la similitud de interés de usuarios en un grupo o entre-grupo se utilizan para determinar grupos de aprendizaje (Dascalu et al., 2014). Alumnos se agrupan automáticamente, por elección de los profesores, o de acuerdo con un algoritmo especializado basado en dichos indicadores. La diversidad y la similitud se determinan a partir de un perfil de aprendizaje (intereses explícitos de los usuarios) y de su historia durante un proceso de aprendizaje.

La similitud entre el usuario nuevos y usuarios con perfil de preferencias, clasificados en sus correspondientes cluters $\mathscr{C}$, se calcula utilizando la ecuación de similitud de Coseno. Las tuplas resultantes, ecuación 3.8, permiten definir el conjunto que denominamos *top-3 cluster*, los tres cluster más similares para un nuevo usuario que denotamos por $\mathscr{C}^*_{nu}$ y se definen en la ecuación (3.9).

$$\{(\mathscr{C}_1, sim(\overrightarrow{nu}, \overrightarrow{\mathscr{C}}_1)), (\mathscr{C}_2, sim(\overrightarrow{nu}, \overrightarrow{\mathscr{C}}_2)), \ldots, (\mathscr{C}_m, sim(\overrightarrow{nu}, \overrightarrow{\mathscr{C}}_m))\} \tag{3.8}$$

$$\mathscr{C}^*_{nu} = \{\mathscr{C}_{k=a,b,c} \mid sim(\overrightarrow{nu}, \overrightarrow{\mathscr{C}}_k) > sim(\overrightarrow{nu}, \overrightarrow{\mathscr{C}}_j),\ \forall j \neq k\} \tag{3.9}$$

3.3.2. Predicción de *rating* para nuevos ítems

Para cada nuevo ítem $ni \in \mathscr{I}_N$ encontraremos los ítems de $i_v \in \mathscr{I}_{\mathscr{C}}$ más similares, que han sido visitados por los usuarios $v \in \mathscr{C}$. Utilizamos diferentes umbrales de similitud, consistentes en un porcentaje p de términos compartidos por los ítems similares i_v y ni, este conjunto de ítems parecidos lo denotaremos por $\mathscr{I}^p_{\mathscr{C},ni}$, y estará definido como el conjunto $\{i_v \in \mathscr{C} \mid$ número de términos de $i_v > p \cdot ($número de términos de $ni)\}$. Luego, para cada nuevo ítem (ni) se calcula el *rating* $R_{\mathscr{C},ni}$, determinado a partir de los usuarios con perfil de preferencias, tal como se define en la ecuación 3.10.

$$R_{\mathscr{C},ni} = \frac{\sum\limits_{\substack{v \in \mathscr{U}_{\mathscr{C}} \\ i_v \in \mathscr{I}^p_{\mathscr{C},ni}}} r_{v,i_v}}{|\mathscr{I}^p_{\mathscr{C},ni}|} \tag{3.10}$$

Finalmente, considerando sólo *top-3 cluster* ($\mathscr{C}^*_{nu}$), promediamos los *ratings* (calificaciones), otorgadas por los usuarios de estos cluster (3.11).

$$R_{\mathscr{C}^*_{nu,i}} = avg(R_{\mathscr{C},i}),\ \forall\, \mathscr{C} \in \mathscr{C}^*_{nu} \tag{3.11}$$

Puesto que votos de nuevos usuarios (nu) están en escala de 1 a 5, y se diferencian de los criterios de calificación del *dataset*, visitas de usuario, generalmente 0 o 1 visita por recurso para cada usuario, definimos una regla de equivalencia. Por lo tanto, para un usuario $nu \in \mathscr{U}_N$, que evalúa un nuevo ítem ni con nota $r_{nu,ni}$, la regla o formula de equivalencia será $e(r_{nu,ni})$ donde e esta definida por la ecuación 3.12.

$$e(r_{nu,ni}) = \begin{cases} 0 & \text{, if} \quad r_{nu,ni} = 1 \vee r_{nu,ni} = 2 \\ 0{,}5 & \text{, if} \quad r_{nu,ni} = 3 \\ 1 & \text{, if} \quad r_{nu,ni} = 4 \vee r_{nu,ni} = 5 \end{cases} \tag{3.12}$$

Con el fin de determinar el efecto de nuestro enfoque, implementamos una estrategia de filtrado colaborativo, teniendo en cuenta sólo el conjunto de nuevos usuarios, para determinar el impacto del error MAE (J. L. Herlocker et al., 2004), de los ítems recomendados. La Tabla 3.12 muestra los resultados, en base a cuatro métricas de similitud diferentes. Podemos observar, resultados muy pobres, y probablemente es debido al tamaño de la muestra.

TABLA 3.12. Resultados de MAE cuando recomendamos recursos basados en filtrado colaborativo y datos de nuevos usuarios, usando distintas métricas de similitud.

Ranking	Métrica de similaridad	MAE
1^{ero}	Spearman	1,227935814
2^{do}	Pearson	1,44987662
3^{ero}	Manhattan	1,472972216
4^{to}	Distancia Euclidiana	1,507429683

Determinamos los ratings o votos $R_{\mathscr{C}^*_{nu},i}$ para los ítems $ni \in \mathscr{I}_N$ que fueron evaluados por los nuevos usuarios, $nu \in \mathscr{U}_N$, teniendo en cuenta la regla de equivalencia e y el *perfil curricular* de los nuevos profesores, como se describió en la sección 3.3.1. Se determinó MAE para cada categoría usando la regla de equivalencia.

La tabla 3.13 presenta nuestros resultados. Los votos dados a los ítems los dividimos en tres categorías, ≥ 1, ≥ 3 y ≥ 4. En la tabla, las columnas de porcentaje, indican porcentaje de palabras que comparten los nuevos ítems ni, con los ítems i_v ya visitados del cluster. La "similaridad" entre ítems, la consideramos en base a los términos que comparten. A menor porcentaje (umbral) de palabras a compartir se encuentran más ítems similares. De la tabla podemos deducir las siguientes conclusiones:

- Para metadatos de *socio-pedagógicos* se encuentra ítems similares con umbral sólo hasta el 30 % de palabras, es decir no hay objetos similares que compartan más del 30 % de las palabras.
- Para metadatos *contenido* es posible encontrar objetos similares que cumplen con sólo un umbral del 20 %, menor al de metadatos *socio-pedagógicos* .
- Para metadatos *curriculares* se encuentran objetos similares compartiendo hasta un 80 % de los terminos.
- Hay mayor precisión de MAE, para aquellos ítems cuyo voto fue 4 o 5.

