UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 1

DEPARTAMENTO DE TÉCNICAS DE PROGRAMACIÓN

PROCEDIMIENTO PARA LA RECOMENDACIÓN DE DOCUMENTOS DE ARCHIVO

Tesis presentada en opción al título de Máster en Informática Aplicada

Autor: Ing. Marlon Jorge Remedios González

Tutor: Dr.C. Yusnier Valle Martínez

La Habana, Junio 2014

A mi madre Ana y resto de mi numerosa familia.

A mi esposa, Maria T.

DECLARACIÓN DE AUTORÍA Y AGRADECIMIENTOS

Yo Marlon Jorge Remedios González, con carné de identidad 85012521223, declaro que soy el autor principal del resultado que expongo en el presente trabajo titulado “Procedimiento para la recomendación de documentos de archivo”, para optar por el título de Máster en Informática Aplicada.

El trabajo fue desarrollado individualmente durante el período comprendido entre el 2012-2014.

Deseo agradecer al Dr.C. Yusnier Valle Martínez, por ayudarme en la formación como ingeniero y máster. Además deseo agradecer a los doctores Luis Martínez y Jorge Gallardo de la Universidad de Jaen, quienes han contribuido con mi superación profesional. A todos ellos y a los que no menciono, que han estado presentes y seguirán ahí para cuando lo necesite, mis más sinceros agradecimientos.

Finalmente declaro que todo lo anteriormente expuesto se ajusta a la verdad, y asumo la responsabilidad moral y jurídica que se derive de este juramento profesional.

Y para que así conste, firmo la presente declaración jurada de autoría en La Habana a los ___ días del mes de _____ del año _____.

_____

Firma del Maestrante

RESUMEN

El presente trabajo hace un análisis de los principales métodos de recomendación de elementos y la explotación de las características de la norma que rige la descripción archivística. Se exponen las principales técnicas de recomendación en enfoques basados en el contenido y colaborativos. A partir del estudio realizado, se presenta un procedimiento que establece cómo debe diseñarse un Sistema de Recomendación en entornos que pretenden difundir acervos históricos. El procedimiento presentado cuenta con cinco componentes (Fuente, Representación, Analizador, Motor de Recomendación y Selección) para cada uno de ellos se describe sus responsabilidades, mostrando cómo se relacionan entre sí y explotan características del entorno para mitigar el problema del arranque en frío. Para validar la propuesta se implementó una herramienta según lo establecido en el procedimiento y los resultados de los experimentos son expuestos y discutidos en el último apartado de la investigación, demostrando que las características de los entornos regidos por documentos de archivos que se insertan en el procedimiento, mejoran la cobertura del sistema y la precisión en la recomendación.

Tabla de contenido

Índice de Figuras

Índice de Tablas

INTRODUCCIÓN

Con el crecimiento de internet y el gran cúmulo de información, se hace difícil y poco preciso para los usuarios acceder rápidamente a la información que necesitan. Muchas son las técnicas de recuperación de información que se utilizan para facilitar la navegabilidad de los usuarios en los diferentes entornos por los que transitan, con el fin de acceder a lo que realmente les interesa.

Los usuarios necesitan un soporte personalizado para atravesar la enorme cantidad de información disponible, de acuerdo a sus intereses y gustos (Burke 2002; Ricci, Rokach et al. 2011). Varias fuentes de información incluyen Sistemas de Recomendación (SR) como una forma de personalización del contenido de los usuarios (Resnick and Varian 1997). El problema de recomendación de elementos[1] ha sido estudiado extensivamente y dos paradigmas han surgido: Sistemas de Recomendación Basados en el Contenido (SRBC) y Sistemas de Recomendación Colaborativos (SRC) (Balabanović and Shoham 1997).

De diferentes formas se han utilizado los acercamientos mencionados anteriormente (SRBC y SRC). En algunos entornos se han combinado las características de cada una de estas técnicas. Estos enfoques híbridos tienen la finalidad de obtener sistemas más precisos o eliminar problemas específicos de un sistema en concreto (Burke 2002).

ANTECEDENTES Y ESTADO ACTUAL DEL TEMA

En la actualidad la personalización es un mecanismo muy utilizado en diversas áreas. El objetivo de la personalización está centrado en adaptar un servicio a los gustos y necesidades particulares de un usuario (Howe A 2008). Una de las herramientas más utilizadas y que mejores resultados ha proporcionado, con el fin de obtener personalizaciones de elementos según el escenario donde se produzcan, han sido los Sistemas de Recomendación (Burke 2000; Burke 2002; Niu, Yan et al. 2002; Mak, Koprinska et al. 2003).

En la década del 90 uno de los primeros acercamientos a los SR fueron los servicios de noticia que brindaba “newsgroups” a sus usuarios, permitiéndoles acceder solamente a las noticias que considerasen que eran de interés para el usuario (Sahlgren 2006). El primer SR que surgió fue el llamado “Tapestry” (Bonhard, Harries et al. 2006), el cual usaba retroalimentación de las noticias y artículos leídos por algunos usuarios para luego decidir cuál de estos elementos sería relevante para otros usuarios que no habían accedido aún a la información.

Otra de las formas de filtrado de información electrónica apareció con la propuesta por Allen (Cooley, Mobasher et al. 1997; Carlson 2003) para la creación de modelos de usuarios o la aproximación de “The information Lens system” (Cooley, Mobasher et al. 1997) donde, a partir del contexto, los usuarios podían definir reglas para el filtrado del correo electrónico.

[1] En el resto del documento se hará referencia al término elemento como instancia del objeto que se desee tratar. Ejemplos: documentos, mapas, informes, fotos, películas, entre otros.

Actualmente el significado del término “sistema de recomendación” hace referencia a cualquier sistema que produce recomendaciones individuales como salida o que tiene el efecto de guiar al usuario de un modo personalizado a objetos útiles o interesantes dentro de un gran espacio de posibles opciones. Estos sistemas son muy atractivos en contextos donde la cantidad de información que se ofrece al usuario supera ampliamente cualquier capacidad de exploración (Howe A 2008).

Como se mencionó anteriormente, dos de los principales enfoques más utilizados son los Sistemas de Recomendación basados en el Contenido y los Sistemas de Recomendación Colaborativos. Los primeros tratan de recomendar elementos similares a aquellos que los usuarios han valorado en el pasado (Lops, de Gemmis et al. 2011). Por otra parte, el acercamiento de filtrado colaborativo (o social) (Linden, Smith et al. 2003; Deshpande and Karypis 2004), se basa en las valoraciones de un usuario, así como en la de otros usuarios en el sistema. La idea principal es que la valoración de un usuario sobre un elemento va a ser similar a la valoración de otro usuario si ambos han valorado elementos similares (Desrosiers and Karypis 2011).

Los sistemas de recomendación han sido usados en una gran variedad de aplicaciones, tales como la recomendación de libros y CD (del inglés Compact Disk) (Mooney and Roy 2000; Linden, Smith et al. 2003), música, películas (Miller, Albert et al. 2003; Bell and Koren 2007), noticias (Billsus, Brunk et al. 2002) y páginas web (Mobasher, Dai et al. 2002). En cada una de estas se realiza una caracterización del entorno y análisis de las propiedades de los elementos a recomendar para obtener recomendaciones más precisas, a partir de la selección de enfoques o hibridación de estos.

FORMULACIÓN DEL PROBLEMA

El diseño de un Sistema de recomendación requiere, más que la elección de un acercamiento (SRBC, SRC) o incluso la hibridación de estos, de técnicas que permitan la caracterización del entorno. Las aplicaciones que brindan un número finito de servicios, regidos por un gran cúmulo de información, que pretenden adjuntar un Sistema de Recomendación, van a depender de la integración de un conjunto de técnicas que permitan proponer con mayor exactitud y precisión elementos que puedan resultar de interés a los usuarios que hagan uso de la misma.

Actualmente la Archivística -según la Real Academia Española (RAE), es el estudio teórico y práctico de los principios, procedimientos y problemas concernientes al almacenamiento de documentos, buscando que dicha documentación se mantenga en el tiempo, pueda ser consultada y clasificada-, se ha convertido en uno de los objetivos de la era digital. Esto trajo consigo que la archivística tuviese que replantearse la forma de almacenar y gestionar los documentos archivísticos, que según (Ruiz 2001), es toda expresión testimonial en cualquier lenguaje, forma o soporte (forma oral o escrita, textual o gráfica, manuscrita o impresa, en lenguaje natural o codificado, en cualquier soporte documental así como en cualquier otra expresión gráfica, sonora, en imagen o electrónica), generalmente en ejemplar único, (aunque puede ser multicopiado o difundido).

Existen varios países e instituciones que han presentado su acervo histórico en diferentes aplicaciones en la gran red de redes. Cuba[2] por ejemplo cuenta con sitios web para su Archivo Histórico Nacional (AHN) o para archivos en algunas de sus provincias como Villa Clara[3]. Por otra parte, España[4] es otro de los países que cuenta con una aplicación alojada en Internet (Portal de Archivos Españoles) con el objetivo de presentar y difundir su acervo histórico. Estados Unidos[5] y Venezuela[6] también cuentan con un portal de documentos históricos. Algunos otros sitios dedicados a representar los Archivos Históricos se pueden consultar en (Semeraro 2009). Cada una de estas aplicaciones pone de cara al público los documentos archivísticos en diferentes formas y cuentan, naturalmente, con servicios diferentes aunque muchos de ellos converjan hacia un mismo fin.

Cada una de estas aplicaciones permite la recuperación de información en diferentes formas, pero debido al gran número de elementos presentes se torna difícil presentar lo que realmente es de interés para el usuario. En este sentido cada uno de estos portales incorporan elementos que facilitan, hasta cierto punto, la navegabilidad del usuario, como por ejemplo los últimos documentos añadidos, las últimas descripciones consultadas, refinar los criterios de búsqueda de diferentes formas predefinidas por el sistema, entre muchas otras estrategias. Aunque son varias las facilidades que proveen los elementos anteriormente mencionados, ninguno de estos sitios cuentan con un SR capaz de identificar las necesidades de los usuarios a partir de la organización de la información y las diferentes formas en las que el usuario transita por el gran cúmulo de descripciones archivísticas que ellos presentan.

Las aplicaciones con el fin de gestionar, presentar y difundir los acervos históricos cuentan con una gran variedad de elementos. En cada una de sus colecciones pueden aparecer elementos como cartas, documentos, informes, mapas, registros de audio, entre otros tipos de documentos de archivo. Muchas de estas aplicaciones en dependencia del objetivo (político, social o cultural) presentan colecciones de imágenes solamente o colecciones de audio, mapas, cartas, entre otras. Esto trae consigo la necesidad de caracterizar cada uno de los tipos de elementos o el conjunto de ellos que pueden estar presentes en las colecciones de archivística con la finalidad de facilitar su presentación a los usuarios que necesitan su búsqueda o consulta.

Debido al impacto que tiene la difusión de la información en la sociedad actual, es de suma importancia la aplicación de un conjunto de técnicas, que rijan en ciertas formas, los aspectos fundamentales para el diseño de Sistemas de recomendación. Teniendo en cuenta el poco desarrollo de Sistemas de recomendación que exploten las características de Archivística y la importancia que tiene guiar a los usuarios en las aplicaciones caracterizadas por un ilimitado número de información, se plantea como **problema de la investigación** ¿Cómo diseñar un Sistema de recomendación en aplicaciones que pretenden difundir acervos históricos?

[2] AHN disponible en www.arnac.cu
[3] Archivo de Villa Clara disponible en www.archivohistorico.villaclara.cu
[4] PARES disponible en pares.mcu.es
[5] Archivo EU disponible en www.archives.gov/historical-docs
[6] AGN (Archivo General de la Nación) disponible en www.sahisweb.gob.ve

El **objeto** en el cual se enmarca el estudio en aras de dar solución al problema, tanto desde el punto de vista teórico como práctico, va a estar dado por el diseño de Sistemas de recomendación en aplicaciones que pretenden difundir acervos históricos.

Con el propósito de brindar una solución efectiva al problema, se plantea como **objetivo general** definir un procedimiento que permita diseñar un Sistema de recomendación para aplicaciones que pretenden difundir acervos históricos.

Para orientar la labor investigativa, la pregunta general formulada con anterioridad se desglosa en las siguientes interrogantes científicas:

- ¿Cómo representar de manera uniforme los documentos de archivo, a partir de las características que rigen su descripción, para garantizar el desarrollo de los componentes fundamentales de un Sistema de recomendación?
- ¿Cómo aprovechar las características de los entornos regidos por documentos de archivo en la aplicación de un Sistema de recomendación?
- ¿Cómo integrar los principios que rigen la descripción de documentos de archivo en el diseño de un Sistema de recomendación, contribuyendo a las propiedades que evalúan su desempeño?

Como **tareas de investigación** se proponen las siguientes:

- Identificación de las variantes de solución existentes y tendencias actuales para el diseño de un Sistema de Recomendación, a partir del estudio profundo de los referentes teóricos que preceden el presente trabajo.
- Caracterización de los diferentes acercamientos, algoritmos, hibridaciones y procedimientos utilizados en el diseño de Sistemas de Recomendación recogidos en la bibliografía actual, que puedan servir de precedente para la investigación.
- Análisis de cada una de las técnicas y normas que rigen la salvaguarda y difusión del acervo histórico a partir de los estándares para la descripción, intercambio y representación de los documentos de archivo.
- Desarrollo de un procedimiento que permita obtener un Sistema de Recomendación para insertar dentro de aplicaciones con fines de difusión de los acervos históricos.
- Evaluación de los resultados obtenidos haciendo uso de métricas de evaluación para las propiedades de los Sistemas de recomendación, a través de la implementación de una herramienta de software que permita realizar las pruebas al procedimiento desarrollado.

La **contribución** fundamental de este trabajo es el desarrollo de un procedimiento para el diseño de SR según la caracterización y explotación de los entornos archivísticos que pretenden difundir sus acervos históricos.

Como **resultados teóricos** se espera:

- Una visión ampliada en el desarrollo de Sistemas de recomendación a partir del uso de los principales acercamientos clásicos (SRBC, SRC), así como la hibridación de estos.

- Un procedimiento que facilite el diseño de un Sistema de recomendación en entornos que pretenden la difusión de acervos históricos.
- Documentación clara y precisa que tipifica las diferentes métricas de evaluación de los Sistemas de Recomendación.

ESTRUCTURA DEL DOCUMENTO

El presente documento se encuentra divido en tres capítulos:

Capítulo 1: Sistemas de Recomendación y Archivística: Un acercamiento a la actualidad

En este capítulo se realiza una caracterización de los principales conceptos asociados al dominio del problema. Se hace una revisión bibliográfica de los principales modelos, técnicas, algoritmos y acercamientos para el diseño de Sistemas de recomendación. Aparejado a esto se caracterizan además las diferentes formas, que aparecen en la bibliografía, en que se pueden hibridar los acercamientos en los Sistemas de recomendación para solapar los problemas que por separado acarrean. Aparece además la caracterización de las principales normas y principios que rigen la archivística.

Capítulo 2: Sistema de Recomendación en Entornos Archivísticos: Presentación de la solución propuesta

En este capítulo se describen los principales componentes del procedimiento propuesto. Inicialmente se propone el esquema general de funcionamiento del procedimiento con cada uno de los componentes que lo conforman. Se exponen y explican cada una de las responsabilidades asignadas, así como la explotación de las características de los entornos de archivos. Se deja claro además, cómo se relacionan con los restantes componentes de la solución propuesta.

Capítulo 3: Validación del procedimiento

En este capítulo se hace una caracterización de las métricas de evaluación de los Sistemas de recomendación. Se presentan además, los resultados de las pruebas realizadas a la aplicación resultante haciendo uso de la propuesta desarrollada en la presente investigación.

CAPÍTULO 1: SISTEMAS DE RECOMENDACIÓN Y ARCHIVÍSTICA: UN ACERCAMIENTO A LA ACTUALIDAD.

Con el objetivo de facilitar la comprensión del alcance de la investigación, en el presente capítulo se exponen conceptos fundamentales asociados al dominio del problema planteado. Se caracterizan los principales enfoques usados en la actualidad para el desarrollo de un Sistema de recomendación, así como diferentes formas de hibridarlos para mitigar los problemas que acarrean por separado. Se describe además un conjunto de normas y principios relevantes para la gestión archivística que facilitan la comprensión del entorno.

1.1 ARCHIVÍSTICA

La Archivística es una disciplina relativamente moderna y con el nombre de Archivología, nace en el siglo XIX como una técnica empírica para arreglo y conservación de los archivos. Su configuración como disciplina independiente y su consideración como ciencia auxiliar es bastante reciente (Herrera 1986).

Herrera (Herrera 1991) define la Archivística como: La ciencia que estudia la naturaleza de los archivos, los principios de su conservación y organización y los medios para la utilización. De igual manera Cruz Mundet en (Cruz Mundet 1994) refleja la definición dada por el Consejo Internacional de Archivo (C.I.A.[7]): Disciplina que trata de los aspectos teóricos y prácticos de los archivos y de su función.

Este amplio campo específico de la Archivística tiene que contar con la ayuda de otras ciencias que, como auxiliares, son indispensables para su completo desarrollo. Y también precisa para su desenvolvimiento de los conocimientos de otros profesionales relacionados con la informática (Herrera 1986). En este sentido (Cruz Mundet 1994) hace referencia a la aplicación de las nuevas tecnologías: Todavía no se ha llegado al extremo de fabricar un androide archivero que sustituya por completo el trabajo humano, sin embargo las nuevas tecnologías brindan la posibilidad de automatizar muchos procesos con indudables ventajas: la gestión administrativa, el almacenamiento y sustitución de soportes, la difusión, entre otras.

Un concepto muy relacionado en esta área es el de Archivo: los archivos fueron considerados desde la antigüedad como “lugar de preservación de los documentos bajo la jurisdicción de una autoridad pública”, el lugar que dotaba a los documentos de veracidad y les otorgaba la capacidad de servir como evidencia y memoria continua de las acciones (Ahn, Brusilovsky et al. 2007).

Son los archivos entendidos como conjunto de documentos portadores de información, los que centran prioritariamente la atención de esta disciplina, convirtiéndolos en su objeto. Es importante no perder de vista la triple dimensión del objeto de archivística:

Archivo – Documentos de archivo – Información

[7] C.I.A. Dictionary of Archival Terminology, Mürichen, New York, London, Paris, 1984.

La finalidad no es otra que el servicio de los archivos a la sociedad, materializado en el ofrecimiento de la información, ya sea a las instituciones productoras o a los ciudadanos, sean o no estudiosos (Herrera 1986).

1.1.1 LOS ARCHIVOS: CARACTERÍSTICAS Y FUNCIONES

Una de las dimensiones del objeto de la archivística como se menciona anteriormente es precisamente los archivos. La definición de archivos para muchas personas se puede describir por un conjunto de adjetivos: sótanos, suciedad, amontonamiento, desorden, oscuridad, serían los términos que fueran unidos a la idea de archivo. De igual manera (Cruz Mundet 1994) plasma otros conceptos errados cuando se habla de los archivos: depósitos de documentos arrumbados, desorganizados, cuando en puridad no se les debería dar esa consideración.

En (Cruz Mundet 1994) se realiza un análisis de la definición de archivos a partir de definiciones dadas por un conjunto de autores, alguna de ellas:

Jenkinson decía que los archivos son documentos acumulados en cualquier fecha por un proceso natural en el curso de la tramitación de los asuntos de cualquier tipo, público o privado y conservados después para su consulta, bajo la custodia de las personas responsables de los asuntos en cuestión o por sus sucesores (Jenkinson 1980).

El Diccionario de Terminología Archivística del C.I.A, lo define con tres acepciones: 1. Conjunto de documentos sean cuales sean su fecha, su forma y su soporte material, producidos o recibidos por toda persona física o moral y por todo servicio u organismo público o privado en el ejercicio de su actividad, son conservados por sus creadores o por sus sucesores para sus propias necesidades, ya trasmitidos a la institución de archivos competente en razón de su valor archivístico. 2. Institución responsable de la acogida, tratamiento, inventariado, conservación y servicio de los documentos. 3. Edificio o parte de edificio donde los documentos son conservados y servidos.

Para Herrera en (Herrera 1991), archivo es uno o más conjuntos de documentos, sea cual sea su fecha, su forma y soporte material, acumulados en un proceso natural por una persona o institución pública o privada en el transcurso de su gestión, conservados, respetando aquel orden, para servir como testimonio e información para la persona o institución que lo produce, para los ciudadanos o para servir de fuentes de historia.

El archivo se debe definir a partir de sus características, recogidas en cada una de las definiciones de los autores anteriormente mencionados y los introducidos en (Herrera 1986).

Cruz Mundet en (Cruz Mundet 1994), resume las cinco características principales a partir de las definiciones de los autores que presenta, como: ¿Qué compone el archivo?, ¿Quién crea, produce o genera un archivo?, ¿Cómo se forma un archivo? y por último, no es suficiente con que sean documentos producidos por cualquier entidad en el desarrollo de su actividad. Para que se pueda hablar de archivo, los documentos han de estar organizados y su información recuperable para su uso.

Las funciones de archivos son: reunir, conservar, ordenar, describir y utilizar los documentos. Estas funciones fueron resumidas en (Sheth and Maes 1993) a: recoger, conservar y servir los documentos. Las funciones de los archivos se encuentran tras los objetos mencionados de Archivística y se pueden resumir en los siguientes aspectos (Cruz Mundet 1994):

- Organización y puesta en servicio de la documentación administrativa, durante ese periodo de máxima utilidad para la gestión administrativa de las oficinas y para la toma de decisiones.
- Asegurar la transferencia periódica al archivo de los documentos que ya no son de uso corriente por parte de las oficinas.
- Aplicar los principios y técnicas modernos de valoración para, transcurrido un tiempo, seleccionar los documentos que por su valor van a ser conservados indefinidamente, y destruir el resto.
- Clasificar los fondos y mantener ordenada la documentación en sus distintas etapas, de acuerdo con los principios de la archivística.
- Describir la documentación para hacer fácilmente accesible la información, mediante los distintos instrumentos de descripción documental y valiéndose de las ventajas ofrecidas por las nuevas tecnologías.

1.2 DOCUMENTO ARCHIVÍSTICO

Diferentes autores reconocidos en el área de la archivística, presentados por Mayra Mena Múgica en su tesis doctoral (Múgica 2006), muestran varias definiciones para Documento Archivístico (DA), que en algunas bibliografías hacen referencia a este término, solo como documento. En (Múgica 2006) se realiza un estudio profundo del alcance del significado de DA y se converge hacia los conceptos que se muestran a continuación por ser representativos de otros dados por Cruz Mundet (1996):

Lo que define, en suma, la naturaleza del documento archivístico es justamente su funcionalidad como instrumento y como testimonio, prueba o evidencia de los actos o transacciones de la sociedad.

Otro acercamiento hacia la comprensión del significado de Documento de archivo es la que da CIA presentada en (Múgica 2006) que define a este como:

Información registrada (documento(s)) independientemente de su forma o medio, creada, recibida o mantenida por una entidad, institución, organización o individuo en el curso de sus obligaciones legales o en sus transacciones de negocios.

En este trabajo se presentan además algunas definiciones relevantes, un ejemplo de esto es la caracterización de los Documentos de Archivos Electrónicos:

…un documento archivístico electrónico, si es para ser útil a la ciencia debe ser continuamente extendido, éste debe ser almacenado y sobre todo consultado…

Lo que puede derivarse de la naturaleza del documento de archivo, sea electrónico o no, desde la variedad de definiciones usadas tanto en la investigación como en la práctica archivística (…) es que un documento de archivo está siempre asociado con una acción o evento, como un agente, producto o subproducto; un documento de

archivo incluye, como mínimo, un conjunto identificable de metadatos que sirven para aportar evidencia acerca de la acción o el evento (Gilliland-Swetland 2005).

1.3 LA DESCRIPCIÓN ARCHIVÍSTICA

Después de esclarecer los conceptos que forman parte del objeto de la archivística, esta sección se va a centrar en la fase de elaboración del documento dentro de la descripción archivística. Además, en los principales metadatos que incluye la Norma Internacional para la Descripción Archivística ISAD (G), junto al cuadro de clasificación de los documentos que se proponen en ella, van a ser examinados debido al impacto que presentan en la descripción del entorno de la investigación.

La finalidad de la descripción archivística es identificar y explicar el contexto y el contenido de los documentos de archivo con el fin de hacerlos accesibles. Esto se consigue mediante la elaboración de unas representaciones precisas y adecuadas que se organizan de acuerdo con unos modelos predeterminados. Los procesos descriptivos pueden empezar con anterioridad o ser simultáneos a la producción de los documentos y continuar a lo largo de todo su ciclo vital. Estos procesos permiten establecer los controles intelectuales necesarios para que las descripciones fiables, auténticas, significativas y accesibles puedan mantenerse a través del tiempo.

La automatización de las técnicas archivísticas y de la descripción en especial, ha sido el motor de arranque para que la normalización pase de ser una aspiración a convertirse en una realidad con perfiles cada vez más nítidos (Ahn, Brusilovsky et al. 2007).

El Consejo Internacional de Archivos viene desarrollando desde 1989 diversas iniciativas encaminadas a establecer la normativa para la descripción archivística. En enero de 1992 durante la reunión celebrada en Madrid, se adoptó el texto de la declaración de principios sobre la descripción archivística. Se redactó el Proyecto ISAD (G): Norma Internacional General de Descripción Archivística, que el citado congreso animó a desarrollar, cuya versión definitiva se aprobó a comienzos de 1993 (Estocolmo).

1.3.1 NORMA INTERNACIONAL GENERAL DE DESCRIPCIÓN ARCHIVÍSTICA: ISAD (G)

Esta norma contiene reglas generales para la descripción archivística que pueden aplicarse con independencia del tipo documental o del soporte físico de los documentos de archivos. Este conjunto de reglas generales forman parte de un proceso dirigido a:

- Garantizar la elaboración de descripciones coherentes, pertinentes y explícitas;
- Facilitar la recuperación y el intercambio de información sobre los documentos de archivos;
- Compartir los datos de autoridad y hacer posible la integración de las descripciones procedentes de distintos lugares en un sistema unificado de información.

ISAD (G) se encuentra estructurada en siete áreas de información descriptiva y 26 campos de descripción de los cuales seis deben utilizarse necesariamente en todos los casos:

- El código de referencia.
- El título.
- El productor.
- La(s) fecha(s).
- La extensión de la unidad de descripción.
- Nivel de descripción.

Esta norma tipifica una organización multinivel donde se describe al fondo como un todo; este debe representarse en una descripción utilizando los elementos de la descripción. Si es necesario describir las partes que integran el fondo, estas pueden describirse por separado. La suma total de todas estas descripciones, jerárquicamente unidas entre sí (Figura 1.1[8]), representan el fondo y las partes descritas. Esto se denomina descripción multinivel.

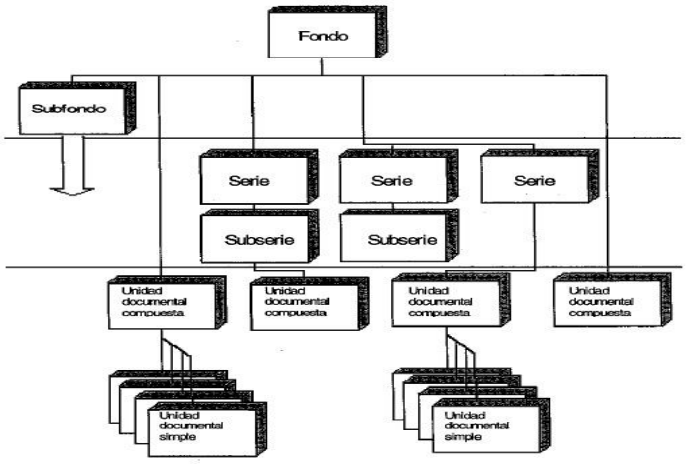

Figura 1.1. Sistema de organización por niveles.

Al establecer una jerarquía de descripciones deben aplicarse las cuatro reglas fundamentales que se especifican a continuación:

- Descripción de lo general a lo particular.
- Información pertinente para el nivel de descripción.
- Vinculación de las descripciones.
- No repetición de la información.

El fondo constituye el nivel más amplio de descripción, y las partes los niveles sucesivos, cuya descripción a menudo solo resulta plenamente significativa si se contempla en el contexto de la descripción de todo el conjunto del fondo. Pueden existir descripciones a nivel de fondo, a nivel de serie, a nivel de expediente y a nivel de documento. Cada uno de estos niveles puede a su vez sub-dividirse dependiendo de la complejidad de la estructura administrativa y de las funciones de la organización

[8] Tomada de la Norma Internacional de Descripción de Archivo ISAD (G)

que generó los documentos de archivos, así como de la propia organización de la documentación.

1.4 SISTEMA DE RECOMENDACIÓN

En la actualidad muchas son las definiciones que aparecen para concretar el significado de los Sistemas de recomendación. En este sentido, como se introdujo anteriormente, según (Cordón 2008), el significado de "Sistema de recomendación" se puede ver como cualquier sistema que produce recomendaciones individuales como salida, o que tiene el efecto de guiar al usuario de un modo personalizado a objetos útiles o interesantes dentro de un gran espacio de posibles opciones. Por otra parte se puede ver como herramientas de software y técnicas que proveen sugerencias o recomendaciones acerca de los ítems que pueden ser útiles para un usuario (Resnick and Varian 1997; Mahmood and Ricci 2009). El objetivo de un Sistema de recomendación es generar recomendaciones significativas a un conjunto de usuarios para los artículos o productos que podrían interesarles (Resnick, Iacovou et al. 1994).

Aunque cada una de las definiciones expuestas con anterioridad convergen hacia un mismo fin, en la presente investigación se puede pensar en un Sistema de Recomendación como una solución que enriquece los diferentes entornos, centrándose en facilitar el tránsito de los usuarios por un gran cúmulo de información, entregando a este último una serie de elementos relevantes según sus afinidades. Desde este punto de vista, más allá de la implementación de técnicas que soportan el funcionamiento de los SR se hace imprescindible para su diseño, la identificación de las características particulares del dominio y los elementos que se pretenden recomendar.

En la actualidad son muchos los sistemas que incorporan SR. Debido al constante crecimiento en la red de redes el número de aplicaciones web, en las diferentes áreas, que cuentan con las ventajas del uso de un SR crece de manera exponencial. Esto está dado principalmente, por la cantidad de información que generalmente se le presenta a un usuario. Algunos ejemplos del funcionamiento de los SR en el mundo real son las sugerencias de libros de Amazon[9], las películas en Netflix[10], los videos en YouTube[11] o música en Last.fm[12].

Como ya se ha mencionado con anterioridad los SR ayudan a las personas a encontrar elementos en dominios regidos por un gran cúmulo de información. Estas herramientas tienen como objetivo proporcionarles a los usuarios elementos útiles de acuerdo a sus necesidades o gustos. Normalmente, las recomendaciones se obtienen a partir de una o varias fuentes de información: preferencias de los usuarios entre los elementos alternativos, preferencias de los usuarios respecto a los atributos de los elementos, preferencias o elecciones de otras personas, valoraciones de expertos o características individuales que pueden predecir las preferencias (Noguera, Barranco et al. 2012).

[9] www.amazon.com
[10] www.netflix.com
[11] www.youtube.com
[12] www.last.fm

En este sentido en dependencia de la fuente de información y la técnica que se utilice para clasificar los elementos se pueden mencionar varios diseños de sistemas de recomendación:

- *Sistema de Recomendación Basado en el Contenido (Pazzani, Muramatsu et al. 1996):* Este tipo de sistema de recomendación computa las recomendaciones de acuerdo a las características de los elementos que el usuario prefirió en el pasado.
- *Sistema de Recomendación Colaborativo (Goldberg, Nichols et al. 1992):* Usa las calificaciones de los usuarios para agrupar a estos en grupos, teniendo en cuenta el grado de similitud que existe entre ellos. Las recomendaciones se infieren entonces, teniendo en cuenta las valoraciones de los usuarios que pertenecen al mismo grupo.
- *Sistema de Recomendación Demográfico (Krulwich 1997):* Se clasifican los usuarios en grupos demográficos basados en los atributos personales. Los usuarios reciben recomendaciones según los grupos en los que se encuentran clasificados.
- *Sistema de Recomendación basados en el Conocimiento (Burke 2000):* Estos sistemas infieren las recomendaciones basados en las necesidades de los usuarios y en el conocimiento de las características de los elementos. Frecuentemente las recomendaciones se determinan haciendo uso de sistemas basados en casos. Un ejemplo de aplicación de lo explicado anteriormente es la de pedir al usuario que especifique un elemento de interés y el sistema entonces infiere un perfil para el usuario teniendo en cuenta el producto que mejor coincide en el espacio de búsqueda.
- *Sistema de Recomendación Basado en la Utilidad (Guttman 1998):* Determinan las recomendaciones basados en el cálculo de la utilidad de cada elemento de acuerdo a los intereses del usuario.
- *Sistema de Recomendación Híbrido (Basu, Hirsh et al. 1998; Burke 2002; Liu, Lai et al. 2009):* Los sistemas mencionados con anterioridad sufren de limitaciones bajo ciertas circunstancias. Los SR Híbridos tratan de eludir estas limitaciones combinando dos o más acercamientos diferentes.

Es importante señalar que la presente investigación va a centrarse en los SRBC, los SRC y las posibles hibridaciones que se pueden obtener a partir de estos enfoques. Esto está basado en el principal resultado que se espera de la investigación: Caracterización de los entornos regidos por documentos de archivo, con la finalidad de difundir los distintos acervos históricos, para el diseño de un SR.

1.4.1 TAREAS DE UN SISTEMA DE RECOMENDACIÓN

Los usuarios que desean hacer uso de un Sistema de recomendación pretenden por lo general que este soporte de manera efectiva las tareas o metas que los SR persiguen. En este sentido (Burke 2007) en un artículo que se ha convertido en una referencia clásica en este campo (Ricci, Rokach et al. 2011) se definen 11 tareas que ayudan a la implementación de un SR:

- *Encontrar un elemento relevante:* Recomendar a un usuario algunos elementos como una lista ordenada teniendo en cuenta qué tan relevante es el elemento para el usuario (por ejemplo, en una escala de uno a cinco estrellas).