3.4. Fase 4: Objetivo O1

En el capítulo 3 logramos determinar comunidades, grupos o cluster de interés. Es más, para los usuarios del *dataset* que compartían gustos similares asociamos los términos más representativos. A modo de ejemplo, para el cluster $\mathscr{C}_1$ que agrupa dos vecindades con $p = 85\,\%$ o más de vecinos en común (tabla 3.6), presentamos los pesos de sus metadatos en la tabla 3.7. Estos metadatos se pueden visualizar en la figura 3.6. Claramente para *contenido curricular*, el cluster mayoritariamente manifiesta interés por la asignatura de *Lenguaje y Comunicación* (metadatos de asignatura), y en cuanto al nivel donde ocupan estos recursos es 5^{to} *Básico* (metadatos de nivel). En esta fase, profundizaremos el análisis

TABLA 3.13. Resultados obtenidos de MAE para la predicción de voto considerando diversos umbrales de similitud.

Categoría de Metadata	Voto dado por nu a ni	Umbrales de Similaridad						
		20 %	30 %	40 %	**50 %**	60 %	70 %	80 %
Curricular	>=4	0.4974	0.4748	0.4775	**0.4723**	0.4936	0.4814	
	>=3	0.6100	0.5936	0.5910	**0.5776**	0.5988	0.6085	
	>=1	1.0133	0.9886	0.9930	**0.9825**	1.0105	1.1638	1.4606
Socio-Pedagógicos	>=4	0.5546	0.4139					
	>=3	0.6906	0.5773					
	>=1	1.1431	0.8915					
Contenido	>=4	0.5440						
	>=3	0.7019						
	>=1	1.0784						

de los metadatos asociados a las comunidades de interés y centraremos nuestro análisis en los metadatos curriculares, pues son aquellos que han permitido mejores resultados en el problema de arranque en frío.

Como se indico en la sección 3.1.2 el sistema escolar en Chile, esta dividido en educación pre-escolar, primaria, secundaria y terciaria. Nuestro interés está en educación primaria (o básica) y educación secundaria (o media). En los ocho niveles en educación básica (1^{ero} a 8^{vo} básico), los alumnos cursan[5] asignaturas de *Artes Visuales, Ciencias Naturales, Educación Física y Salud, Historia, Geografía y Ciencias Sociales, Inglés, Lenguaje y Comunicación, Matemática, Música, Orientación, Tecnología* y *Lengua indígena* como se aprecia en la Tabla 3.14. Al llegar los alumnos a enseñanza media, que contempla 4 años de estudio, pueden optar por *educación científico humanista* (*c-h*) o *educación técnico profesional* (*t-p*). La educación *c-h* contempla las asignaturas de *Artes Musicales, Artes Visuales, Educación Física, Filosofía y Psicología* (sólo para 3^{ero} y 4^{to} medio), *Historia-Geografía y Ciencias Sociales, Inglés, Lenguaje y Comunicación, Matemática, Educación Tecnológica* (sólo para 1^{ero} y 2^{do} medio) y *Ciencias Naturales* (*Biología, Física*, y *Química*), la Tabla

[5]El detalle de los contenidos de las asignaturas se puede encontrar en la página del ministerio `http://www.curriculumenlineamineduc.cl/605/w3-channel.html`.

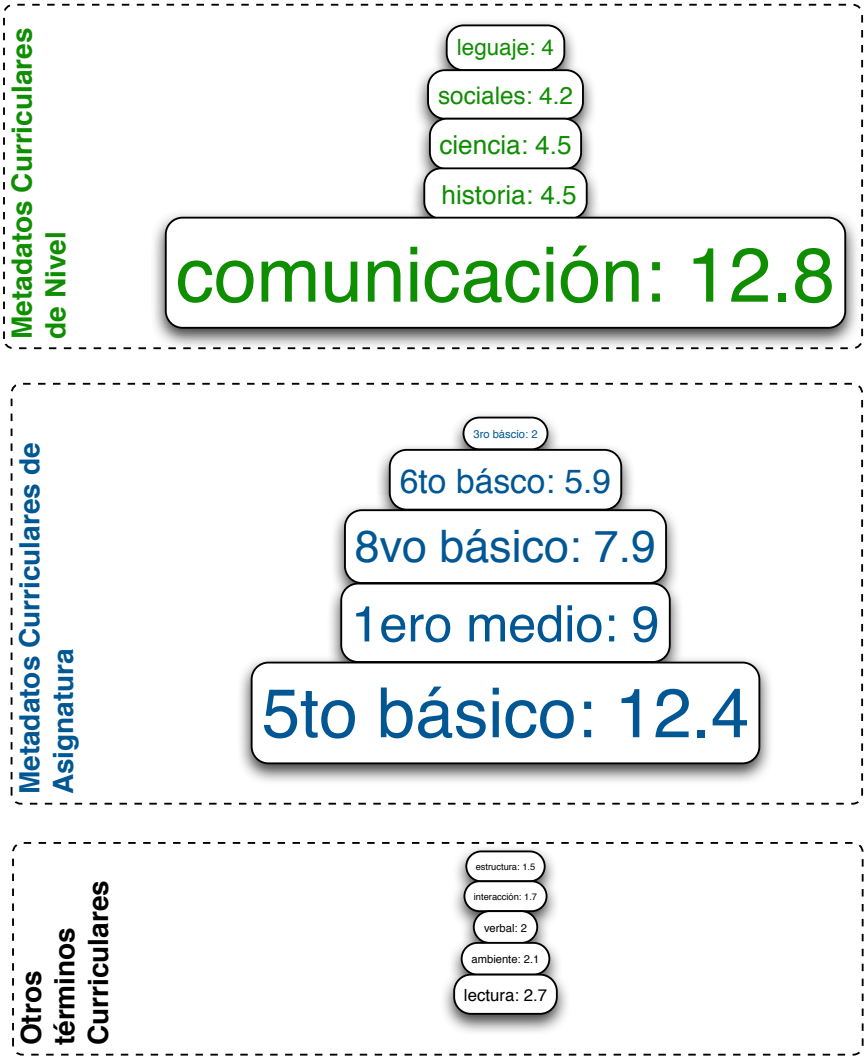


FIGURA 3.6. Vista de los metadatos como nube de tags para los metadatos de la categoría Metadata Curricula del cluster $\mathscr{C}_1$.

3.15 resume esta información. La educación *t-p* tiene un tronco en común con la enseñanza *c-h*, aunque el foco está en la formación técnica, por lo que no detallamos asignaturas debido a la diversidad de alternativas que hay.

TABLA 3.14. Asignaturas para educación básica

Asignatura	Niveles que abarca
Artes Visuales	1^{ero} a 8^{to} Básico
Ciencias Naturales	1^{ero} a 8^{to} Básico
Educación Fśica y Salud	1^{ero} a 8^{to} Básico
Historia, Geografía y Ciencias Sociales	1^{ero} a 8^{to} Básico
Inglés	5^{to} a 8^{to} Básico
Lenguaje y Comunicación	1^{ero} a 8^{to} Básico
Matemáticas	1^{ero} a 8^{to} Básico
Música	1^{ero} a 8^{to} Básico
Orientación	1^{ero} a 8^{to} Básico
Tecnología	1^{ero} a 8^{to} Básico

TABLA 3.15. Asignaturas para educación media *c-h*

Asignatura	Niveles que abarca
Artes Musicales	1^{ero} a 4^{to} Medio
Artes Visuales	1^{ero} a 4^{to} Medio
Educación Física	1^{ero} a 4^{to} Medio
Filosofía y Psicología	3^{ero} y 4^{to} Medio
Historia, Geografía y Ciencias Sociales	1^{ero} a 4^{to} Medio
Inglés	1^{ero} a 4^{to} Medio
Lenguaje y Comunicación	1^{ero} a 4^{to} Medio
Matemáticas	1^{ero} a 4^{to} Medio
Educación Tecnológica	1^{ero} y 2^{do} Medio
Ciencias Naturales Biología	1^{ero} a 4^{to} Medio
Ciencias Naturales Física	1^{ero} a 4^{to} Medio
Ciencias Naturales Química	1^{ero} a 4^{to} Medio

La información de temas y contenidos de las asignaturas, las podemos obtener del *Ministerio de Educación* a través de la página Web para temas curriculares[6]. Los detalles de contenido de las asignaturas son importante para nuestro trabajo y en especial en esta sección ya que queremos entender la naturaleza de los metadatos relevantes de las

[6] `http://www.curriculumenlineamineduc.cl/`

comunidades de profesores identificadas. Por ejemplo, a partir de las *Bases Curriculares y Programas de Estudios* disponibles en el sitio Web del Ministerio podemos analizar las palabras que están asociadas a una asignatura. La tabla 3.16 muestra un ejemplo de palabras contenidas en las descripciones de la asignatura de *matemática* para 1^{ero} básico en las 4 unidades que contempla esta asignatura para el nivel (1^{ero} básico). Para efectos de este análisis se creó un diccionario de estas palabras para cada asignatura.

TABLA 3.16. Palabras que describen la asignatura de matemáticas para 1^{ero} básico.

Unidades	Palabras/Terminos
1^{era} Unidad	números, contar, ordenar, patrón, igualdad, largo, corto, bajo, alto, fechas.
2^{da} Unidad	números hasta 100, unidades, decenas, figuras 3D y 2D.
3^{era} Unidad	conteo hacia adelante, conteo hacia atrás, restar, cálculo mental, preguntas, tablas de conteo, pictogramas.
4^{ta} Unidad	líneas rectas, líneas curvas, pictogramas.

A continuación procederemos a analizar los metadatos de los recursos de aprendizaje asociados a los clusters $\mathscr{C}$ identificados en la tabla 3.6. Nótese que se produjeron 1.376 clusters diferentes, por lo que para manejar esta cantidad de grupos hemos realizado un análisis en base a los términos de su representación vectorial ($\overrightarrow{\mathscr{C}}$), que fué presentado en la sección 3.1.6. En la tabla 3.7 se presentó un ejemplo de este análisis para dos *clusters*. Este análisis parte agrupando los clusters en una jerarquía adicional, donde se buscan clusters similares pero no en función de los usuarios comunes sino en función de los términos o palabras que describen los recursos de estas comunidades.

Así, vamos a definir un vocabulario V como el conjunto de términos ponderados que representan la temática de un grupo de *clusters* similares. Dos *clusters* $\mathscr{C}_i$ y $\mathscr{C}_j$ son similares de acuerdo a sus términos si y sólo si, coinciden en un porcentaje p en los top n términos de mayor peso en $\overrightarrow{\mathscr{C}_i}$ y $\overrightarrow{\mathscr{C}_j}$. El peso ponderado de cada término se calcula sumando los pesos de cada ocurrencia del término en los *clusters* similares, y dividiendo la suma entre el número total de palabras de los clusters. En la tabla 3.7 se presenta un análisis de vocabularios para diferentes valores de p. Podemos notar que para un valor de $p = 60\,\%$, se encuentra el menor número de vocabularios compartidos (34) y el menor número (29) de vocabularios

que representan a clusters únicos (no son compartidos por otros clusters). Por esta razón consideraremos este valor de p para el resto del análisis.

Umbral de coincidencia													
90%		**85%**		**80%**		**75%**		**70%**		**65%**		**60%**	
Nro Clusters	**Nro Vocab.**	**Nro Clusters**	**Nro Vocab.**	**Nro Clusters**	**Nro Vocab.**	**Nro Clusters**	**Nro Vocab.**	**Nro Clusters**	**Nro Vocab.**	**Nro Clusters**	**Nro Vocab.**	**Nro Clusters**	**Nro Vocab.**
19	1	30	1	77	1	175	1	154	1	230	1	419	1
9	2	29	1	68	1	80	1	127	2	226	1	341	1
8	1	23	1	63	1	59	1	107	1	168	1	115	1
7	1	21	1	33	1	44	2	79	1	71	1	95	1
6	4	17	1	30	1	36	1	52	1	67	1	72	1
5	6	14	2	25	1	35	1	35	1	60	1	54	1
4	13	11	4	22	1	32	2	32	1	51	1	48	1
3	15	10	2	20	2	26	1	30	1	50	1	38	1
2	**91**	9	1	17	2	24	1	25	2	42	1	26	1
		8	4	16	1	22	3	23	3	38	1	15	3
		7	4	14	2	20	1	22	1	26	1	10	1
		6	8	13	3	19	2	17	3	23	1	9	1
		5	11	12	2	18	1	16	1	21	1	7	1
		4	13	10	2	17	1	15	1	20	1	6	3
		3	29	9	3	15	1	14	1	19	1	5	2
		2	**103**	8	4	12	3	13	1	15	1	4	4
				7	6	11	1	12	1	14	1	3	4
				6	3	10	3	11	2	13	1	**2**	**6**
				5	13	9	1	10	1	10	2		
				4	13	8	2	9	2	9	1		
				3	34	7	5	8	1	8	3		
				2	**86**	6	6	7	6	7	3		
						5	8	6	3	6	3		
						4	14	5	6	5	4		
						3	25	4	8	4	6		
						2	**55**	3	17	3	7		
								2	**30**	**2**	**13**		
	134		**186**		**183**		**143**		**99**		**60**		**34**
Total Vocabularios compartidos													
1	991	1	647	1	347	1	161	1	80	1	39	1	29
Total Vocabularios únicos													

FIGURA 3.7. Total de Vocabularios obtenidos para diversos valores de coincidencia de palabras del 60 % al 90 %

La figura 3.8 presenta un ejemplo del Vocabulario 12. La primera fila presenta los identificadores de los clusters que comparten este vocabulario. La primera columna presenta los

términos, la segunda el número de clusters similares, la tercera la suma de los pesos para el término por cluster, y la cuarta columna el peso promedio de los términos.