- *Encontrar todos los elementos relevantes:* Recomendar todos los elementos que pueden satisfacer algunas necesidades de los usuarios. En algunos casos es ineficiente encontrar solamente algunos elementos relevantes. Especialmente cuando la misión del SR es de suma importancia o cuando el número de elementos es pequeño. Examinar detenidamente todos los posibles elementos a recomendar podría derivar ciertos beneficios, tales como, la clasificación de los elementos que se muestran, atendiendo al grado de relevancia para el usuario o la explicación adicional que los SR brindan junto a sus recomendaciones.
- *Anotación en el contexto:* Teniendo en cuenta el contexto existente, por ejemplo, una lista de elementos, se puede enfatizar en algunos de ellos en función de las preferencias del usuario a largo plazo. Un sistema que recomienda programas de televisión puede anotar que programas de televisión aparecen en la guía de programación electrónica.
- *Recomendar una secuencia*: En lugar de centrarse en la recomendación de un solo elemento, la idea es presentar una serie de elementos que son agradables en su conjunto.
- *Recomendar un conjunto:* Sugerir un grupo de elementos que se integren bien. Por ejemplo un plan de viaje puede estar compuesto de varios lugares de interés, destinos y lugares de acogida. Teniendo en cuenta estos elementos el usuario puede escoger un destino turístico único.
- *Navegar:* En esta tarea el usuario puede navegar por diferentes entornos sin la intención de acceder a ningún elemento, en este sentido el SR debe ser capaz de ayudar al usuario para ver qué elementos son de mayor interés para él.
- *Encontrar recomendaciones creíbles:* Algunos usuarios no confían en los sistemas de recomendación por lo que juegan con ellos para ver que tan buenos son en la formulación de las recomendaciones. Es por esto que el sistema también debe ofrecer funciones específicas para que los usuarios prueben su comportamiento.
- *Mejorar el perfil:* Esto está relacionado con la capacidad del usuario para proporcionar información al SR referente a sus gustos y preferencias, es una tarea fundamental que es estrictamente necesaria para proporcionar recomendaciones personalizadas. Si el sistema no tiene conocimiento sobre el usuario activo entonces solo les proporcionará las mismas recomendaciones que se entregan a un usuario promedio.
- *Expresarse:* Para algunos usuarios las recomendaciones no son de interés en lo absoluto. Más bien, lo que es importante para ellos es contribuir con sus valoraciones y expresar sus opiniones y creencias. La satisfacción de los usuarios en esa actividad puede actuar como una palanca para el usuario que sostiene firmemente a la aplicación.
- *Ayudar a otros:* A varios usuarios les satisface la idea de contribuir con sus valoraciones, porque creen que la comunidad se beneficia con su contribución. Esto puede ser una motivación para toda la información que rutinariamente no es usada por los sistemas de recomendación.
- *Influir en los demás:* En los SR de la Web, existen usuarios que tienen como objetivo influenciar explícitamente a otros usuarios para que adquieran determinado producto. De esta manera, pueden aparecer usuarios que usen el sistema de forma intencionada para promover o personalizar ciertos elementos.

El diseño del SR soportado en la caracterización de entornos archivísticos, puede centrar sus principales técnicas y componentes en cada uno de los acercamientos anteriormente descritos. Se podría incorporar la presentación de los elementos relevantes para un usuario a partir de la clasificación de los elementos haciendo uso de estereotipos gráficos, ocultando de esta forma el valor de utilidad determinado por la técnica que se emplee. Realizando un análisis de la cantidad de documentos archivísticos presentes en cada uno de los sistemas o fondos documentales explorados por el SR se pueden analizar todos los elementos (mientras el número de elementos no sobrepase un umbral que se determine según las características del entorno).

En este mismo sentido la técnica a usar para la recomendación de elementos o las hibridaciones de estas descansan en tareas como, la obtención de recomendaciones creíbles, navegar o recomendar un conjunto de elementos, por mencionar algunas. La retroalimentación del SR en aras de mejorar la precisión en que se recomiendan los elementos puede enfocarse en la ayuda de otros usuarios, la mejora de los perfiles o sencillamente evitar la influencia de terceros para resaltar o poner en desventaja ciertos elementos.

En las próximas secciones se discutirán diseños que explotan la mayoría de los aspectos vistos hasta este punto, con la finalidad de identificar la forma adecuada de obtener recomendaciones de interés para el usuario final.

1.5 SISTEMAS DE RECOMENDACIÓN BASADOS EN CONTENIDOS

Como se ha mencionado con anterioridad, este diseño realiza un estudio de los gustos de los usuarios teniendo en cuenta las características de los elementos que ha valorado en el pasado. Así, para cada elemento no valorado se puede calcular un valor de utilidad que indica si este es acorde con los gustos o no de los usuarios.

La manera en que más se han estudiado los SRBC así como los SRC se presenta en la Figura 1.2:

		Elementos					
		1	2	... *i* ...			*m*
Usuarios	1	5	3		1	2	
	2		2				4
	.			5			
	.	3	4		2	1	
	u					4	
	.				3	2	
	.						
	n	3	5		?	1	

Figura 1.2. Matriz de valoraciones.

Las preferencias conocidas de los usuarios son representadas en una matriz de *n* usuarios y *m* elementos, donde cada celda $r_{u,i}$ corresponde con la valoración de un usuario *u* para un elemento *i*. Normalmente las valoraciones están esparcidas, es decir los usuarios no valoran todos los elementos. La tarea de la recomendación es predecir qué valoración le daría el usuario a los elementos no valorados. Habitualmente la

predicción se realiza para todos los elementos que no han sido votados por el usuario y los mayores valores son los presentados como recomendación. El usuario bajo estas consideraciones para la recomendación es llamado usuario activo (Melville and Sindhwani 2010).

En (Lops, de Gemmis et al. 2011) se propone una arquitectura base sobre la cual debe estar sustentado el diseño de un SRBC. Este proceso de recomendación es desarrollado en tres pasos, donde cada paso es manejado por un componente diferente:

Analizador de Contenido: Cuando la información no tiene una estructura (por ejemplo, texto), se necesita de alguna especie de etapa de pre-procesamiento para extraer la información en forma estructurada. La principal responsabilidad de este componente es representar de forma adecuada el contenido de los elementos procedentes de cierta fuente de información para los siguientes pasos del proceso. Esta representación es la entrada al componente Construcción del Perfil y al Componente de Filtrado.

Construcción del Perfil: Este módulo recoge datos representativos de las preferencias del usuario y trata de generalizar estos datos, con el fin de construir el perfil de usuario. Usualmente la estrategia de generalización se realiza a través de técnicas de aprendizaje automático, que son capaces de inferir un modelo de los intereses del usuario a partir de los elementos que les gustaron o no en el pasado.

Componente de Filtrado: Este módulo explota el perfil de usuario para sugerir elementos relevantes, comparando la representación del perfil de usuario contra la del elemento que se pretende recomendar. El resultado es un valor de relevancia que puede ser binario o continuo (este es determinado haciendo uso de alguna métrica de similitud (Holte and Yan 1996)).

El primer paso en el proceso de recomendación es el que realiza el Analizador de Contenido, que como ya se mencionó toma las descripciones de los elementos que provee alguna fuente de información para procesarlo y obtener la representación de los elementos de forma estructurada. Con el objetivo de construir y actualizar el perfil del usuario activo, las interacciones de este con los elementos son coleccionadas y almacenadas. Esta interacción se denomina anotaciones o retroalimentación, las cuales junto a la representación de los elementos son explotadas durante el proceso de construcción del perfil para predecir la relevancia de un nuevo elemento presentado. Los usuarios pueden definir sus áreas de interés a través de un perfil inicial sin necesidad de facilitar retroalimentación alguna.

Para construir el perfil del usuario activo el componente responsable aplica algoritmos de aprendizaje supervisado que posteriormente será usado por el Componente de Filtrado. A partir de la descripción de un nuevo elemento el Componente de Filtrado determina si es relevante para el usuario activo comparando esta descripción del elemento con el perfil del usuario. En este sentido, los elementos determinados de mayor relevancia, son incluidos en la lista de elementos que será presentada al usuario activo.

Las preferencias de los usuarios con el tiempo pueden cambiar, por lo que es necesario además mantener y actualizar la información en el componente de Construcción del Perfil. Después de obtener una retroalimentación de los elementos presentados por el sistema el proceso de aprendizaje debe realizarse de nuevo en aras de actualizar los intereses de los usuarios.

La mayoría de los SRBC utilizan modelos de recuperación relativamente simples, tales como los basados en comparación de palabras claves o los modelos basados en espacios vectoriales. Estos últimos son una representación espacial de documentos de texto, donde cada documento es representado como un vector de n-dimensiones en el espacio. Cada dimensión del vector representa un término de todo el vocabulario que provee la colección de documentos. Cada documento d_j es representado como un vector de pesos de n-dimensiones, siendo $d_j = \{w_{1j}, w_{2j}, \dots, w_{kj}\}$, donde w_{kj} es el peso para el termino t_k (termino perteneciente al diccionario que se obtiene de todos los documentos) en el documento d_j.

La representación de los documentos de esta forma plantea dos cuestiones, ponderación de los términos y la medición de la similitud vectorial. El esquema de ponderación más utilizado es TF-IDF (del inglés Term Frecuency – Inverse Document Frecuency) donde TF se determina como:

$$TF(t_k, d_j) = \frac{f_{k,j}}{max_z f_{z,j}}$$

Donde $f_{k,j}$ es la frecuencia del termino *k* en el documento *j* y el dividendo de la expresión es el termino de mayor frecuencia sobre todos los términos del vocabulario que aparecen en el documento d_j. Por otra parte IDF se determina como:

$$IDF(t_k) = log \frac{N}{n_k}$$

Donde *N* es el número total de documentos y n_k el número de documentos donde el termino t_k ocurre al menos una vez. De manera general se puede decir entonces que $TF\text{-}IDF = TF(t_k, d_j) * IDF(t_k)$. Con el fin de obtener valores de pesos en el intervalo [0,1] y los documentos sean representados por vectores de igual longitud, se determina el peso normalmente a partir de la normalización del coseno:

$$w_{k,j} = \frac{TF\text{-}IDF(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} TF\text{-}IDF(t_s, d_j)^2}}$$

Luego, se hace necesario el uso de una métrica de similitud para determinar la cercanía entre los documentos. Una de las más usadas es la similitud del coseno (Melville and Sindhwani 2010):

$$sim(d_i, d_j) = \frac{\sum_k w_{k,i} * w_{k,j}}{\sqrt{\sum_k {w_{k,i}}^2} * \sqrt{\sum_k {w_{k,j}}^2}}$$

Los SRBC basados en espacio vectorial representan los elementos y el perfil de usuarios como vectores de términos de peso, de ahí que determinar si cierto elemento es de interés o no para el usuario, se realiza aplicando la similitud del coseno.

En el estado del arte que realiza (Lops, de Gemmis et al. 2011) se pueden apreciar varios sistemas basados en palabras clave en diferentes áreas de aplicación tales como música, noticias, comercio electrónico, películas, entre otras. Cada dominio presenta diferentes problemas lo que hace que requieran de diferentes soluciones.

Además de estos métodos tradicionales heurísticos, otras técnicas para la recomendación de elementos basadas en el contenido han sido usadas, tales como los clasificadores Bayesianos (Pazzani, Muramatsu et al. 1996) y varias técnicas basadas en máquinas de aprendizaje, incluyendo clustering, arboles de decisión y redes neuronales artificiales. Estas técnicas difieren de las provenientes de la Recuperación de la Información en que ellas determinan la utilidad de la predicción basado en modelos de aprendizaje que usan valores estadísticos y no métodos heurísticos tales como la medida de similitud del coseno. Por ejemplo, basados en un conjunto de páginas Web que fueron valoradas como relevantes e irrelevantes por el usuario, (Sahlgren 2006) usa un clasificador Bayesiano para clasificar las páginas no valoradas. Específicamente el clasificador es usado para estimar la probabilidad de que cierta página p_j pertenezca a cierta clase C_i (relevante o irrelevante) dado un conjunto de palabras claves $k_{i,j}, \dots, k_{n,j}$ de la página:

$$P(C_i | k_{i,j}, \& \dots \&, k_{n,j})$$

(Sahlgren 2006) asumió además que las palabras claves son independientes y que la probabilidad expuesta en la formula anterior es proporcional a:

$$P(C_i) \prod_x P(k_{x,j} | C_i)$$

Mientras que esta asunción no aplica necesariamente en muchas aplicaciones, resultados experimentales demuestran que el clasificador Bayesiano de naîve produce altos valores de exactitud en la clasificación (Sahlgren 2006). El clasificador Bayesiano de naîve, ha sido usado en varios Sistemas de recomendación basados en el contenido, tales como Syskill & Webert (Pazzani, Muramatsu et al. 1996), NewsDude (Billsus and Pazzani 1999), Daily Learner (Billsus and Pazzani 2000), LIBRA (Mooney and Roy 2000) and ITR (Semeraro 2009).

Otro ejemplo de clasificador, linear en este caso, es el método de Rocchio. En este algoritmo se representan los documentos como vectores, de esta manera cada par de documentos con contenido similar tendrán vectores similares. Cada una de las dimensiones del vector corresponderá con términos que aparecen en el documento. El peso de cada componente se determina haciendo uso del esquema TF-IDF ya analizado. Luego, el aprendizaje es alcanzado combinando vectores (de ejemplos positivos y negativos) en un vector prototipo para cada clase del conjunto de clases C. Para clasificar un nuevo documento *d*, primero se determina la similitud entre los vectores prototipos de cada una de las clases y el nuevo documento *d* (por ejemplo,

usando la similitud del coseno). Luego, *d* es asignado a la clase cuyo vector prototipo tenga mayor valor de similitud.

Formalmente, el método de Rocchio calcula un clasificador $\overrightarrow{c_i} = \langle w_{1i}, \ldots, w_{|T|i} \rangle$ para la categoría $c_i$determinado por:

$$w_{ki} = \beta \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|POS_i|} - \gamma \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|NEG_i|}$$

Donde w_{kj} es el TF-IDF del término t_k en el documento d_j, POS_i y NEG_i son el conjunto de ejemplos positivos y negativos en el conjunto de entrenamiento para una clase determinada c_j, β y γ son parámetros de control que permiten modificar la importancia de todos los ejemplos positivos y negativos. Para asignar un documento d_j a una clase $\hat{c}$, se determina la similitud entre cada vector prototipo $\overrightarrow{c_i}$ y el vector del documento $\overrightarrow{d_j}$ y $\hat{c}$ va ser de c_i con el valor más alto. Este clasificador ha sido usado en SR, tales como YourNews (Ahn, Brusilovsky et al. 2007), Fab (Balabanović and Shoham 1997) and NewT (Sheth and Maes 1993).

Otras soluciones basadas en el contenido presentan modelos más complejos como es el Análisis Semántico haciendo uso de Ontologías o Fuentes de Conocimiento Enciclopédicas. El primero de los mencionados permite perfiles más precisos que contienen referencias a los conceptos definidos en las bases de conocimientos externas. La motivación fundamental de este enfoque es el reto de proporcionar un Sistema de recomendación con la cultura y los conocimientos lingüísticos que caracterizan la capacidad de interpretar los documentos en leguaje natural y el razonamiento de su contenido. Por otra parte, el análisis semántico a partir de fuentes de conocimiento enciclopédicas se centra en la idea de que la construcción del perfil de usuario se puede ver beneficiada de la infusión del conocimiento desde el exterior, respecto a la utilización clásica de los conocimientos extraídos de los propios documentos (Lops, de Gemmis et al. 2011).

La adopción de un SRBC presenta varias ventajas cuando se compara con los SRC:

- *Independencia del usuario:* Los SRBC solamente tienen en cuenta las calificaciones proporcionadas por el usuario activo para construir su propio perfil.
- *Transparencia:* Las explicaciones de como el SRBC obtuvo la lista de recomendaciones que se presentan al usuario, se pueden dar de manera explícita a partir de las características de los elementos o la descripción de estos. Estas características se pueden tomar como indicadores para decidir si confiar o no en la recomendación.
- *Elemento nuevo:* Son capaces de recomendar elementos que no han sido aún valorados por ningún usuario.

Sin embargo los SRBC presentan algunas limitaciones:

- *Análisis de Contenido Limitado:* Los diseños basados en el contenido tienen un límite natural en el número y tipo de características que representan a los elementos. En este sentido, no importa si la técnica que se aplique sea manual

o automática, esto no es suficiente para obtener los intereses del usuario. El conocimiento del dominio en muchas ocasiones es necesario para poder determinar cuáles atributos resultan relevantes para los elementos.

- *Exceso de Especialización:* No tienen ningún método inherente para garantizar la aparición de un elemento no esperado. Es decir, el sistema sugiere elementos cuyos resultados en la tarea de la predicción sean altos cuando se compara con el perfil del usuario, por lo que el usuario solo recibirá como recomendación elementos que sean similares a aquellos que ya haya valorado.
- *Usuario Nuevo:* El sistema debe tener calificaciones suficientes para realmente poder entender las necesidades e intereses de los usuarios y proporcionar información precisa. Por lo tanto, cuando pocas calificaciones están disponibles, para un nuevo usuario, el SRBC no va a ser capaz de proporcionar recomendaciones fiables.