Vocabulario 12			
[66, 89, 792, 1032]			
Palabras	Clusters	Peso total	Peso promedio
chile	4	5.64101	1.41025
forma	4	5.27447	1.31862
8B basico	4	5.17280	1.29320
3B basico	4	4.74072	1.18518
4B medio	4	4.71724	1.17931
sonoro	4	4.52370	1.13093
acercamiento	4	4.45883	1.11471
sonido	4	4.27932	1.06983
region	4	3.99140	0.99785
ambiente	3	7.62196	2.54065
recurso	3	5.41147	1.80382
entorno	3	4.19594	1.39865
6B basico	3	3.92402	1.30801
escenico	3	3.81454	1.27151
cine	2	4.54940	2.27470
video	2	4.48104	2.24052
aviso	2	4.44766	2.22383
publicitario	2	4.44766	2.22383
natural	2	3.82947	1.91473
espacio	2	2.63385	1.31692

FIGURA 3.8. Los 20 términos top para el Vocabulario 12 ordenados por el Peso promedio

La figura 3.9 muestra un fragmento de estos vocabularios, presentando los top-20 términos más relevantes de 5 vocabularios incluyendo el peso promedio de cada palabra. Las palabras han sido coloreadas de acuerdo a su pertenencia a alguna de las temáticas descritas en el diccionario ejemplificado en la Tabla 3.16. El mapa de colores se presenta en la figura 3.10. Se considera el color verde para la(s) asignatura de *Ciencias Naturales* (Biología, Química o Física), color azul para *Historia, Geografía y Ciencias Sociales*, etc. Los términos asociados al nivel (ej. 1^{1ero} básico) se han dejado sin colorear.

Vocabulario 1		Vocabulario 2		Vocabulario 3		Vocabulario 4		Vocabulario 5	
Palabras	Peso promedio	Palabras	Peso promedio	Palabras	Peso promedio	Palabras	Peso promedio	Palabras	Peso promedio
interaccion	4.7609	lectura	3.3128	8B basico	1.9262	estructura	3.5218	interaccion	4.2739
organismo	4.4321	escriturar	3.2636	america	3.0137	interaccion	3.4261	numero	3.9964
ambiente	4.2387	oral	2.8593	vision	2.6056	ser	3.4223	organismo	4.0368
lectura	4.1459	natural	2.1680	panoramico	2.5804	vivo	3.3685	ambiente	3.7921
estructura	5.0615	8B basico	1.7361	natural	2.3092	materia	2.9700	lectura	4.4258
ser	4.9620	2B medio	2.1521	4B medio	2.2996	6B basico	2.0032	estructura	4.4582
escriturar	4.0113	ambiente	4.4593	chile	2.1675	transformacion	3.2841	escriturar	4.0823
vivo	4.8502	numero	4.0379	6B basico	1.8859	lectura	5.6191	ser	4.2687
oral	3.6262	3B medio	1.7795	independencia	5.2270	organismo	2.9187	8B basico	1.9454
numero	4.2355	interaccion	5.5127	7B basico	1.8689	ambiente	2.7762	oral	3.8497
funcion	4.1637	7B basico	1.7508	ambiente	4.6879	3B medio	1.9242	vivo	4.1259
8B basico	2.0130	4B medio	2.0143	2B medio	2.1856	3B basico	1.8804	funcion	3.8262
7B basico	2.0302	organismo	5.3372	interaccion	5.5019	4B basico	1.9567	matematico	2.8575
2B medio	2.4085	6B basico	1.6789	lectura	4.1936	8B basico	1.8931	diversidad	5.6180
transformacion	5.9137	naturales	2.0962	organismo	4.9050	funcion	3.0542	2B medio	2.5618
materia	5.4795	matematico	2.4596	numero	4.3262	aplicar	2.3213	6B basico	1.9157
4B medio	2.5570	sociales	1.9421	escriturar	3.7557	fisico	2.9820	7B basico	1.8921
chile	2.7411	biologia	2.2131	4B basico	1.8647	4B medio	2.3499	cultural	3.2768
matematico	3.3718	4B basico	1.5284	oral	3.3196	numero	4.9003	3B medio	2.1589
natural	2.1943	social	1.5922	3B basico	1.6484	7B basico	1.7733	geometria	5.7529

FIGURA 3.9. Los 20 términos top para 5 *vocabularios*

FIGURA 3.10. Mapa de colores para la clasificación de términos

En la tabla 3.9 podemos apreciar que el Vocabulario 1 incluye 8 palabras asociadas con la asignatura de *Ciencias Naturales* y 4 asociadas con *Matemáticas*, lo que permite concluir que este grupo esta compuesto mayoritariamente por intereses en *Ciencias Naturales* y *Matemáticas*. Al comparar estos vocabularios en función de la temática que abarcan mayoritariamente, podemos obtener 9 grandes categorías como se puede ver en la Figura 3.11. Así, la primera categoría presenta 1 vocabulario conteniendo 20 términos que se refieren principalmente a *Artes Visuales* (6 términos), y Ciencias Naturales en segundo lugar (5 términos) considerando también los niveles de 8^{vo}, 3^{ero} y 6^{to} básico, así como 4^{to} medio. El análisis del resto de las categorías nos indica que en promedio el 54 % de los términos corresponden a conceptos curriculares relacionados con asignatura y nivel.

Categoría	Total Vocabularios	Total términos	Términos Asignatura 1	Niveles	Total términos asignatura	Total términos nivel	% términos asignatura y nivel
Artes Visuales	1	20	acercamiento, escenico, cine, video, aviso, publicitario	8vo basico, 3ero basico, 4to medio, 6to basico	**6**	**4**	75%
Ciencias Naturales			acercamiento,, ambiente, recurso, entorno, natural		5		
Historia, Geografía y Ciencias Sociales	1	20	contemporaneo, america, tierra, panoramico	2B medio, 6B basico, 4B medio, 4B basico, 8B basico, 3B basico, 1B basico, 7B basico	4	8	75%
Ciencias Naturales			natural, planeta, tierra		3		
Historia, Geografía y Ciencias Sociales	3	60	chile (3), america (3), social (3), sociales (2), panoramico (2)	2do medio (3), 4to medio (3), 3ero medio (2), 4to basico (2), 7mo basico (2), 8vo basico (2)	13	14	63%
Lenguaje y Comunicación			lectura (3), escriturar (3), oral (3), lengua (2)		11		
Ciencias Naturales	7	140	interaccion (7), organismo (6), estructura (6), ser (6), vivo (6), ambiente (5), materia (4), fisico (3)	4to medio (5), 3ero basico (5), 8vo basico (4)	43	14	59%
Matemáticas			numero (7), funcion (6), transformacion (6), matematico (6)		25		
Ciencias Naturales	7	140	interaccion (7), organismo (7), ambiente (6), materia (4), natural (3), quimico (3), estructura (3)	4to medio (6) 2domedio (5) 8vo basico (3) 3ero medio (3) 6to basico (3)	33	20	55%
Lenguaje y Comunicación			lectura (7), escriturar (7), oral (7), lengua (3)		24		
Lenguaje y Comunicación	4	80	lectura (4), escriturar (4), oral (4), argumentacion (3), texto (2)	2ero medio (4), 3ero medio (4), 4to medio (4), 4to basico (4), 1ero basico (4)	17	20	55%
Historia, Geografía y Ciencias Sociales			geografia (3), contemporaneo (2), chile (2)		7		
Ciencias Naturales	4	80	ambiente (4), interaccion (4), organismo (3), natural (3), vision (2), diversidad (2)	2do medio (4), 4to medio (4), 3ero medio (3), 8vo basico (3)	18	14	48%
Geografía y Ciencias Sociales			panoramico (2), america (2), sociales (2)		6		
Lenguaje y Comunicación	3	60	escriturar (3), lectura (3), oral (3), obra (1), ficticio (1), comunicar (1)	3ero medio (2), 7mo basico (2), 4to medio (2), 4to basico (2)	12	8	47%
Ciencias Naturales			natural (2), interaccion (2), entorno (2), modelo (2)		8		
Lenguaje y Comunicación	4	80	escriturar (4), lectura (4), numero (3), oral (2), literario (1)	2B medio (4), 7B basico (4), 4B medio (4), 4B basico (3)	14	15	46%
Matemáticas			numero (3), matematico (3), aplicar (2)		8		
Total	34	680			249	117	54%

FIGURA 3.11. Resumen de los términos asociados a asignatura y nivel por asignatura.

4. PROPUESTA DE METADATA PARA EDUCACIÓN

En la sección 2.11 se presentaron las definiciones de *metadatos según* los estándares IEEE-LOM, y *Dublin Core*. Antes de seguir, debemos indicar que las iniciativas de definir metadatos para recursos digitales al estilo de IEEE-LOM, Dublin Core o SCORM (Committee, 2002; D. C. M. Initiative et al., 2000; A. D. L. Initiative, 2001), provienen de la industria y son masivamente usados. Sin embargo, son costosas pues requieren conocimiento experto para hacer las anotaciones; pesadas pues implican muchos *tags*; ambiguos y altamente complejos por ejemplo no es claro el valor que debe asociarse a tags tales como *complejidad semantica* (Millard et al., 2009), lo que eleva su costo (Motelet, 2007). Pero aún más importante, carecen de soporte para la representación de temas pedagógicos asociados a un objeto de aprendizaje tales como requisitos de aprendizaje y los objetivos de aprendizaje que el recurso cubre (Nitto, Mainetti, Monga, Sbattella, y Tedesco, 2006).

De los experimentos desarrollados y del análisis de resultados observamos que los medadatos relevantes para profesores están relacionados principalmente con los medatados curriculares, particularmente *asignatura* y *nivel*, tales como *Matemáticas, Lenguaje y Comunicación, Ciencias,* 1^{ero} *medio,* 8^{vo} *básico*, por ejemplo. Si analizamos los metadatos IEEE-LOM y Dublin Core (sección 2.11), vemos que los metadatos se reducen a los se muestran en la figura 4.1 que contiene categorías metadatos y ejemplos de los valores de tales metadatos de acuerdo a los 34 vocabularios identificados en la sección 3.4.

En resumen la tabla de la Figura 4.1 nos indica que metadatos importantes en recursos digitales están relacionados con:

1. *Descripción*: categoría presente en cualquiera de los estándares IEEE-LOM o *Dublin Core*. Los términos explícitos que aparecen en estas categorías suelen ser muy diversos, pero en general hacen referencias a tópicos de enseñanza asociados a las asignatura en los distintos niveles del curriculum nacional.

Estándar	Categoría	Sub categoría y metadatos		
Dublin Core	**CD. Lenguaje**	inglés, castellano		
	DC. Descripción	acercamiento, crear, educación, moral, independencia, social		
	DC. Materia	artística, biología, naturales, cinética, química, geografía, sociales, castellano, lenguaje, matemática, historia, inglés		
	DC. Cobertura	1ero básico, 2do básico, 2do medio, 3ero básico, 3ero medio, 4to básico, 4to medio, 5to básico, 6to básico, 7mo básico, 8vo básico		
	DC. Format	sonido, video, cine, texto		
IEEE LOM	**Categoría technical (*técnica*).**	**4.1 Formato** sonido, video, cine, texto		
	5. Categoría educational (*educativa*).	**5.1 Tipo de Interactividad:** crear, argumentación, comprensión, definir, escriturar, lectura, oral, aplicar	**5.6 Contexto (Nivel)** 1ero básico, 2do básico, 2do medio, 3ero básico, 3ero medio, 4to básico, 4to medio, 5to básico, 6to básico, 7mo básico, 8vo básico	
	Categoría clasificación (classification)	**9.1 Identificador (Tema):** artística biología naturales cinética química geografía sociales castellano lenguaje matemática historia inglés	**9.4 Palabras Claves:** auditivo, sonido, sonoro, artista, aviso, cine, cultural, enlace, escénico, espacio, publicitario, acercamiento, calor, diversidad, entorno, estructura, físico, fuerza, materia, modelo, movimiento, natural, organismo, planeta, ser, vida, visión, vivo, ambiente, chile, contemporáneo, época, independencia, mundo, panorámico, región, regional, social, tierra, viaje, comunicar, discurso, lectura, lengua, literario, obra, análisis, función, geometría, numero, transformación, valor, alternativo, variedad	**9.5 Descripción:** acercamiento, crear, educación, moral, independencia, social

FIGURA 4.1. Esquema metadatos Dublin Core e IEEE-LOM que caracterizan a los términos asociados a las comunidades de interés de profesores.

2. *Cobertura* o *contexto* según estándar *Dublin Core* o IEEE-LOM: se reduce a metadatos de nivel educacional, en la curriculum nacional, para los 8 niveles de educación básica y los 4 niveles de educación media. Estos metadatos son estáticos mientras no cambie el curriculum nacional.

3. *Formato* o *Categoría Técnica* también presentes en los estándares IEEE-LOM y *Dublin Core*: en este caso, y con el advenimiento de nuevas tecnologías es una categoría de metadatos que toma relevancia por la mayor demanda de multimedia, y de formatos por parte de profesores y alumnos.
4. *Materia* o *Identificador* según el estándar: se reduce principalmente los nombres de las asignaturas presentes en los distintos niveles del curriculum nacional, y son estáticos mientras no existan cambios en el curriculum.