La explotación de un SRBC como punto de partida para cualquier sistema podría resultar interesante. El beneficio de cada una de sus ventajas, junto a la utilización de elementos que rijan el entorno donde se pretenda alojar el SR, podría convergir hacia resultados satisfactorios. Algunas de las limitaciones de los SRBC se mitigan con el enfoque colaborativo.

1.6 SISTEMAS DE RECOMENDACIÓN COLABORATIVOS

Los SRC o el Filtrado Colaborativo trabaja recolectando información de los usuarios a partir de las valoraciones de los elementos de un determinado dominio y brinda las recomendaciones teniendo en cuenta la similitud entre las valoraciones de varios usuarios. Los métodos de filtrado colaborativos pueden dividirse en los métodos basados en la Vecindad y los basados en el Modelo. Los primeros, comúnmente, también se les conocen como los acercamientos basados en memoria (Breese, Heckerman et al. 1998).

En el método basado en la Vecindad las valoraciones de los usuarios sobre los elementos almacenados en el sistema son usadas directamente para predecir la valoración para un elemento nuevo. Esto puede realizarse en dos formas conocidas como las recomendaciones basadas en usuarios y las basadas en elementos (Desrosiers and Karypis 2011).

Los enfoques mencionados anteriormente pueden ser generalizado por un algoritmo que se resume en los siguientes tres pasos:

- *Primero:* Asignar pesos a todos los usuarios respecto a la similitud de estos con el usuario activo.
- *Segundo:* Seleccionar los k usuarios que tengan los mayores valores de similitud respecto al usuario activo normalmente denominado Vecindad.
- *Tercero:* Determinar la predicción de la valoración para el elemento en cuestión a partir de las calificaciones de la vecindad seleccionada.

Para una mejor comprensión de la aplicación de estos pasos a seguir para predecir la valoración de un elemento se asumirá que la matriz que sustenta cada momento del proceso es la representada en la Figura 2.

En el primer paso, el peso $w_{a,u}$ es una medida de similitud entre el usuario *u* y el usuario activo *a*. En la sección anterior se introdujo la similitud del coseno, existen otras definidas en la literatura como la correlación de Spearman, correlación T de Kendall, Diferencias de las Medias Cuadráticas y Entropía (Herlocker, Konstan et al. 1999; Su and Khoshgoftaar 2009). Estudios realizados muestran que la correlación de Pearson es la que mejores resultados brinda (Breese, Heckerman et al. 1998):

$$w_{a,u} = \frac{\sum_{i \in I}(r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i \in I}(r_{a,i} - \bar{r}_a)^2 \sum_{i \in I}(r_{u,i} - \bar{r}_u)^2}}$$

Donde *I* es el conjunto de elementos valorados por ambos usuarios (*u* y *a*), $r_{u,i}$ es la valoración que recibe el elemento *i* del usuario *u* y $\bar{r}_u$ es la valoración media dada por el usuario *u*.

Luego, se seleccionan los k usuarios más cercanos al usuario activo como se especificó en el paso dos, aunque es válido añadir que solo los usuarios que hayan valorado el elemento -para el cual se desea calcular la predicción- serán seleccionados. Si se denomina este conjunto de vecinos como $N_i(u)$ se puede predecir el valor de $r_{u,i}$ (calificación promedio dada a *i* por estos vecinos) como (Desrosiers and Karypis 2011):

$$\hat{r}_{ui} = \frac{1}{|\, N_i(u)|} \sum_{v \in N_i(u)} r_{vi}$$

Esto presenta un problema y es que no tiene en cuenta que los vecinos seleccionados pueden tener diferentes niveles de similitud. Una solución común a este problema es la ponderación de la contribución de cada vecino de acuerdo a su similitud con el usuario activo de la siguiente manera:

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i(u)} w_{uv} r_{vi}}{\sum_{v \in N_i(u)} |w_{uv}|}$$

Este resultado también presenta un problema importante: no tiene en cuenta el hecho de que los usuarios pueden utilizar diferentes valores de clasificación para cuantificar el mismo nivel de apreciación para un elemento. Por ejemplo, un usuario puede dar el valor más alto de calificación a un número pequeño de elementos, siendo estricto en este sentido, mientras que otro puede valorar con la máxima escala aquellos elementos que simplemente sean de su agrado. Este problema suele afrontarse normalizando $r_{u,i}$. Dos de los más populares esquemas de normalización de la predicción son: mean-centering y Z-score. Estudios realizados por (Howe A 2008) muestra que el uso de Z-score tiene mejores beneficios. En este sentido normalizando la predicción $r_{u,i}$ con Z-score:

$$\hat{r}_{ui} = \bar{r}_u + \sigma_u \frac{\sum_{v \in N_i(u)} w_{uv}(r_{vi} - \bar{r}_v)/\sigma_v}{\sum_{v \in N_i(u)} |w_{uv}|}$$

Hasta este punto se ha caracterizado el acercamiento basado en usuarios más cercanos. Cuando se aplica este enfoque a millones de usuarios, el sistema no

presenta una buena escalabilidad debido a la complejidad computacional que arroja determinar los usuarios similares al usuario activo. Si el entorno cuenta con una diferencia significante entre usuarios y elementos, siendo estos últimos menos frecuentes (Desrosiers and Karypis 2011), una alternativa propuesta en (Linden, Smith et al. 2003) es el acercamiento basado en el elemento. Otro criterio que se puede seguir es la estabilidad, en este sentido es preferente hacer uso de los sistemas colaborativos basados en elementos si la dinámica en los cambios del medio respecto a los elementos es significativamente baja, es decir que no varíen en el tiempo de manera exponencial los elementos presentes. De manera análoga al enfoque usuario-usuario, según las características del acercamiento basado en elementos se puede determinar la similitud entre dos elementos aplicando Pearson como sigue:

$$w_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U}(r_{u,i} - \bar{r}_i)^2}\sqrt{\sum_{u \in U}(r_{u,j} - \bar{r}_j)^2}}$$

Donde *U* es el conjunto de usuarios que han valorado ambos elementos *i* y *j*, $r_{u,i}$ es la valoración dada por el usuario u sobre el elemento *i*, $\bar{r}_i$ es el promedio de las valoraciones de *i* a través de todos los usuarios.

Luego la predicción de la valoración para el elemento deseado, aplicando de igual forma la normalización Z-score queda:

$$\hat{r}_{ui} = \bar{r}_i + \sigma_i \frac{\sum_{j \in N_u(i)} w_{ij}(r_{uj} - \bar{r}_j)/\sigma_j}{\sum_{j \in N_u(i)} |w_{ij}|}$$

La manera, antes descrita, en que se determina la predicción de las valoraciones donde estas son calculadas como la media de las valoraciones de la vecindad, esencialmente resuelve el problema de la regresión. Por otro lado vecindad basada en la clasificación encuentra la valoración más similar dada por el usuario *u* a un elemento *i*, teniendo la vecindad más cercana de u que ha valorado el ítem i. El voto v_{ir} dado por *k*-NN de *u* para la valoración *r* ϵ *S* puede ser obtenido como la suma de los pesos de similitud de la vecindad que han valorado el elemento *i*:

$$v_{ri} = \sum_{v \in N_i(u)} \delta\,(r_{vi} = r) w_{uv}$$

Donde $\delta(r_{vi} = r)$ es 1 si $r_{vi} = 1$ y 0 en cualquier otro caso. Una vez calculado este para cada posible valoración, la predicción para el ítem va ser simplemente, el v_{ir} donde el valor es mayor.

Un método basado en clasificación que considere la normalización de las valoraciones también puede ser definido. Sea S´ el conjunto de posibles valores a normalizar, la predicción de la valoración puede ser definida como:

$$\hat{r}_{ui} = h^{-1}(\arg max_{\, r \in S'} \sum_{v \in N_i(u)} \delta(h(r_{vi}) = r) w_{uv})$$

Donde *h* es la función de normalización de la valoración, que se puede determinar de la misma manera que se analizó para la normalización de valoración haciendo uso de mean-centering o z-score en el cálculo de este por regresión.

Seleccionar que tipo de implementación, regresión o clasificación basada en la vecindad, depende en gran medida de la escala de valoración del sistema. Si la escala de valoración en el sistema es continua, por ejemplo, puede tomar valores entre -10 y 10, entonces el método de regresión es más apropiado. Por otra parte, en sentido contrario, si la escala de las valoraciones solo cuenta con pequeños valores, por ejemplo "bueno" o "malo", o si los valores no pueden ser ordenados en una forma clara, entonces métodos de clasificación son preferidos. Desde el momento en que los esquemas de normalización tienden a mapear las valoraciones a una escala continua, en los métodos de clasificación la normalización puede ser más difícil de manejar (Desrosiers and Karypis 2011).

Ambos acercamientos presentados en esta sección, para el filtrado colaborativo basado en la vecindad, presentan dos defectos importantes:

- *Cobertura limitada:* La similitud entre dos usuarios se determina solo para aquellos usuarios que hayan valorados los mismos elementos. Esta suposición es muy limitante ya que los usuarios pueden haber valorado muy pocos o ningún elemento en común y aun así pueden ser similares. Otro problema es que solo los elementos valorados por los vecinos serán aquellos que en un futuro podrán ser recomendados.
- *Sensibilidad a la escasez de datos:* Es un problema común en la mayoría de los Sistemas de recomendación propiciado por el hecho de que los usuarios normalmente solo valoran una pequeña proporción de los elementos disponibles (Sarwar, Karypis et al. 2000). Esto se intensifica para los elementos recién añadidos al sistema, por lo general no tienen ninguna valoración en lo absoluto, problema que se conoce como arranque en frío (Schein, Popescul et al. 2002).

Una solución a estos problemas es completar la matriz de valoraciones con valores por defecto (Billsus and Pazzani 2000). Estos valores se pueden determinar a partir del valor medio del rango de valoración y el valor de valoración del usuario medio o aplicar un enfoque basado en contenido (Melville, Mooney et al. 2002). Finalmente, también se puede utilizar la similitud basada en el contenido en "vez de" o en "adición a" las valoraciones para determinar la vecindad. Sin embargo estas soluciones presentan sus propias limitaciones como por ejemplo, dar un valor por defecto a las valoraciones ausentes puede inducir ruidos en la recomendación, además como se vio en los SRBC la información de los elementos puede estar ausente para determinar la valoración o la similitud. Para resolver estos problemas en (Desrosiers and Karypis 2011) se presentan dos enfoques: métodos de reducción de la dimensión y basados en grafos. En (Koren and Bell 2011) se presentan técnicas avanzadas para el diseño de SRC.

1.7 SISTEMAS DE RECOMENDACIÓN HÍBRIDOS

Los Sistemas de recomendación híbridos combinan diferentes diseños de recomendación para eliminar problemas de un sistema en concreto o para aumentar la

precisión de las recomendaciones. Algunas de las hibridaciones que se le realizan a los SRC se implementan para mitigar uno de los problemas descritos con anterioridad, el arranque en frío. (Burke 2002; Burke 2007) presenta siete formas diferentes en que se pueden combinar los SR:

Por Pesos: El valor de la recomendación de un elemento se obtiene ponderando los diferentes resultados obtenidos por los SR. No obstante existen ocasiones en las que el resultado de una recomendación no se puede ponderar ya que el SR utilizado no ofrece un valor que expresa el grado de similitud del elemento. En estos casos, en lugar de ponderar los resultados se realiza la unión de los valores obtenidos por los distintos SR y el conjunto que se obtiene es llamado candidatos. Por ejemplo el sistema “P-Tango” (Claypool, Gokhale et al. 1999) usa un SRBC y SRC e inicialmente asigna el mismo peso a todos los elementos, a medida que los usuarios van interactuando con el sistema estos valores se van ajustando.

Conmutados: El sistema utiliza un criterio para establecer que sistema de recomendación utilizar en cada momento. Existen dos variantes:

- A partir de los resultados obtenidos se determinan que resultados mostrar.
- Se selecciona que Sistema de recomendación utilizar antes de procesar cualquier información.

Por ejemplo, Daily Learner utiliza un Sistema de recomendación basado en contenido para la recomendación de noticias, pero cuando este no tiene la suficiente confianza para realizar una recomendación, se cambia a un enfoque colaborativo (Burke 2002).

Mezclados: La utilización de más de un método de recomendación se utiliza simultáneamente, es decir, diferentes recomendaciones se presentan al mismo tiempo. El sistema PTV (Cotter and Smyth 2000) utiliza esta forma para recopilar recomendaciones sobre programas de televisión. Hace uso de diseños basados en el contenido para mostrar descriptores textuales de los programas y un enfoque colaborativo sobre las preferencias de los usuarios para la recomendación de los programas.

Combinación de Características: Las propiedades o rasgos de un SR son usadas, mediante una adaptación a otro tipo de Sistemas de recomendación. Esto puede traer consigo que no se vea como un Sistema de recomendación híbrido pero se considera lo contrario por hacer uso de varias fuentes del conocimiento: la combinación de características toma parte de la lógica de las recomendaciones de un tipo de diseño en lugar de tomar un componente como parte de la hibridación.

En Cascada: Se utiliza la salida de un diseño de recomendación cualquiera como entrada para aplicar cualquier otro acercamiento de recomendación. Se puede ver como un proceso de refinamiento de cada una de las iteraciones que este esquema pueda desarrollar.

Aumento de cualidades: Estas técnicas toman las salidas de un Sistema de recomendación que presenta elementos similares (objetos comparables) como incremento de las propiedades definidas por el segundo enfoque a utilizar para los elementos a recomendar. Por ejemplo se puede pensar en la confección de un

Sistema de recomendación basado en contenido que esté regido por libros. En este sentido se podrían tomar las recomendaciones de Amazon que presenta sus elementos por “títulos relacionados” y “Autores que también pueden resultar de interés” e incrementar la información de entrada para el sistema que se pretende implementar. Esta forma de hibridar añade calidad a las recomendaciones.

Meta Niveles: Esta técnica puede confundirse con la técnica de aumentos de cualidades, la principal diferencia es que en este modelo las recomendaciones que brinda el primer SR utilizado será el reemplazo de toda la base de conocimiento del diseño de recomendación que se va a utilizar a continuación. Por ejemplo, en (Basu, Hirsh et al. 1998) se utilizan técnicas de clasificación bayesianas para crear perfiles de preferencias de usuarios que se basan en el contenido.

1.8 CONCLUSIONES PARCIALES

Los conceptos asociados al dominio del problema, esclarecen cuales deben ser los principios a seguir para diseñar el SR en aplicaciones que pretenden difundir acervos históricos. La descripción de la norma ISAD (G) permite resaltar, además de una estructura y organización detallada de los elementos referentes a la descripción archivística, componentes que se pueden aprovechar (por ejemplo, la organización multiniveles, las áreas de descripción o los campos obligatorios para el intercambio de descripciones) para modelar los perfiles de usuarios, obtener recomendaciones más precisas e incrementar la fiabilidad de las recomendaciones.

Tanto los diseños de recomendación basados en el contenido, como los colaborativos presentan limitaciones propias. Algunas soluciones en función de reducir o eliminar las deficiencias de cada uno de los enfoques vistos, para obtener recomendaciones, se basan en los modelos de hibridación que se presentaron en el capítulo. Esto da una idea de cómo debe ser analizado la integración de las características del entorno y los ítems a recomendar.

CAPÍTULO 2 SISTEMA DE RECOMENDACIÓN EN ENTORNOS ARCHIVÍSTICOS: PRESENTACIÓN DE LA SOLUCIÓN PROPUESTA

La esquematización y caracterización de los componentes que representan una aplicación, sistema o entorno determinado, no pueden verse como un elemento aislado de la actividad científica. En el presente trabajo el término procedimiento hace referencia a cada uno de los componentes que representan el esquema de funcionamiento de la solución propuesta, junto a su descripción, forma en que se retroalimentan y se relacionan con los demás.

En el presente capítulo se explican cada uno de los componentes que conforman el esquema de funcionamiento del procedimiento propuesto. En primer lugar se presenta el modelo general con cada una de las partes que se integran en la solución. Para cada componente del esquema se hace una descripción de sus responsabilidades, junto a los datos que requiere de entrada y lo que provee como salida. En cada apartado se refleja cómo se benefician cada uno de los componentes del procedimiento propuesto, de las características del entorno (archivística). Cada componente del procedimiento que se presenta se nutre de los principales principios establecidos por la norma internacional ISADG (G), en aras de mejorar los resultados que se pretenden obtener.

2.1 DESCRIPCIÓN DEL PROCEDIMIENTO PROPUESTO

La recomendación de elementos parte de la representación de los elementos que se brindan por las distintas fuentes de información en dependencia del contexto donde se pretendan implementar. Una vez caracterizados los elementos se precisa de la descripción de los usuarios. Con la representación de los ítems y la comprensión de las preferencias de los usuarios el sistema de recomendación es capaz de determinar los elementos que pueden resultar de interés para un usuario. El resultado arrojado por el motor de recomendación es presentado al usuario, el cual puede emitir su criterio de las recomendaciones recibidas. Con la retroalimentación que obtiene el sistema, los perfiles de los ítems y los usuarios son actualizados en aras de alcanzar mayor precisión en el proceso de recomendación.

Figura 2.1. Esquema general de funcionamiento del procedimiento.

2.2 CARACTERIZACIÓN DE LOS COMPONENTES DE LA PROPUESTA DE SOLUCIÓN

En esta sección para cada componente presentado en la Figura 2.1 se va a realizar una descripción de cada una de sus partes (entrada, representación, análisis, salida). Como se puede apreciar lo primero que se necesita es la fuente de información, la cual se pretende representar y analizar para entregar a cada usuario aquella que represente sus intereses.

2.2.1 FUENTE DE INFORMACIÓN

La aplicación de cualquier Sistema de recomendación debe caracterizar los ítems a presentar a los usuarios. Es común que esta premisa se aplique con mayor medida en los SRBC, donde cada ítem puede ser descrito por el mismo número de atributos con valores conocidos. Por ejemplo, si se desea recomendar películas el conjunto de atributos que se puede tener en cuenta para describirlas son: actores, director, género, título, etc. Aunque en el caso de páginas Web, noticias, correos o documentos no es apropiado pues estos, normalmente son descritos por textos no estructurados (Lops, de Gemmis et al. 2011).

Aun cuando no se implemente un SRBC o no se cuente con una representación estructurada del contenido a recomendar, la representación uniforme de los elementos que rigen un entorno dado puede facilitar la presentación de los ítem recomendados y hasta cierto punto solventar el problema del arranque en frío. Se puede pensar en la categorización de los ítems o incluso en la agrupación en diferentes clases según la forma en que se describan, presentando al usuario el conjunto de elementos que pertenezcan a una misma clase.

Como se analizó en el Capítulo 1, en las entidades que tienen como fin la salvaguarda y difusión de los acervos históricos, se pueden encontrar documentos (cartas, libros, informes, manifiestos, cintas de audio, mapas, fotografía). La Norma Internacional General de Descripción Archivística ISAD (G) describe a un documento como la información que ha quedado registrada de alguna forma con independencia de su soporte o características.

Es evidente que la variedad de tipologías documentales que se pueden encontrar en un Archivo es uno de los elementos a tener en cuenta a la hora de representar la información. Esta diversidad de información, es una característica fundamental que debe ser considerada para el diseño de un SR. Aunque la gran mayoría de la información que se encuentra en un Archivo es textual, gráfica o auditiva, suponiendo inicialmente una dispersión en la tipología documental, ISAD (G) provee dentro de sus regulaciones características que hacen que la información se pueda representar uniformemente a partir de descripciones textuales. La presente investigación se centra en dos de los campos estipulados en ISAD (G) y en una de las reglas de descripción para los niveles de organización:

- título
- alcance y contenido
- descripción de cada uno de los niveles padres

Se puede ver la información contenida en los archivos como un conjunto de elementos que pertenecen a conjuntos disjuntos atendiendo a su soporte, ver Figura 2.2, haciendo uso de los descriptores textuales mencionados anteriormente se puede ver cada ítem como un documento de archivo independiente del soporte.

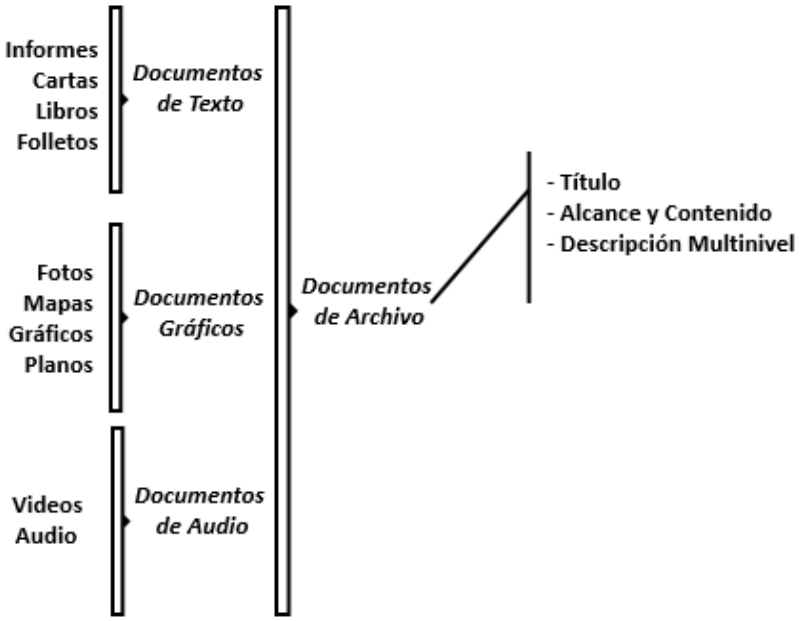

Figura 2.2. Representación de los documentos de archivo.

De aquí que la fuente de información en entornos archivísticos, independientemente del sistema o su forma de almacenamiento, debe proveer al menos los descriptores anteriormente mencionados. El primero de estos (título) pertenece al área de identificación (contiene la información esencial para identificar la unidad de descripción) de la norma ISAD (G) y pretende denominar la unidad de descripción. El segundo descriptor (alcance y contenido) perteneciente al área de contenido y estructura (contiene información relativa al objeto y organización de la unidad de descripción) proporciona información necesaria para apreciar el valor potencial de la unidad de descripción. Da una visión de conjunto (por ejemplo, periodos de tiempo, ámbito geográfico) y realiza un resumen del contenido (por ejemplo, tipos documentales, materia principal) de la unidad de descripción, apropiados al nivel de descripción. Por último, la descripción de cada uno de los niveles padres que contienen la unidad, viene dado por la característica multinivel (ver Figura 1.1) que propone la norma atendiendo a las reglas que especifica:

- Descripción de lo general a lo particular
- Información pertinente para el nivel de descripción
- Vinculación de las descripciones
- No repetición de la información

En el ejemplo, que se muestra a continuación (tomado de ISAD (G)) se puede comprender como se propone que quede representado cada una de las Unidades Documentales Simples. En este solo se muestran los campos que son de interés para el módulo Fuente de Información.

Descripción según ISAD (G):

Fondo

Title: Methodist Church (Canadá) Missionary Society fonds.

Scope and Content: Fonds Consists of the following series: General Board of Missions, 1965-1925; correspondence of the General Secretaries, 1968-1923; foreign mission records, 1888-1950; home mission records, 1906-1927; financial records, 1899-1930; quarterly returns of oboriginal institutes and day schools, 1902-1923; printed ephemera; and constitution and financial records of the Superannuation Fund for Lary Missionaries of Foreign Fields, 1919-1929.

Serie

Title: Records re foreign missions.

Scope and Content: Series consists of records re the following missions: West China. 1891-1931: West China Union University, 1896-1950; and Japan, 1873-1925.

Subserie

Title: West China Mission collection.

Scope and Content: Subseries consists of correspondence of the General Secretaries of the Methodist Church (Canadá) Missionary Society; copybook of W.J. Mortimore; minutes of the West China Mission Council; reports, financial records, property registers, manuscripts of historical and biographical studies, and other material relating to the evangelistic, pastoral, educational and medical work of the West China Mission.

Unidad Documental Compuesta

Title: Canadian Methodist Mission Property Register, West China.

Scope and Content: File consist of Canadian Methodist Mission Property Register pages, West China for Chengtu College University and Chengtu City.

Unidad Documental Simple (1)

Title: Chengtu, College University, No. 1, University Site, East of Administration Building skirting east and west road to Silk School with some breaks, 1914.

Unidad Documental Simple (2)

Title: Plan of Chengtu, College University, No. 1, University Site, East of Administration Building skirting east and west road to Silk School with some breaks [cartographic material].

En el ejemplo que se expone se tiene un Fondo documental compuesto por una Serie que contiene una Subserie, esta última cuenta con una Unidad Documental Compuesta por dos Unidades Documentales Simples. Para este ejemplo la *Unidad*

Documental Simple (1) quedaría representada en el componente Fuente de Información por el título de la misma, en este caso el *alcance y contenido* es el vacío y por último la suma de las descripciones de cada uno de los niveles antecesores representará el tercer campo (descripción niveles):

UDS (1) Representado en el componente Fuente de Información:

Título:

Chengtu, College University, No. 1, University Site, East of Administration Building skirting east and west road to Silk School with some breaks, 1914.

Alcance y Contenido: []

Descripción Niveles:

File consist of Canadian Methodist Mission Property Register pages, West China for Chengtu College University and Chengtu City.

Subseries consists of correspondence of the General Secretaries of the Methodist Church (Canadá) Missionary Society; copybook of W.J. Mortimore; minutes of the West China Mission Council; reports, financial records, property registers, manuscripts of historical and biographical studies, and other material relating to the evangelistic, pastoral, educational and medical work of the West China Mission.

Series consists of records re the following missions: West China. 1891-1931: West China Union University, 1896-1950; and Japan, 1873-1925.

Fonds Consists of the following series: General Board of Missions, 1965-1925; correspondence of the General Secretaries, 1968-1923; foreign mission records, 1888-1950; home mission records, 1906-1927; financial records, 1899-1930; quarterly returns of oboriginal institutes and day schools, 1902-1923; printed ephemera; and constitution and financial records of the Superannuation Fund for Lary Missionaries of Foreign Fields, 1919-1929.

Como se puede apreciar la *Descripción Niveles* en la representación de la UDS (1) no es más que la suma de los *Alcance y Contenido* de los niveles antecesores (Ver ejemplo tomado de ISAD (G)). Si se representa la UDS (2) excepto el título de la misma, los restantes valores almacenados (Alcance y Contenido, Descripción Niveles) son exactamente los mismos lo que evidencia la estrecha relación que existe entre estos documentos. Si para un usuario resulta de interés algunos de estos documentos evidentemente se puede pensar en la recomendación del otro.

El procedimiento que se propone en el trabajo de investigación brinda algunos de los elementos que se consideran de mayor relevancia para la caracterización de los ítems que provee el entorno. Esto no quiere decir o imponer, una normativa para la fuente de información en los distintos sistemas de gestión archivística, al contrario, muestra cómo se puede hacer uso de las características de los documentos de archivos según lo tipificado en la norma ISAD (G) para describir o comprender la información existente.

2.2.2 REPRESENTACIÓN DE LA INFORMACIÓN

En el apartado anterior se introdujeron algunos elementos que van guiando el proceso de recomendación en la archivística a partir de la representación de los elementos. La representación de la información para el diseño de cualquier Sistema de recomendación juega un papel fundamental.

En el componente Representación (Figura 2.1) es el encargado de llevar la información no estructurada, proveniente de la fuente de información, a una representación estructurada que permita a los restantes componentes la correcta interpretación de la misma. Por otro lado este componente tiene la responsabilidad de determinar, para cada ítem representado, los k-vecinos más cercanos atendiendo las ventajas que provee la representación multinivel de la norma de descripción de archivos ISAD (G). De igual manera es responsable de estructurar las interacciones de los usuarios con el sistema. A partir de lo descrito se pueden resumir las tareas del componente de Representación del procedimiento propuesto como:

- Representación de los ítems.
- Determinar y almacenar los k-vecinos para cada ítem.
- Representación de las interacciones y perfiles de usuarios.

La representación de los ítems, como se había mencionado en el apartado anterior, tiene su base en la estructura uniforme que se alcanza para cada documento de archivo (Figura 2.3). Para los documentos de archivos se torna compleja la definición de atributos, por lo que es conveniente el uso de técnicas, para la representación de estos, provenientes de la Recuperación de Información (Baeza-Yates and Ribeiro-Neto 1999). Para la representación de los documentos de archivo el procedimiento que se defiende en la investigación propone el uso de la representación basada en palabras claves VSM (del inglés Vector Space Model).

VSM es una representación espacial de los documentos de textos. En este modelo, cada documento es representado como un vector de n-dimensiones, donde cada dimensión corresponde con un término de todo el vocabulario de la colección de documentos (Lops, de Gemmis et al. 2011). Supongamos que contamos con *k* documentos que pertenecen a una colección *D*, para cada documento que pertenece a *D* se determinan las palabras claves, conformando así la representación de cada documento en la colección.

t_1	t_2	t_3	...	t_{k-1}	t_k

Figura 2.3 Representación vectorial basada en palabras claves.

Retomando el ejemplo de documento de archivo visto en el acápite anterior, la representación de la UDS 1 (*Chengtu, College University, No. 1, University Site, East of Administration Building skirting east and west road to Silk School with some breaks, 1914.*) estaría conformada por el vector de *n* dimensiones donde cada dimensión va a estar compuesta por las palabras que ocurren en el Título, Alcance y Contenido y Descripción Niveles. Esta representación facilita la implementación de un enfoque basado en el contenido sobre el procedimiento que se propone. En este sentido se

hace necesario el almacenamiento además de las anotaciones de los usuarios para luego poder inferir los intereses de este. Para la representación propuesta y retomando lo explicado en el Capítulo 1 (Ver Figura 1.2) el componente cuenta con una matriz de *M* x *N* filas y columnas donde el número de filas va a estar dado por la cantidad de usuarios y el número de columnas por la cantidad de documentos de archivo. De esta forma en cada $C_{m,n}$ celda de la matriz se corresponde con la valoración dada por el usuario U_m al documento D_n.

Estas representaciones no solo facilitan el cálculo de la utilidad de un documento para un usuario a partir de los intereses de este representado en la matriz, sino que garantiza además el uso de diseños colaborativos para la predicción de documentos de archivos que pueden resultar de interés para el usuario activo.

Una vez representados los documentos de la colección se pueden desarrollar otras de las tareas del componente, determinar la vecindad para cada documento. Esta vecindad puede ajustarse en dependencia de la cantidad de colecciones documentales existentes. Es válido detenerse a analizar el tamaño de la colección de documentos, pues si el número de documentos es significativo el cálculo de la vecindad según los principales algoritmos existentes puede ser costoso. En este sentido se puede aprovechar la representación multinivel propuesta en ISAD (G) con el fin de determinar los vecinos del documento en cuestión. A continuación se ejemplifica lo que se propone:

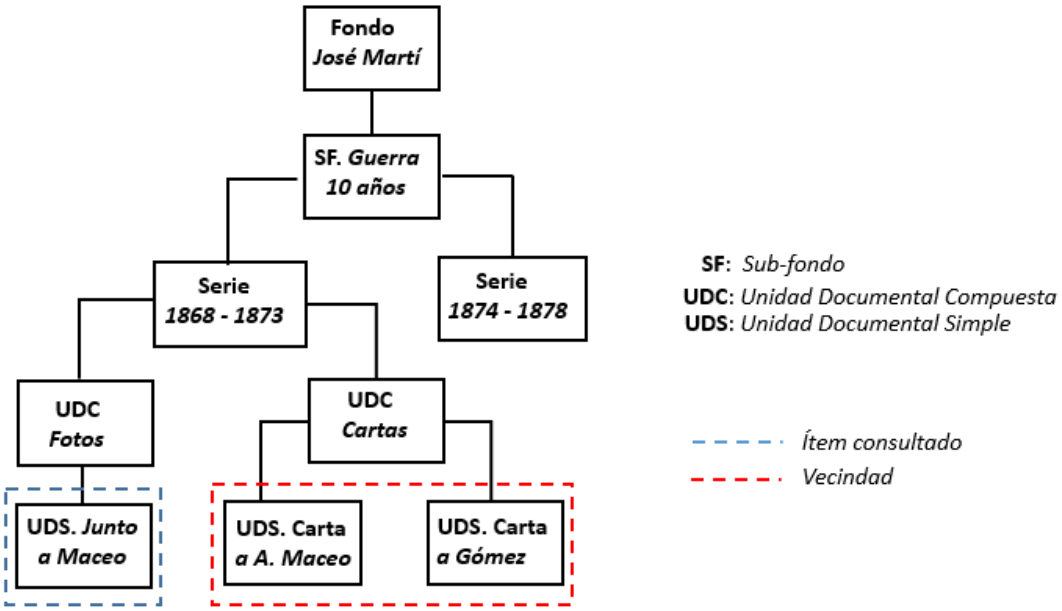

Figura 2.4. Selección de la vecindad inmediata.

En la Figura 2.4 se muestra una representación hipotética de un archivo con la representación de cada uno de los niveles de organización. En esta se ejemplifica para un documento cualquiera (resaltado en azul) como se puede acotar la vecindad (resaltado en rojo) a la que llamaremos vecindad inmediata atendiendo a la descripción multinivel propuesta por ISAD (G). De esta forma para cada documento se

puede determinar los k vecinos más cercanos en la vecindad inmediata sin necesidad de explorar el archivo completo. De manera general podemos resumir la determinación de los k vecinos para cada ítem como:

- Determinar la vecindad inmediata para cada documento.
- Calcular la distancia entre el documento actual y los documentos existentes en la vecindad inmediata.
- Ordenar de menor a mayor las distancias determinadas en el paso anterior.
- Almacenar los vecinos inmediatos ya ordenados.