De seguro hay más categorías de metadatos presentes en los vocabularios, sin embargo del análisis de los vocabularios las categorías presentes en la tabla de la figura 4.1 son los más relevantes. Términos explícitos en estas categorías encontradas son más bien estáticos, como *cobertura*/*contexto* o *materia*/*tema*. Otros valores de categorías de metadatos tienen un carácter más dinámico, como *descripción*, pues los términos de los vocabularios estarán relacionados con los recursos asociados a temas específicos de los cursos que pueden variar por cambios de programas o por temas distintos a cubrir por época del año escolar.

5. CONCLUSIONES

En este trabajo hemos analizado el impacto de la metadata en la búsqueda y descubrimiento de recursos de aprendizaje para profesores de enseñanza básica y media en Chile. Para ello se ha utilizado un dataset, derivado de un repositorio de material educativo ampliamente utilizado en Chile, con datos de 5 años. Nuestro data set confirma la tendencia observada en data set educativos similares, es decir, se caracterizó una alta escasez de evaluaciones. El dataset contenía tanto recursos de aprendizaje anotados con metadata, como información de los profesores respecto de las búsquedas que han realizado (visitas a recursos Web de aprendizaje) en el portal Educar Chile. El análisis de estas visitas utilizando técnicas de recomendación y agrupación (clustering) ha permitido identificar comunidades de interés así como la metadata que caracteriza a los recursos de aprendizaje de su interés. También llevamos a cabo un experimento con profesores en etapa de formación, con el fin de obtener obtener perfiles de profesores, para validar la eficacia de la metadata identificada en una tarea exigente de recomendación como es el problema del *cold start* o arranque en frío.

5.1. Preguntas de investigación e Hipótesis

En esta sección presentamos las principales conclusiones de este trabajo con respecto a preguntas de investigación e hipótesis planteadas en la tesis.

Con respecto a las preguntas (R1) ¿Qué información deben considerar los sistemas de recomendación en el sistema de educación de media y básica del sistema escolar Chileno? y (R3) ¿Cuál es el impacto de considerar el contexto de la búsqueda y la metadata de los recursos digitales en la calidad de la recomendación?, se concluye lo siguiente:

Cómo se explicó antes, en esta tesis se identificaron las comunidades de interés común de profesores utilizando una técnica de agrupación jerárquica, depués se definieron vectores con los términos usados por los recursos visitados por cada grupo diferenciándolos por categoría (curricular, socio-pedagógica, contenido libre). También se utilizó la metadata diferenciada por categorías para determinar la similaridad de un nuevo item con los items

del grupo al que pertenece el profesor enfrentando así el problema de arranque en frío. Los resultados presentados en la Tabla 3.13 demuestran que al utilizar la *metadata curricular* como elemento para determinar tanto la pertenencia a la comunidad de profesores como la similaridad a los recursos visitados, la predicción de la preferencia por nuevos recursos de parte de nuevos profesores es significativamente mejor para umbrales de similaridad del 50 %.

Con respecto a la pregunta (R3) ¿Qué características del contexto del profesor son relevantes para la búsqueda de recursos educativos?, se concluye lo siguiente:

El análisis de los términos que describen los recursos (y por lo tanto el interés) visitados por las comunidades de profesores identificadas en la Fase 1, sub sección 3.1.5 se presenta en la Fase 4, sub-sección 3.4. En la tabla 3.9 se puede apreciar un fragmento de los 35 vocabularios identificados, las 20 palabras más utilizadas están referidas a la *asignatura* y *nivel*, ambos corresponden a la metadata curricular.

La respuesta a las preguntas de investigación R1, R2 y R3 permiten concluir que la Hipótesis 1 de esta tesis es correcta, efectivamente *los metadatos curriculares (sector/asignatura, nivel) influencian positivamente el descubrimiento de recursos de aprendizaje por parte de profesores*.

Con respecto a la pregunta (R4) ¿Cómo se debe representar esta información?, se concluye lo siguiente:

El análisis de los metadatos relevantes de la Fase 4, en contraste con los estándares para anotar metadata en educación presentado en el capítulo 4, nos permiten concluir que las categorías que pueden utilizarse para representar esta metadata son genéricas (ej. contexto, palabra clave). Allí se identifican los metadatos que podrían utilizarse para contener la metadata relevante encontrada en este trabajo. Se nota, particularmente en el caso de IEEE-LOM, que son metadatos genéricos que pueden contener diferente tipo de información por lo que su valor de discriminación es muy subjetivo (dependería de los valores de los metadatos). En el caso de Dublin Core existe un metadato que puede usarse directamente

para representar la asignatura (subject), pero en el estándar (que es más genérico que IEEE-LOM) también se plantea como un metadato general (i.e. se describe el "asunto" que cubre el recurso pero no necesariamente, la asignatura). Esto nos permite concluir que la hipótesis dos es correcta, es decir, *existen metadatos que son importantes para los profesores de educación de media y básica del sistema escolar Chileno, que no están contemplados como metadata estándar*.

Con respecto a la pregunta (R5) ¿Cómo se puede obtener información (para la búsqueda) sin imponer sobrecarga de trabajo al profesor?, se concluye lo siguiente:

En la subsección 3.2 se presentó un cuestionario para determinar el perfil del profesor, este cuestionario se diseñó y validó en conjunto con expertos de la Facultad de Educación. Los datos del cuestionario fueron utilizados en la subsección 3.3.1. Como se puede apreciar en el cuestionario presentado en las Figuras 3.2 y 3.3, la información necesaria para que el método propuesto en esta tesis para lidiar con el arranque en frío funcione es bastante estable, los profesores deberían caracterizar su perfil una vez, al entrar al sitio Web que contiene los recursos.

Finalmente, la estrategia propuesta para lidiar con el problema de arranque en frío nos permite concluir que la Hipótesis 3 es correcta, es decir, *si es posible prescindir de la historia de consumo de un usuario, para recomendar con alta precisión, recursos novedosos (no antes vistos) a un profesor*.

5.2. Contribuciones

Los sistemas de recomendación pueden ser utilizados como un medio para hacer frente a la gran cantidad de recursos disponibles en la Web, mediante el reconocimiento de los objetivos de las comunidades específicas de la práctica, como los profesores, si los algoritmos de recomendación consideran las propiedades y valores de la comunidad. Los metadatos asociados a los ítems se convierten en una pieza fundamental para hacer frente a la cantidad y diversidad de material. Nadolski (Nadolski et al., 2009) evalúa los efectos del uso de una ontología con un modelo de preferencias de usuario y descubre que una estrategia

basada en esta ontología es más precisa, aunque es costosa y compleja. Otros investigadores se centran en un enfoque más liviano ligero cuando se enfrentan a la complejidad de metadatos IEEE-LOM (Tiropanis et al., 2009).