Es importante tener en cuenta que la vecindad inmediata, en algunos casos, no se obtiene con la facilidad que se muestra en la figura anterior. Como se puede apreciar en la Figura 2.4, algunos documentos de archivos no tienen nodos al mismo nivel con el mismo padre (hermanos) o nodos primos (hijos de un nodo hermano del padre del documento en cuestión), en estos casos se puede aprovechar la descripción multinivel de los documentos de archivos para determinar dicha vecindad. Es posible que en ocasiones se requiera de la exploración de otros niveles (por ejemplo, Serie 1890 – 1985. Ver Figura 2.4). Esta exploración en el árbol se debe realizar ascendiendo por cada uno de los niveles, mientras el nivel correspondiente no sea mayor que el nivel donde se encuentra el sub-fondo que contiene el ítem explorado. Luego para cada nivel se realiza una búsqueda en profundidad hasta alcanzar todas las UDS correspondientes al nivel.

En aras de comprender cada una de las responsabilidades del componente que se describe en este apartado, se describe a continuación la tercera tarea que se adjunta al mismo; representación de las interacciones y perfiles de usuarios. En el capítulo uno, se explicaron algunas de las formas en las que se puede retroalimentar un sistema de recomendación para la actualización de los perfiles de usuarios independientemente del enfoque que se pretenda implementar (basado en el contenido colaborativo, basado en el contexto, en la utilidad o cualquier otro, incluso la hibridación de estos).

Existen dos formas en las que se pueden almacenar la interacción de los usuarios (retroalimentación) con los elementos del sistema. Cuando el sistema requiere una evaluación explicita del usuario, esta técnica es nombrada normalmente como retroalimentación explícita; la otra técnica es nombrada retroalimentación implícita (Rich 1979). La presente investigación se centra en las formas en que se pueden obtener las evaluaciones de manera explícitas, aunque es válido aclarar que un estudio se puede realizar en estos contextos (caracterizados por documentos de archivo) para recolectar las anotaciones de manera implícita (se obtienen a partir del monitoreo y análisis de las actividades de los usuarios).

Las anotaciones de manera explícita indican que tan relevante o interesante resulta un ítem para un usuario (Rich 1979). Existen tres acercamientos principales para obtener retroalimentación por parte del usuario de manera explícita:

- *me gusta/ no me gusta:* los elementos son clasificados como relevantes o no relevantes adoptando una simple escala binaria (Billsus and Pazzani 1999).
- *rating:* una discreta escala numérica es usualmente adoptada para valorar los elementos, alternativamente se pueden asociar simbologías o clasificaciones

con la escala numérica (Shardanand and Maes 1995), un ejemplo de esto es la usada por Syskill & Weber (Pazzani, Muramatsu et al. 1996) donde los usuarios tienen la posibilidad de clasificar las páginas web como calientes, tibias o frías; otro ejemplo es el uso de estrellas.

- *comentarios de textos:* Comentarios referentes a un simple documento, son recolectados y presentados a los usuarios para facilitar el proceso de toma de decisión (Resnick, Iacovou et al. 1994), por ejemplo en amazon.com o ebay.com la retroalimentación de los cliente pueden ayudar a los usuarios para valorar cuales son los productos preferidos por la comunidad.

A partir de la explicación anterior la tercera responsabilidad que se adjunta al módulo es el almacenamiento de cada una de las valoraciones de los usuarios del sistema como se muestra en la Figura 2.3. Es decir que para cada usuario existente o nuevo en el sistema se construye un vector con *K* dimensiones, donde *K* es el número de documentos valorados por este. Esto permitirá la construcción de los perfiles de usuarios para implementaciones de enfoques como SRBC o SRC (Ver Figura 1.2).

En caso de que se pretenda adjuntar un Sistema de recomendación con cualquier otro enfoque, de igual forma este componente es el responsable de almacenar y construir los perfiles de usuarios. Por ejemplo si se piensa en diseñar un SR basado en el conocimiento, es responsabilidad de este componente determinar, inferir, construir, como sea concebido, el perfil de usuario a partir de la información que se pida al usuario activo. De igual forma es responsable de construir y actualizar la base de conocimiento del sistema en cuestión.

De manera general se puede decir que al finalizar la implementación de este componente el sistema puede contar con:

- Cada uno de los ítem representados vectorialmente, donde cada dimensión representa una palabra extraída de la representación de la información en el componente anterior. Esto permitirá implementación de un SRBC o el cálculo de las distancias que separa un documento del resto de los documentos presentes en el sistema.
- Para cada ítem representado, el componente provee la vecindad inmediata de cada uno de estos, aprovechando la concepción multinivel de las descripciones de los fondos documentales propuesta en ISAD (G). Esto hasta cierto punto resuelve uno de los problemas que presentan los Sistemas de recomendación (el arranque en frío). Con la determinación de la vecindad inmediata de cada documento de archivo se pueden recomendar de manera inmediata, sin necesidad de ningún cálculo ni consultas a los perfiles de usuarios, documentos relacionados con un ítem que un usuario activo este consultando en cualquier momento.
- La construcción, almacenamiento y actualización de los perfiles de usuarios permite el diseño de enfoques basados en el contenido, colaborativo o cualquier hibridación entre estos. Además los datos almacenados en él, contribuyen al Analizador de Contexto para la toma de decisiones.

2.2.3 ANALIZADOR DEL CONTEXTO

Hasta este punto, el procedimiento descrito cuenta con una caracterización de la información, lo que permite su representación y análisis de las interacciones de los usuarios. En el apartado anterior se analizaron un conjunto de responsabilidades del componente Representación. Este último provee información necesaria para el funcionamiento del módulo que se describe en el presente acápite, Analizador del Contexto (AC).

El Analizador del Contexto, sin quitarle importancia al resto de los componentes del procedimiento que se propone, se puede ver como el núcleo fundamental del esquema de funcionamiento. Este componente interactúa con el módulo Representación y Motor de Recomendación, decidiendo cuando se deben actualizar las fuentes de recomendación y como se toma la información que será presentada al usuario, haciendo uso del Motor de Recomendación o pasando directamente a la Selección y recomendación de elementos. Todo lo mencionado, más la información que maneja, tanto para mostrar a un usuario o para actualizar el esquema de funcionamiento de algunos módulos, hacen de este componente uno de los de mayor importancia.

Resumiendo sus responsabilidades, se puede plantear que el componente Analizador del Contexto debe ser capaz de:

- Decidir el origen de los datos que serán mostrados al usuario.
- Almacenar las anotaciones de los usuarios.
- Proveer la información necesaria para actualizar el Motor de Recomendación y el módulo de Representación.

Para poder cumplir con la primera responsabilidad: decidir el origen de los datos que serán mostrados al usuario, el analizador debe tener en cuenta el estado del contexto. En la Figura 2.5 se representa el funcionamiento del mismo. Teniendo en cuenta el estado en el cual se encuentra, solicita la vecindad inmediata al módulo Representación especificando el documento (en la figura se representa como *D1*). En este sentido el módulo le da como respuesta el listado que conforma dicha vecindad. Por otra parte requiere la predicción del Motor de Recomendación a partir del perfil de usuario activo. Este último componente (Motor de Recomendación) devuelve el listado con los posibles elementos de interés para el usuario. El Analizador de Contexto combina la predicción y la vecindad entregando todos los elementos al módulo Selección y Representación.

El Analizador tiene en cuenta el estado de la información con que cuenta el sistema del usuario activo. Para un usuario no caracterizado (información insuficiente en el perfil de usuario) que está registrado o no en el sistema, el Motor de Recomendación retorna un listado de elementos vacíos. En este sentido el componente de Selección y Recomendación solo recibirá, entonces, la vecindad inmediata. Puede resultar que para un documento de archivo la vecindad esté conformada por el vacío, si tal es el caso, el Analizador de Contexto solo proveerá la predicción devuelta por el Motor de Recomendación al Componente de Selección y Recomendación. Si en ambos casos los listados de documentos resultan el vacío entonces la información que debe gestionar el Analizador procederá de los documentos de archivos más consultados,

visitados o aquellos que estén próximos al documento según la representación vectorial (Ver sección 2.2.2 Representación de la Información).

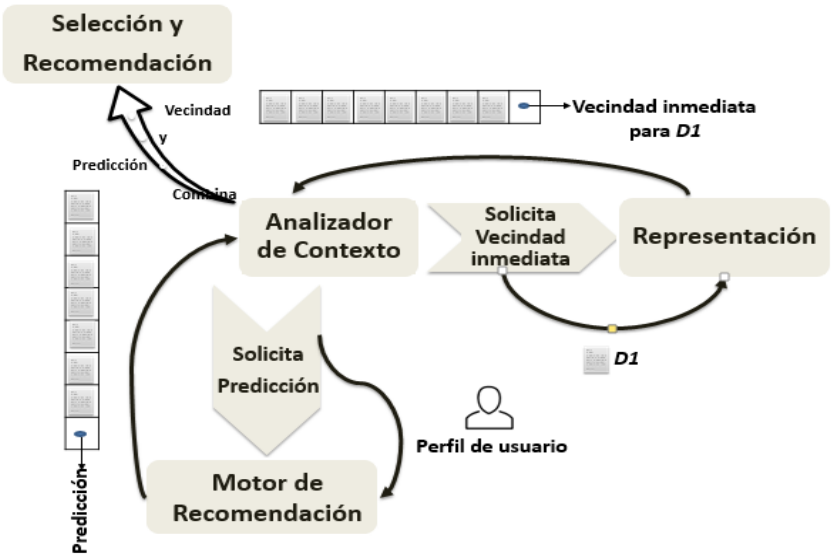

Figura 2.5. Esquema de funcionamiento del AC.

Para cada usuario activo que se pretenda proveer de recomendaciones, el AC selecciona cada uno de los elementos que el usuario considera relevante (en una escala de valoración de 1-5 puntos, valoraciones de 4 y 5 pueden ser consideradas relevantes, mientras que de 1 a 3 se puede considerar no relevante). Para cada uno de los documentos de archivos relevantes para el usuario activo el analizador solicita al módulo de Representación de la Información los vecinos de dicho documento según lo descrito en la sección 2.2.2. Según la descripción multinivel, de los documentos de archivo, plasmada en ISAD (G) estos documentos deben resultar de interés para el usuario por lo que serán combinados con las recomendaciones arrojadas por el MR antes de mostrarse al usuario.

El sistema que se diseñe debe permitir a los usuarios, como se había visto anteriormente la interacción con estos, es decir, debe ser capaz de retroalimentarse para poder inferir los intereses de los usuarios. Es por esto que para los usuarios registrados en el sistema la aplicación debe permitir la valoración de los mismos. En este sentido cada vez que un usuario dé su criterio de los documentos con los cuales interactúe, el componente debe almacenar dichas anotaciones y actualizar el perfil de usuario, cumpliendo de esta forma con la segunda de las responsabilidades asignadas.

Otra de las responsabilidades que no puede verse separada de la anteriormente descrita es el suministro de información necesaria al Motor de Recomendación para el cálculo de las recomendaciones. Es por esto que el Analizador debe registrar los nuevos ingresos de usuarios al sistema y almacenar sus anotaciones. También debe llevar un control para los nuevos documentos de archivos y descripciones que se inserten en la aplicación, pues esta información es igual de relevante que la retroalimentación con los usuarios. Toda esta nueva información el Analizador la suministra al componente de Representación para que pueda actualizar cada uno de los datos que dicho componente gestiona.

Toda la información que maneja el Analizador garantiza la recomendación de elementos a los usuarios bajo cualquier circunstancia. Por ejemplo, si el cálculo de recomendaciones no se puede realizar debido a que la información que se maneja del usuario o de los documentos (según el enfoque se compute en el motor de recomendación) o la vecindad inmediata no es representativa, el módulo con el almacenamiento de los documentos consultados puede inferir los más relevantes por la comunidad y proveerlos como recomendación.

2.2.4 MOTOR DE RECOMENDACIÓN: FILTRADO DE LA INFORMACIÓN RELEVANTE

El componente Motor de Recomendación es el lugar responsable de alojar el enfoque seleccionado. El procedimiento que se propone no fuerza la implementación de un enfoque particular. En el presente apartado se exponen las consideraciones que deben guiar el núcleo del SR, independientemente de la forma en que se determine los elementos relevantes. En el primer capítulo se analizaron diferentes enfoques de SR. Cada uno de los principales acercamientos ya sea, basado en el contenido, colaborativo o la hibridación de estos pueden aplicarse en el procedimiento que caracteriza la presente investigación.

Determinar el tipo de SR que se debe adjudicar al componente debe estar basado en un análisis de la información que exista en el entorno en cada momento. Por ejemplo, pensar en SRBC puede ser una buena elección si se cuenta con información necesaria para describir un usuario a partir de su perfil de usuario y no con suficientes valoraciones para determinar la vecindad si se piensa en un acercamiento colaborativo. La implementación de este se hace posible desde el momento en que se almacenan las representaciones de los documentos de archivos sin importar su tipología documental (ver apartado 2.2.1 Fuente de información). Cuando se tiene caracterizado a un usuario a partir de las valoraciones positivas que ha dado en el pasado entonces aplicando un SRBC (ver Capitulo 1 sección 2.2) y explotando la organización multinivel para las descripciones de documentos de archivos propuesta por ISAD (G) se pueden mitigar las principales limitaciones de los SRBC vistas en el capítulo anterior:

Primero: Al tener conocimiento del entorno y contar con una organización multinivel que garantiza la interrelación entre los documentos con antecesores comunes, es muy probable que los elementos recomendados al usuario sean de su interés. Esto elimina de forma parcial además, el problema del usuario nuevo, puesto que con una sola valoración el procedimiento es capaz de recomendar elementos con una alta probabilidad de aceptación.

Segundo: Haciendo uso de la vecindad inmediata para los elementos valorados en el pasado por el usuario de manera positiva, garantiza la aparición de elementos no esperados. Es decir, el sistema no recomendará solamente elementos con atributos similares a los anotados por el usuario anteriormente.

Por otro lado, una vez que el sistema cuente con un número significativo de usuarios y valoraciones, que permitan recomendar elementos con una alta precisión haciendo uso de un enfoque colaborativo (ver Capitulo 1 sección 1.6) la implementación de este también toma ventaja de la caracterización de los elementos en el componente

Representación. Esto de igual forma que en la aplicación de un SRBC, puede solventar la cobertura limitada y la sensibilidad a la escasez de los datos en los SRC.

En el estudio del arte realizado en la presente investigación se mencionan las diferentes formas en que se pueden hibridar los enfoques ya mencionados (SRBC y SRC) la asunción de cualquieras de las formas que se presentan en el capítulo anterior también pueden tomar ventaja de lo mencionado, por lo que se puede afirmar que computar un Sistema de recomendación híbrido en el Motor de Recomendación es permitido y soportado por el procedimiento propuesto.

Independientemente del enfoque que se desarrolle en este módulo, el Analizador de Contexto le informará al Motor de Recomendación cuando debe recalcular y proveer las recomendaciones para ser presentadas al usuario activo. Este toma toda la información necesaria del componente Representación para computar las solicitudes recibidas por el Analizador de Contexto, el cual una vez ejecutado proveerá los elementos necesarios al componente Selección Recomendación.

2.2.5 SELECCIÓN DE LA INFORMACIÓN Y RECOMENDACIÓN

En las secciones anteriores se ha analizado como se obtienen los elementos de interés para un usuario a partir de las características de los entornos regidos por documentos de archivo y la explotación de estas. Desde el primer componente, que provee toda la información (documentos de archivo) hasta el cálculo de la recomendación en el Motor de Recomendación se han descrito todos los elementos del contexto que favorecen a cada uno de los componentes vistos.

Este componente (Selección y Recomendación) de alguna forma debe ser capaz de seleccionar los elementos acordes con la situación del contexto. Es por esta razón que en todo momento debe conocer las condiciones bajo cual recibe la información del AC. Por ejemplo, no debe seleccionar de igual forma los elementos a mostrar a un usuario poco caracterizado (el sistema no tiene o tiene muy poca información de él) que la información que muestra para un usuario completamente descrito por su perfil de usuario. Desde este punto de vista el proceso de selección se realiza atendiendo a un contexto determinado el cual va estar regido por los distintos escenarios que se analizaron en el componente Analizador:

- Se cuenta con la predicción del Motor de Recomendación y con la vecindad inmediata: En este escenario la mayor parte de los elementos que se le presentaran al usuario se tomarán del listado de documentos que proviene del Motor de Recomendación, añadiendo algunos elementos de la vecindad inmediata.
- Solo recibe la vecindad inmediata.
- Solo recibe la predicción del Motor de Recomendación.
- Recibe los documentos de mayor afinidad por la comunidad de usuarios en el sistema.

Aunque no se pretende vincular los elementos de diseño de un sistema con la propuesta de la investigación, en todos los casos el máximo de elemento a mostrar debe seleccionarse atendiendo a las características de las interfaces de usuarios. De igual forma cada una de las variantes en la que se mezclan los documentos de archivo

que el modulo recibe, dígase cantidad, forma en que se combinan o se presentan al usuario, deben ser ajustadas por el entorno donde se implemente el procedimiento.

Es responsabilidad implícita para el componente Selección y Representación mantener el control de los documentos presentados a usuarios del sistema. La finalidad de esta tarea converge a la actualización de los perfiles de usuarios y sus interacciones con los documentos en el archivo, así como refinar los elementos presentados a un mismo usuario en intervalos de tiempos distantes.