Una contribución de esta tesis es la identificación de la influencia que los metadatos *curriculares* tienen para mejorar la calidad de recomendación y potencialmente las búsquedas de recursos por parte de profesores, particularmente asignatura y nivel. Sin embargo, en esta tesis se encontró que estos metadatos no deben considerarse como categorías exclusivas en la discriminación de datos. Por ejemplo, si vemos lo que sucede con el *nivel*; para el cluster $\mathscr{C}_1$ en la tabla 3.7, el metadato más importante de nivel es 5^{to} *básico*, mientras que para los cursos superiores (6^{to}, 8^{vo} básico, y 1^{ero} medio) la relevancia disminuye fuertemente pero no es inexistente. Esto puede sugerir que un solo recurso puede ser reutilizado en varios niveles con propósitos diferentes; un recurso podría ser fundamental en algunos niveles, pero introductorio en otros. Lo mismo sucede en la tabla 3.9, podemos notar que se forman vocabularios donde las comunidades se refieren a dos asignaturas simultáneamente aunque una tiene mayor importancia que la otra. Es importante considerar la combinación adecuada de la relevancia de un recurso educativo según la materia y el nivel, en lugar de considerar a éstos metadatos como discriminadores absolutos (por ejemplo, considerar que un profesor puede estar interesado en un sólo nivel o una sola materia).

Otra contribución de esta tesis constituye la estrategia para enfrentar el problema de arranque en frío (*cold start problem*) en el contexto de educación, caracterizado por una alta dispersión en evaluaciones o visitas de items, mediante la explotación de un perfil de profesor y de la identificación de comunidades de interés común. Los metadatos y nuestro enfoque nos permiten estimar evaluaciones para nuevos usuarios y nuevos elementos. Una vez más, los metadatos curriculares se convierten en una pieza clave para minimizar el error de predicción de estas evaluaciones.

Finalmente, en esta tesis se identifica una falencia de los estándares utilizados popularmente para anotar recursos de aprendizaje en relación a su capacidad de representar la metadata que en esta invetigación se ha encontrado como relevante para la recomendación

de recursos. Si bien se pueden utilizar algunos tags, estos son demasiado generales como para servir de elementos de clasificación útiles para contenido relacionado con educación básica y media en Chile. Dada la naturaleza de las prácticas docentes en latinoamérica para este nivel escolar, se cree que este trabajo podría ser extendido para otros países de Latinoamérica. Dada la inversión que varios gobiernos hacen en proveer recursos de aprendizaje electrónicos, este hallazgo cobra mayor relevancia.

REFERENCIAS

Adomavicius, G., y Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, *17*(6), 734-749.

Berners-Lee, T. (1997). *Metadata architecture.*

Burke, R. (2007). Hybrid Web Recommender Systems. *The Adaptive Web*, *4321*, 377–408.

Canini, L., Benini, S., y Leonardi, R. (2013). Affective recommendation of movies based on selected connotative features. *IEEE Trans. Circuits Syst. Video Techn.*, *23*(4), 636–647.

Carrer-Neto, W., Hernández-Alcaraz, M. L., Valencia-García, R., y Sánchez, F. G. (2012). Social knowledge-based recommender system. application to the movies domain. *Expert Syst. Appl.*, *39*(12), 10990–11000.

Chatti, M., Dakova, S., Thus, H., y Schroeder, U. (2013). Tag-based collaborative filtering recommendation in personal learning environments. *IEEE Transactions on Learning Technologies*.

Colombo-Mendoza, L. O., Valencia-García, R., González, A. R., Alor-Hernández, G., y Zapater, J. J. S. (2015). Recommetz: A context-aware knowledge-based mobile recommender system for movie showtimes. *Expert Syst. Appl.*, *42*(3), 1202–1222.

Committee, L. T. S. (2002). *1484.12.1tm ieee standard for learning object metadata.*

Dascalu, M.-I., Bodea, C.-N., Lytras, M., Pablos, P. O. de, y Burlacu, A. (2014). Improving e-learning communities through optimal composition of multidisciplinary learning groups. *Computers in Human Behavior*, *30*, 362 - 371.

Devyver, P. A., y Kittler, J. (1982). *Pattern recognition: A statistical approach*. London: Prentice-Hall.

Ding, L., Finin, T. W., Joshi, A., Pan, R., Cost, R. S., Peng, Y., y cols. (2004). Swoogle: a search and metadata engine for the semantic web. En D. A. Grossman, L. Gravano, C. Zhai, O. Herzog, y D. A. Evans (Eds.), *Proceedings of the 2004 ACM CIKM international conference on information and knowledge management, washington, dc, usa, november 8-13, 2004* (pp. 652–659). ACM.

Drachsler, H., Bogers, T., Vuorikari, R., Verbert, K., Duval, E., Manouselis, N., y cols. (2010). Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia CS*, *1*(2), 2849-2858.

Duval, E., Vuorikari, R., y Manouselis, N. (2009). Special issue on social information retrieval for technology enhanced learning. *J. Digit. Inf.*, *10*(2).

Felfernig, A., y Burke, R. D. (2008). Constraint-based recommender systems: technologies and research issues. En D. Fensel y H. Werthner (Eds.), *Icec* (Vol. 342, p. 3). ACM.

Gayo, J. E. L., Pablos, P. O. de, y Lovelle, J. M. C. (2010). Wesonet: Applying semantic web technologies and collaborative tagging to multimedia web information systems. *Computers in Human Behavior*, *26*(2), 205–209.

Ghazarian, S., Shabib, N., y Nematbakhsh, M. A. (2014). Improving sparsity problem in group recommendation. En F. Cena, A. S. da Silva, y C. Trattner (Eds.), *Hypertext 2014 extended proceedings: Late-breaking results, doctoral consortium and workshop proceedings of the 25th ACM hypertext and social media conference (hypertext 2014), santiago, chile, september 1-4, 2014.* (Vol. 1210). CEUR-WS.org.

Goldberg, K., Roeder, T., Gupta, D., y Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retr.*, *4*(2), 133-151.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., y Tibshirani, R. (2009). *The elements of statistical learning* (Vol. 2) (n.º 1). Springer.

Herlocker, J., Konstan, J. A., y Riedl, J. (2002, October). An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Inf. Retr.*, *5*(4), 287–310.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., y Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*, 5-53.

Ims global learning consortium. (1995).

Initiative, A. D. L. (2001). *Sharable content object reference model - scorm.*

Initiative, D. C. M., y cols. (2000). *Dcmi specifications.*

Jannach, D., Zanker, M., Felfernig, A., y Friedrich, G. (2003). *Recommender systems, an introduction*. Cambridge Press.

Jurafsky, D., y Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Prentice Hall.

Koper, R. (2003). Combining reusable learning resources and services to pedagogical purposeful units of learning. *Reusing online resources: A sustainable approach to eLearning*, 46–59.

Kurilovas, E., Serikoviene, S., y Vuorikari, R. (2014). Expert centred vs learner centred approach for evaluating quality and reusability of learning objects. *Computers in Human Behavior*, *30*(0), 526 - 534.