2.3 CONCLUSIONES PARCIALES

La asignación de responsabilidades por componentes, facilita el proceso de comunicación entre ellos, posibilitando además, el soporte de cada una de las tareas del Sistema de recomendación.

La representación de la información, propuesta en el procedimiento descrito donde un documento de archivo es representado de manera uniforme independientemente de su tipología, permite la aplicación de diseños basados en el contenido.

La utilización de la descripción multinivel planteada en la norma ISAD (G) permite identificar documentos de archivo cercanos a otros. Lo que posibilita la recomendación de elementos para usuarios que no están totalmente descritos en el sistema. Esto posibilita además mitigar parcialmente el problema del arranque en frío y elevar la cobertura del enfoque alojado en el MR.

CAPÍTULO 3: VALIDACIÓN DEL PROCEDIMIENTO

Los Sistemas de recomendación pueden ser encontrados en muchas aplicaciones modernas que presentan al usuario una inmensa cantidad de información. Los diseñadores que pretendan implementar SR cuentan con varios algoritmos, como los tratados en el Capítulo 1, teniendo que decidir cuál de ellos es el más apropiado para el objetivo que persiga. Normalmente ese tipo de decisión está basada en los experimentos, comparando el desempeño de cada uno de los candidatos disponibles para la recomendación.

En el presente capítulo se exponen los resultados obtenidos a partir de las pruebas realizadas al procedimiento, para el diseño del Sistema de recomendación en entornos que pretenden difundir acervos históricos, propuesto en la investigación. Se plantean además, los principales conceptos asociados a las métricas de evaluación usadas para la validación de la solución, la concepción de los experimentos realizados y el análisis de los resultados obtenidos.

3.1 MÉTRICAS DE EVALUACIÓN

La mayoría de los diseños de recomendación han sido evaluados y ubicados de acuerdo a que tan potentes son sus predicciones (su habilidad en predecir exactamente las elecciones de los usuarios). Por lo tanto, se deben identificar el conjunto de propiedades que pueden influenciar el éxito de un SR en el contexto de una aplicación específica. Entonces, se podrá evaluar el desempeño del sistema teniendo en cuenta esas propiedades relevantes.

Cuando se pretende evaluar un Sistema de recomendación, se pueden realizar tres tipos diferentes de experimentos; experimento offline, estudio de usuario y experimento online. El primero de ellos es muy sencillo de implementar, usando un conjunto de datos existentes y un protocolo que modele el comportamiento del usuario para estimar el rendimiento del SR, midiendo por ejemplo la exactitud en la predicción. Por otra parte una opción más costosa es el estudio de usuario, donde a un conjunto pequeño de usuarios se les pide la realización de una serie de tareas en el sistema y luego se les pregunta por su experiencia durante el uso del mismo. Finalmente se puede desarrollar un experimento costoso a larga escala sobre un sistema desplegado (experimento online). Tal experimento evalúa el rendimiento de los SR sobre usuarios reales que son inconscientemente conducidos a la evaluación (Shani and Gunawardana 2011).

3.1.1 MÉTRICAS DE EXACTITUD DE LA PREDICCIÓN

La exactitud de la predicción es una de las propiedades de los SR más discutidas en la literatura. En la mayoría de los casos, el pilar fundamental de los SR es su ingeniería de predicción. Esta ingeniería puede predecir las opiniones de los usuarios sobre los ítems (ej. Las valoraciones sobre documentos). Una suposición válida es asumir que los usuarios van a preferir aquellos SR donde la exactitud de la predicción es mayor (Shani and Gunawardana 2011).

En el procedimiento que se propone, como se analizó en el capítulo anterior, se desea predecir el ratings que provee el Motor de Recomendación para los ítems que un

usuario no ha valorado (ej. De 1 a 5 puntos). Por lo que es deseable medir con que exactitud se proveen las valoraciones y como se afectan las recomendaciones entregadas al usuario una vez el Analizador de Contexto sea ejecutado (ver sección 2.2.3).

Root Mean Squared Error (RMSE), es posiblemente la métrica más usada para evaluar la exactitud de las evaluaciones calculadas. Las valoraciones $\hat{r}_{ui}$ son generadas por el sistema para el conjunto de pruebas T de pares usuarios-elementos $(u\,,i)$ donde el verdadero valor r_{ui} es conocido. Normalmente, r_{ui} es conocido porque en un experimento offline, estos valores son ocultados. El RMSE entre el valor calculado y la real valoración r_{ui} está dado por:

$$RMSE = \sqrt{\frac{1}{T}\sum_{(u,i)\in T}(\hat{r}_{ui} - r_{ui})^2}$$

Otra alternativa de esta media es el *Mean Absolute Error* (MAE) dada por:

$$MAE = \sqrt{\frac{1}{T}\sum_{(u,i)\in T}|\hat{r}_{ui} - r_{ui}|}$$

Comparado con MAE, RMSE penaliza desproporcionadamente los grandes errores, por ejemplo, en un sistema donde se determine el error para 4 elementos RMSE va a preferir un sistema con 2 errores de 3 valoraciones y 0 de cuatro a uno que presente 1 de 3 y 0 de los 3 restantes, mientras que MAE prefiere el otro. Es importante considerar que si la distribución de usuarios o ítems no está balanceada en el sistema, es preferible determinar MAE o RMSE por cada uno de estos (usuarios o ítems) y luego determinar la media para el sistema (Herlocker, Konstan et al. 2004).

3.1.2 MÉTRICAS DE USO DE LA PREDICCIÓN

En muchas aplicaciones, además de medir con que exactitud son determinadas las predicciones en los SR, se debe cuantificar el uso que le da el usuario a las recomendaciones que recibe. En las evaluaciones offline del uso de la predicción, normalmente se tiene el conjunto de ítems que un usuario ha usado (esto se puede ver como el conjunto de ítems que el usuario ha valorado). Si se selecciona un usuario de prueba, se pueden ocultar algunas de las valoraciones que ha dado en el pasado y pedir al sistema de recomendación que determine la predicción para estos elementos. Se tendrán entonces cuatro posibles resultados para los elementos ocultados y recomendados, como se muestra en la Tabla 3.1.

Estos resultados pueden verse de la siguiente manera: Un *tp* (abreviatura del inglés true-positive) va ser un elemento recomendado al usuario por el sistema que pertenezca al conjunto de elementos usados por este. De manera análoga un *fp* (abreviatura del inglés false-positive) es un ítem recomendado que el usuario no haya valorado. Para el caso de los elementos que no son presentados al usuario (no recomendados) se tiene que un *fn* (abreviatura del inglés false-negative) es el ítem que no es presentado al usuario y ha sido usado por este y por último, un *tn* (abreviatura del inglés true-negative) es un elemento no recomendado al usuario y no usado por este.

Tabla 3.1. Posibles resultados de recomendación.

	RECOMENDADOS	**NO RECOMENDADOS**
USADOS	Verdadero-Positivo ***(tp)***	Falso-Negativo ***(fn)***
NO USADOS	Falso-Positivo ***(fp)***	Verdadero-Negativo ***(tn)***

Se puede determinar entonces el número de ítems que pertenece a cada una de estas celdas (Tabla 3.1) y computar las siguientes medidas:

$$\boldsymbol{Precision} = \frac{\#tp}{\#tp + \#fp}$$

$$\boldsymbol{Recall}\ (\boldsymbol{True\ Positive\ Rate}) = \frac{\#tp}{\#tp + \#fn}$$

$$\boldsymbol{False\ Positive\ Rate} = \frac{\#fp}{\#fp + \#tn}$$

Comúnmente se puede esperar un intercambio entre estas medidas, mientras permitir grandes listas de recomendación mejora el *Recall*, la Precisión decae. En aplicaciones donde el número de recomendaciones es ordenado antes de presentar al usuario, el uso de la Precisión es la más útil sobre los N ítems presentados. Además, se pueden computar curvas que comparen la Precisión y el *Recall*, o *True Positive Rate* y *False Positive Rate* (Shani and Gunawardana 2011). Cuáles de estas medidas computar, depende en gran medida de las necesidades de cada dominio. Este cálculo en algunos entornos donde para cada usuario se muestran un número fijo de recomendaciones se puede llevar a cabo de manera tal que, para cada lista de recomendaciones de cada usuario se determinan cada una de las medidas y luego se determina la media de cada una de estas.

Es importante tener en cuenta que, para el caso del procedimiento propuesto, donde los elementos son ordenados y presentados al usuario según el valor de predicción, se debe definir un valor a partir del cual un elemento va a ser considerado relevante o no. Otro ejemplo, es transformar la escala de valoración (ej. 1-5) en una escala binaria donde es convertida cada valoración de 4 o 5 a relevante y todas las valoraciones de 1 a 3 en no relevantes (Dahlen, Konstan et al. 1998). Para la solución propuesta, donde las valoraciones están dadas en una escala de 1-5 se definirán como relevantes las predicciones que estén por encima de 3.5 puntos.

3.1.3 COBERTURA

La exactitud de la predicción en los SR, especialmente en los sistemas de filtrado colaborativo, presenta algoritmos que pueden proveer recomendaciones con una gran calidad, pero solo para una pequeña porción de ítems. El término cobertura puede ser visto como la cobertura en el espacio de ítems o en el espacio de usuarios.

Es común observar que el término cobertura se refiere a la porción de elementos que el SR puede recomendar. Esto es llamado también Catálogo de Cobertura. Una simple medida del Cátalo de Cobertura es el por ciento de todos los ítems que pueden ser recomendados. Una medida más significativa es el por ciento de elementos que pueden ser recomendados durante la experimentación. En la presente investigación, donde son calculadas las predicciones, se está de acuerdo con el hecho de que elementos con puntuaciones por debajo de 3.5 puntos no sean recomendados.

La cobertura puede ser determinada además en función de la proporción de usuarios para los cuales el sistema puede recomendar elementos. En algunos casos el sistema puede no recomendar elementos a usuarios donde las valoraciones determinadas estén por debajo de un umbral (como en el ejemplo anterior, por debajo de 3.5 puntos).

Una de las maneras en las que se puede determinar la Cobertura para un SR es seleccionar de manera aleatoria una muestra de pares usuarios/ítems, calculando la predicción para cada par y luego medir el porcentaje de pares para los cuales la predicción fue determinada. De manera análoga, en la cual se analizan la Precisión y *Recall* o *True Positive Rate* y *False Positive Rate*, la cobertura debe ser medida en combinación con las medidas de exactitud en la predicción (ver sección 3.1.1).

3.1.4 OTRAS MÉTRICAS: ALGUNAS CONSIDERACIONES IMPORTANTES

En adición a las métricas analizadas en los apartados anteriores existen otras que analizan características no menos importantes. Algunas de estas métricas, como por ejemplo, el desempeño del sistema ante la aparición de un nuevo ítem o un nuevo usuario (arranque en frío), la aparición de elementos novedosos o casuales beneficiosos en los elementos presentados al usuario, la diversidad existente en las recomendaciones, la adaptabilidad del sistema, entre otras que evalúan diferentes propiedades de los SR. Es responsabilidad del diseñador del sistema decidir cuál de estas medidas deben ser analizadas en dependencia de las características del contexto donde se pretenda implementar el SR.

En la solución que se propone en el capítulo anterior de la presente investigación se puede apreciar cómo se usan conocimientos del dominio para resolver, por ejemplo, el arranque en frío. La propuesta realizada para mitigar esta problemática puede incidir directamente sobre los valores de métricas vistas en las secciones 3.1.1, 3.1.2 y 3.1.3. Además puede alterar hasta cierto punto los elementos novedosos y casuales satisfactorios que aparecen en las recomendaciones.

Arranque en frío: Puede ser considerado un sub problema de Cobertura sobre un conjunto específico de ítems o usuarios. En adición para conocer qué tan grande es el conjunto de elementos o usuarios que corresponden a dicha limitante (Schein, Popescul et al. 2002). Desde el momento en que el componente Analizador de Contexto (ver Capítulo 2) mezcla la lista que provee el Motor de Recomendación con los ítems según el conocimiento del dominio, la ingeniería del procedimiento propuesto se puede enfrentar al arranque en frío. Pero esto es deseable, aun cuando la Precisión y la Exactitud de la predicción disminuyan, en aras de realzar las características que se especifican a continuación.

Novedad en los ítems recomendados: Las recomendaciones novedosas para un ítem son aquellas que aparecen en los ítems presentados al usuario, y este último no tenía conocimiento de la existencia de estos (Zhang, Callan et al. 2002).

Aparición de elementos casuales beneficiosos en la recomendación: Esta propiedad pretende determinar qué tan sorprendente puede resultar un ítem recomendado al usuario. Por ejemplo, si un usuario ha valorado muchas películas de un actor determinado, recomendar una película de dicho actor, sobre la cual el usuario no tenía conocimiento, puede ser novedoso pero no sorprendente (Murakami, Mori et al. 2008).

En el capítulo anterior se detalla como el procedimiento propuesto toma cierta ventaja de la organización multinivel de los documentos de archivo para mitigar limitantes como la del arranque en frío. Estos métodos propuestos (ver sección 2.2.2) pueden realzar la Novedad y Elementos beneficiosos en la recomendación. Por ejemplo un usuario que haya valorado positivamente una carta de José Martí a María Mantilla puede ser provisto con cartas que él desconocía (Novedoso), en este mismo ejemplo se puede ver que si dentro de la unidad documental que contiene la carta valorada por el usuario no existen otros documentos de archivos descritos, la búsqueda de elementos cercanos subirá un nivel en la organización multinivel propuesta por ISAD (G) y elementos como: discursos celebrados por José Martí en la Guerra de Independencia podrían aparecer en la recomendación; este último es un ejemplo de aparición de elementos casuales en la recomendación.

3.2 EXPERIMENTACIÓN

Para la evaluación de la solución propuesta se realizaron experimentos offline. La experimentación offline, brinda la posibilidad de realizar pruebas a diferentes algoritmos a un bajo costo y responde a preguntas como: ¿Qué tan exactas son las predicciones? o ¿Qué uso le da el usuario activo a las recomendaciones brindadas por el sistema? En este sentido se diseñaron las pruebas sobre 5000 documentos de archivos, para 55 905 valoraciones de 500 usuarios en el sistema. Cada uno de los resultados que aparecen en este apartado está determinado para cada uno de los usuarios existentes y cuantificados como la media de los cálculos arrojados.

En el componente de representación se determinaron las vecindades inmediatas de cada uno de los documentos envueltos en el proceso de pruebas. Para 5000 documentos según la representación multinivel de estos se pudo determinar a 321 elementos las vecindades inmediatas. Las anotaciones de los usuarios están dadas en una escala de 1 a 5 puntos, donde uno es la valoración más baja y cinco la más alta. En el Motor de Recomendación se desarrolló un enfoque colaborativo usuario-usuario. La predicción de los elementos fue determinada por regresión y la distancia entre cada uno de los usuarios fue determinada haciendo uso de la correlación de Pearson. A los usuarios se les muestran todos los elementos que se consideren relevantes después de ejecutado el componente de Selección y Recomendación.

La Tabla 3.2 muestra el comportamiento del filtrado colaborativo alojado en el Motor de Recomendación (MR. ver sección 2.2.4). En ella aparecen la medición del Error Absoluto Medio (MAE) calculado para diferentes valores de K (tamaño de la vecindad). Se muestra además la Cobertura del sistema teniendo en cuenta las recomendaciones

dadas por el MR y la Cobertura del sistema incluyendo la técnica propuesta en el Analizador de Contexto (AC. ver sección 2.2.3).

Tabla 3.2 MAE y Cobertura (MR) y (MR + AC).

K	MAE	Cobertura MR	Cobertura MR+AC
5	0,809695	0,5755712	0,6355712
10	0,788169	0,1000752	0,1600752
15	0,739197	0,0267488	0,0867488

En la gráfica se puede observar con mayor claridad como la cobertura para el sistema aplicando el procedimiento propuesto (Cobertura MR + AC) es mayor que teniendo solamente en cuenta las recomendaciones dadas por el motor de recomendación. Esto está dado por la inclusión en la recomendación de las consideraciones respecto a los documentos de archivos; su descripción y ubicación multinivel.

En la Figura 3.1 se puede apreciar además, como decrece en ambos casos, la cantidad de elementos que el sistema puede cubrir a medida que se incrementa el tamaño de la vecindad seleccionada para predecir las valoraciones de los documentos de archivo. Este comportamiento es normal, debido a que seleccionar un mayor número de vecinos para predecir una valoración implica una mayor probabilidad de que no exista tal número de usuarios (en este caso, usuarios, debido a la implementación en el MR de un enfoque colaborativo usuario-usuario) que hayan valorado ese elemento.

Por otra parte la Figura 3.2 muestra el comportamiento del MAE, de igual forma, para diferentes tamaños de vecindad. En este caso el decrecimiento del error en función del aumento de K es justificado. Determinar un mayor número de vecinos implica que más usuarios van a incidir en la predicción de la valoración, por lo que el Error Absoluto Medio con el cual se va a calcular este valor es menor.