Learning Technology Standards Committee of the IEEE. (2002, July 15). *Draft standard for learning technology - learning object metadata* (Inf. Téc.). New York: IEEE Standards Department.

Lehmann, L., Hildebrandt, T., Rensing, C., y Steinmetz, R. (2008). Capture, management, and utilization of lifecycle information for learning resources. *Learning Technologies, IEEE Transactions on*, *1*(1), 75–87.

Lika, B., Kolomvatsos, K., y Hadjiefthymiades, S. (2014a). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065–2073.

Lika, B., Kolomvatsos, K., y Hadjiefthymiades, S. (2014b). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065–2073.

Mandl, M., Felfernig, A., Teppan, E., y Schubert, M. (2011). Consumer decision making in knowledge-based recommendation. *J. Intell. Inf. Syst.*, *37*(1), 1–22.

Manouselis, N., y Costopoulou, C. (2007). Experimental analysis of design choices in multiattribute utility collaborative filtering. *IJPRAI*, *21*(2), 311-331.

Manouselis, N., y Costopoulou, C. (2008). Experimental analysis of multiattribute utility collaborative filtering on a synthetic data set. *Personalization Techniques and Recommender Systems, Series in Machine Perception and Artificial Intelligence*, *70*, 111–134.

Manouselis, N., Drachsler, H., Verbert, K., y Duval, E. (2013). *Recommender systems for learning*. Springer.

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., y Koper, R. (2011). Recommender systems in technology enhanced learning. En *Recommender systems handbook* (pp. 387–415). Springer.

Manouselis, N., Vuorikari, R., y Van Assche, F. (2010). Collaborative recommendation of e-learning resources: an experimental investigation. *Journal of Computer Assisted Learning*, *26*(4), 227–242.

Millard, D. E., Howard, Y. M., McSweeney, P., Arrebola, M., Borthwick, K., y Varella, S. (2009). Phantom tasks and invisible rubric: The challenges of remixing learning objects in the wild. En U. Cress, V. Dimitrova, y M. Specht (Eds.), *Learning in the synergy of multiple disciplines, 4th european conference on technology enhanced learning, EC-TEL 2009, nice, france, september 29 - october 2, 2009, proceedings* (Vol. 5794, pp. 127–139). Springer.

Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., y Riedl, J. (2003). Movielens unplugged: experiences with an occasionally connected recommender system. En *Iui* (p. 263-266). ACM.

Motelet, O. (2007). *Improving learning-object metadata usage during lesson authoring*. Tesis Doctoral no publicada, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile.

Nadolski, R., Berg, B. van den, Berlanga, A. J., Drachsler, H., Hummel, H. G. K., Koper, R., y cols. (2009). Simulating light-weight personalised recommender systems in learning networks: A case for pedagogy-oriented and rating-based hybrid recommendation strategies. *J. Artificial Societies and Social Simulation*, *12*(1).

Nitto, E. D., Mainetti, L., Monga, M., Sbattella, L., y Tedesco, R. (2006). Supporting interoperability and reusability of learning objects: The virtual campus approach. *Educational Technology & Society*, *9*(2), 33-50.

Ochoa, X., y Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *Int. J. on Digital Libraries*, *10*(2-3), 67-91.

Ochoa, X., Klerkx, J., Vandeputte, B., y Duval, E. (2011). On the use of learning object metadata: The globe experience. En *Towards ubiquitous learning* (pp. 271–284). Springer.

Ortega, F., Bobadilla, J., Hernando, A., y Rodríguez, F. (2014). Using hierarchical graph maps to explain collaborative filtering recommendations. *Int. J. Intell. Syst.*, *29*(5), 462–477.

Popescul, A., Ungar, L. H., Pennock, D. M., y Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. En J. S. Breese y D. Koller (Eds.), *UAI '01: Proceedings of the 17th conference in uncertainty in artificial intelligence, university of washington, seattle, washington, usa, august 2-5, 2001* (pp. 437–444). Morgan Kaufmann.

Rodrıguez, A. C., Gago, J. M. S., Rifón, L. E. A., y Rodrıguez, R. P. (2015). A recommender system for non-traditional educational resources: A semantic approach. *Journal of Universal Computer Science*, *21*(2), 306–325.

Santos, O. C., y Boticario, J. G. (2010). Modeling recommendations for the educational domain. *Procedia Computer Science*, *1*(2), 2793 - 2800.

Segaran, T. (2007). *Programming collective intelligence: Building smart web 2.0 applications*. O'Reilly.

Shepitsen, A., Gemmell, J., Mobasher, B., y Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. En *Proceedings of the 2008 acm conference on recommender systems* (pp. 259–266). ACM.

Sieg, A., Mobasher, B., y Burke, R. (2010). Improving the effectiveness of collaborative recommendation with ontology-based user profiles. En *proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems* (pp. 39–46). ACM.

Song, B., y Gao, J. (2014). The enhancement and application of collaborative filtering in e-learning system. En *Advances in swarm intelligence* (pp. 188–195). Springer.

Steinacker, A., Ghavam, A., y Steinmetz, R. (2001). Metadata standards for web-based resources. *IEEE MultiMedia*, *8*(1), 70–76.

Tang, T. Y., Winoto, P., y McCalla, G. (2014). Further thoughts on context-aware paper recommendations for education. En *Recommender systems for technology enhanced learning* (pp. 159–173). Springer.

Taraghi, B., Saranti, A., Ebner, M., Müller, V., y Großmann, A. (2015). Towards a learning-aware application guided by hierarchical classification of learner profiles. *J. UCS*, *21*(1), 93–109.

Tiropanis, T., Davis, H. C., Millard, D. E., y Weal, M. J. (2009). Semantic technologies for learning and teaching in the web 2.0 era. *IEEE Intelligent Systems*, *24*(6), 49-53.

Valdebenito, D., y Cruzat, C. (2012). *Enlaces, innovación y calidad en la era digital 20 años impulsando el uso de las tic en la educación.*

Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., y Duval, E. (2011). Dataset-driven research for improving recommender systems for learning. En P. Long, G. Siemens, G. Conole, y D. Gasevic (Eds.), *Lak* (p. 44-53). ACM.

Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., y cols. (2012). Context-aware recommender systems for learning: a survey and future challenges. *Learning Technologies, IEEE Transactions on*, *5*(4), 318–335.

Zhang, X., Pablos, P. Ordonez de, y Zhang, Y. (2012). The relationship between incentives, explicit and tacit knowledge contribution in online engineering education project. *The Knowledge Society Selected Papers on Engineering Education from the 4th World Summit on the Knowledge Society*.

Zhao, J.-c., Liu, S.-h., y Zhang, J. (2015). Research on personalized recommendation system on item-based collaborative filtering algorithm. En *Proceedings of the international conference on advances in mechanical engineering and industrial informatics (ameii 2015), zhengzhou, china, apr 11-12, 2015.*