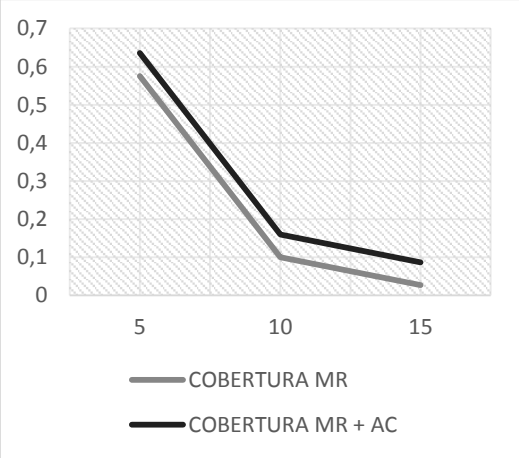

Figura 3.1. Cobertura comparativa (MR + AC) y (MR).

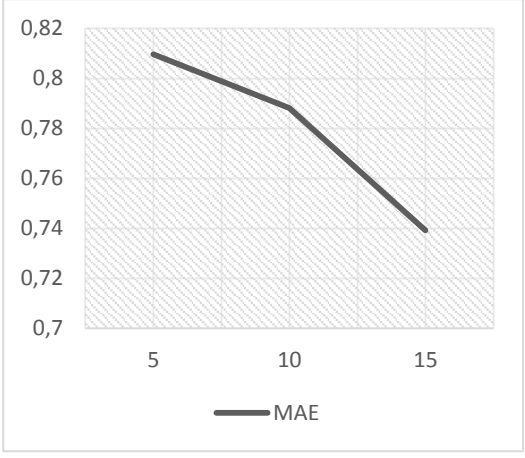

Figura 3.2. Error Absoluto Medio (MAE).

Se debe recordar que estas medidas, MAE y Cobertura, nos dicen qué tan exactas son las predicciones calculadas y para qué por ciento de ítem el sistema puede determinar una predicción. Pero otras propiedades deben ser cuantificadas para luego poder determinar qué configuración es la más adecuada y cómo es el comportamiento del sistema. En este sentido a continuación se muestran en la Tabla 3.3 los resultados obtenidos para medir qué tan usadas son las recomendaciones por los usuarios.

Tabla 3.3. Precisión y Recall de (MR) y (MR + AC).

K	Precisión MR	Recall MR	Precisión MR + AC	Recall MR + AC
5	0,851418	0,822173	0,867575	0,627098
10	0,877822	0,815183	0,884470	0,629401
15	0,144000	0,139000	0,150000	0,128667

En las Figuras 3.3 y 3.4 se puede apreciar mejor las comparativas de estas métricas. Es importante tener en cuenta que la Precisión en la recomendación puede verse como la probabilidad de que un elemento relevante sea recomendado. De cara al usuario, la precisión es muy intuitiva, ya que establecer que un sistema tiene una precisión de un 90% significa que de cada diez elementos nueve son buenas recomendaciones. Por ejemplo, para la tabla anterior para k = 5 la solución propuesta con precisión 0,867575 (87% aproximadamente) implica que cada diez recomendaciones 8,7 son buenas. Por otra parte, el *Recall* es la probabilidad de que elementos relevantes sean recomendados.

Para la realización de este experimento se tomaron todos los usuarios en el sistema, para cada uno de ellos se seleccionaron los elementos valorados de manera positiva (como las valoraciones van a estar dadas de 1 a 5 puntos donde uno es el valor más bajo y 5 el más alto, la conversión a una valoración binaria va a estar dada por: valoraciones de 4 o 5 puntos, se consideran valoraciones positivas; valoraciones de 1 a 3 se consideran valoraciones negativas). Para cada elemento valorado positivamente se seleccionaron los documentos más cercanos a estos (ver sección 2.2.2) y se adicionaron a los elementos valorados por el usuario. Luego se calculó la predicción para cada uno de estos elementos por el MR y se procedió según lo analizado en la sección 3.1.2 del presente capítulo.

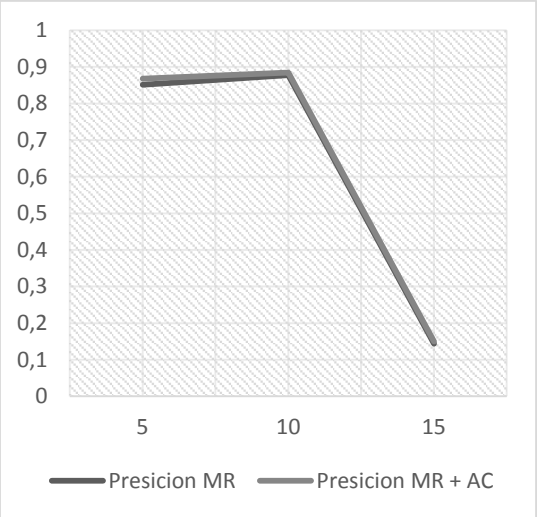

Figura 3.3. Precisión.

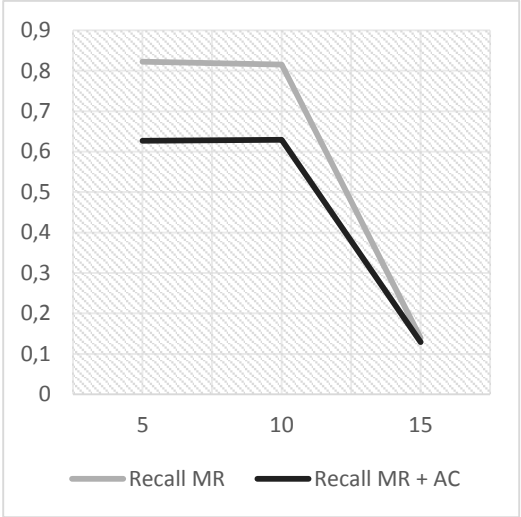

Figura 3.4. Recall.

Como se puede apreciar arriba la Precisión del sistema con la aplicación del procedimiento propuesto es mejor. Este resultado está acorde con la premisa de que "para usuarios que han valorado positivamente un documento D, es muy probable que les resulten relevantes documentos que comparten ancestro con D".

Aunque el *Recall* es mejor para el MR. La disminución del *Recall* con la solución que se propone, está dada por el hecho de que muchos de los elementos que se añaden a las valoraciones del usuario (simulando el uso de estos elementos) no son valoradas positivamente por los usuarios en la vecindad. Como el procedimiento presupone que estos elementos son valorados positivamente por el usuario activo, el número de falsos negativos aumenta y de ahí que el Recall MR + AC en este experimento tenga resultados por debajo del MR.

Es interesante resaltar, que para estos resultados, los mejores valores son arrojados para K = 10, lo que evidencia la cantidad de usuarios con la que se debe contar en la vecindad para determinar las recomendaciones en el MR. Aunque en el caso de la Cobertura y MAE el análisis de esta variable se debe ver desde otro punto de vista. En el caso de la Cobertura, según las características de dominio, es responsabilidad del diseñador decidir si prefiere una amplia cobertura sobre una alta precisión atendiendo al valor de K. De manera análoga debe caracterizar el Error Absoluto Medio.

3.3 CONCLUSIONES PARCIALES

La inclusión de los elementos propuestos por el procedimiento para el diseño del Sistema de recomendación en entornos que pretenden difundir acervos históricos incrementa la Cobertura del sistema.

Se puede apreciar como la precisión del sistema es mayor cuando se mezclan las recomendaciones arrojadas por el MR y los elementos presentados por el AC. Esto tiene mayor impacto, pues los usuarios en la vecindad más cercana, para los datos probados, validan la hipótesis de que elementos en la vecindad inmediata a documentos valorados positivamente resultan de interés para el usuario.

La inclusión de elementos pertenecientes a la vecindad inmediata garantiza la novedad en la recomendación. Por otra parte las pruebas realizadas permiten inferir que para una posible configuración equivalente en un entorno real el número de vecinos para cada usuario debe ser k = 10, valor para el cual los mejores resultados para la Precisión y el *Recall* fueron alcanzados.

CONCLUSIONES GENERALES

El procedimiento propuesto en la presente investigación esclarece como debe diseñarse un Sistema de recomendación, si se pretende adjuntar este a una aplicación que pretende difundir acervos históricos, a partir de las normas que rigen la descripción de los documentos de archivos.

Las pruebas realizadas al procedimiento propuesto demuestran los elevados valores de cobertura y precisión alcanzados, ante diseños como los colaborativos usuarios-usuarios, permitiendo recomendar elementos de interés en entornos regidos por documentos de archivos de manera precisa.

El procedimiento propuesto garantiza la aplicación de enfoques colaborativos, basados en el contenido o la hibridación de estos en el Motor de Recomendación, lo que evidencia la flexibilidad de su diseño.

RECOMENDACIONES

Implementar en el Motor de Recomendación del procedimiento propuesto enfoques basados en contenidos e hibridación de este con un enfoque colaborativo, para analizar el comportamiento de cada uno de ellos.

Definir un valor de utilidad para cada uno de los documentos de archivos según la posición que ocupen respecto a un documento activo en la organización multinivel propuesta por ISAD (G). Esto puede permitir la ponderación de similitud entre documentos en las vecindades inmediatas.

Incluir predictores de tiempo en los perfiles de usuarios y permitir la retroalimentación del sistema de manera implícita a partir del monitoreo de las acciones de los usuarios sobre los documentos de archivos.

REFERENCIAS BIBLIOGRÁFICAS

Ahn, J., P. Brusilovsky, et al. (2007). "Open User Profiles for Adaptive News Systems: Help or Harm?" C.L. Williamson, M.E. Zurko, P.F. Patel-Schneider, P.J. Shenoy (eds.) Proceedings of the 16th International Conference on World Wide Web: 9.

Baeza-Yates, R. and B. Ribeiro-Neto (1999). Modern information retrieval, ACM press New York.

Balabanović, M. and Y. Shoham (1997). "Fab: content-based, collaborative recommendation." Communications of the ACM **40**(3): 66-72.

Basu, C., H. Hirsh, et al. (1998). Recommendation as classification: Using social and content-based information in recommendation. AAAI/IAAI.

Bell, R. M. and Y. Koren (2007). Scalable collaborative filtering with jointly derived neighborhood interpolation weights. Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, IEEE.

Billsus, D., C. A. Brunk, et al. (2002). "Adaptive interfaces for ubiquitous web access." Communications of the ACM **45**(5): 34-38.

Billsus, D. and M. J. Pazzani (1999). "A hybrid user model for news story classification." COURSES AND LECTURES-INTERNATIONAL CENTRE FOR MECHANICAL SCIENCES **99**: 108.

Billsus, D. and M. J. Pazzani (2000). "User modeling for adaptive news access." User modeling and user-adapted interaction **10**(2-3): 147-180.

Bonhard, P., C. Harries, et al. (2006). Accounting for taste: using profile similarity to improve recommender systems. Proceedings of the SIGCHI conference on Human Factors in computing systems, ACM.

Breese, J. S., D. Heckerman, et al. (1998). Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc.

Burke, R. (2000). "Knowledge-based recommender systems." Encyclopedia of library and information systems **69**(Supplement 32): 175-186.

Burke, R. (2002). "Hybrid recommender systems: Survey and experiments." User modeling and user-adapted interaction **12**(4): 331-370.

Burke, R. (2007). Hybrid web recommender systems. The adaptive web, Springer**:** 377-408.

Carlson, C. N. (2003). "Information overload, retrieval strategies and Internet user empowerment."

Claypool, M., A. Gokhale, et al. (1999). Combining content-based and collaborative filters in an online newspaper. Proceedings of ACM SIGIR workshop on recommender systems, Citeseer.

Cooley, R., B. Mobasher, et al. (1997). Web mining: Information and pattern discovery on the world wide web. Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on, IEEE.

Cordón, L. G. P. (2008). Modelos de recomendación con falta de información. Aplicaciones al sector turístico. Departamento de Informática. Jaen, Universidad de Jaen.

Cotter, P. and B. Smyth (2000). Ptv: Intelligent personalised tv guides. AAAI/IAAI.

Cruz Mundet, J. R. (1994). "Manual de archivística." Madrid: Fundación Germán Sánchez Ruiperez/Pirámide.

Dahlen, B. J., J. A. Konstan, et al. (1998). "Jump-starting movielens: User benefits of starting a collaborative filtering system with" dead data." University of Minnesota TR **98**: 017.

Deshpande, M. and G. Karypis (2004). "Item-based top-n recommendation algorithms." ACM Transactions on Information Systems (TOIS) **22**(1): 143-177.

Desrosiers, C. and G. Karypis (2011). A comprehensive survey of neighborhood-based recommendation methods. Recommender systems handbook, Springer**:** 107-144.

Gilliland-Swetland, A. (2005). "Electronic records management." Annual review of information science and technology **39**(1): 219-253.

Goldberg, D., D. Nichols, et al. (1992). "Using collaborative filtering to weave an information tapestry." Communications of the ACM **35**(12): 61-70.

Guttman, R. H. (1998). Merchant differentiation through integrative negotiation in agent-mediated electronic commerce, Citeseer.

Herlocker, J. L., J. A. Konstan, et al. (1999). An algorithmic framework for performing collaborative filtering. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM.

Herlocker, J. L., J. A. Konstan, et al. (2004). "Evaluating collaborative filtering recommender systems." ACM Transactions on Information Systems (TOIS) **22**(1): 5-53.

Herrera, A. H. (1986). Archivística general: teoría y práctica.

Herrera, A. H. (1991). Archivística. Archivística General. Teoría y Práctica D. P. d. Sevilla. España**:** 26.

Holte, R. C. and J. N. Y. Yan (1996). Inferring what a user is not interested in. Advances in Artifical Intelligence, Springer**:** 159-171.

Howe A, F. R. (2008). "Re-considering neighborhood-based collaborative filtering param-eters in the context of new data." Proceeding of the 17th ACM conference on Information and knowledge management: 2.

Jenkinson, S. H. (1980). Selected writings of sir Hilary Jenkinson, Alan Sutton Gloucester, England.

Koren, Y. and R. Bell (2011). Advances in collaborative filtering. Recommender Systems Handbook, Springer**:** 145-186.

Krulwich, B. (1997). "Lifestyle finder: Intelligent user profiling using large-scale demographic data." AI magazine **18**(2): 37.

Linden, G., B. Smith, et al. (2003). "Amazon. com recommendations: Item-to-item collaborative filtering." Internet Computing, IEEE **7**(1): 76-80.

Liu, D.-R., C.-H. Lai, et al. (2009). "A hybrid of sequential rules and collaborative filtering for product recommendation." Information Sciences **179**(20): 3505-3519.

Lops, P., M. de Gemmis, et al. (2011). Content-based recommender systems: State of the art and trends. Recommender Systems Handbook, Springer**:** 73-105.

Mahmood, T. and F. Ricci (2009). Improving recommender systems with adaptive conversational strategies. Proceedings of the 20th ACM conference on Hypertext and hypermedia, ACM.

Mak, H., I. Koprinska, et al. (2003). Intimate: A web-based movie recommender using text categorization. Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on, IEEE.

Melville, P., R. J. Mooney, et al. (2002). Content-boosted collaborative filtering for improved recommendations. AAAI/IAAI.

Melville, P. and V. Sindhwani (2010). Recommender systems. Encyclopedia of machine learning, Springer**:** 829-838.

Miller, B. N., I. Albert, et al. (2003). MovieLens unplugged: experiences with an occasionally connected recommender system. Proceedings of the 8th international conference on Intelligent user interfaces, ACM.

Mobasher, B., H. Dai, et al. (2002). "Discovery and evaluation of aggregate usage profiles for web personalization." Data mining and knowledge discovery **6**(1): 61-82.

Mooney, R. J. and L. Roy (2000). Content-based book recommending using learning for text categorization. Proceedings of the fifth ACM conference on Digital libraries, ACM.

Múgica, M. M. M. (2006). Propuesta de requisitos funcionales para la gestión de documentos archivísticos electrónicos en la administración central del estado cubano.

Departamento de Bibliotecología y Ciencia de la Información Habana, Editorial Universitaria.

Murakami, T., K. Mori, et al. (2008). Metrics for evaluating the serendipity of recommendation lists. New frontiers in artificial intelligence, Springer**:** 40-46.

Niu, L., X.-W. Yan, et al. (2002). Product hierarchy-based customer profiles for electronic commerce recommendation. Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on, IEEE.

Noguera, J. M., M. J. Barranco, et al. (2012). "A mobile 3D-GIS hybrid recommender system for tourism." Information Sciences **215**: 37-52.

Pazzani, M. J., J. Muramatsu, et al. (1996). Syskill & Webert: Identifying interesting web sites. AAAI/IAAI, Vol. 1.

Resnick, P., N. Iacovou, et al. (1994). GroupLens: an open architecture for collaborative filtering of netnews. Proceedings of the 1994 ACM conference on Computer supported cooperative work, ACM.

Resnick, P. and H. R. Varian (1997). "Recommender systems." Communications of the ACM **40**(3): 56-58.

Ricci, F., L. Rokach, et al. (2011). Introduction to recommender systems handbook, Springer.

Rich, E. (1979). "User modeling via stereotypes*." Cognitive science **3**(4): 329-354.

Ruiz, F. F. (2001). "Archivística, Archivo y Documento de Archivo." Revista de Biblioteconomía y Documentación **2**.

Sahlgren, M. (2006). The Word-Space Model: Using Distributional Analysis to Represent Syntag- matic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. Department of Linguistics, Faculty of Humanities. Stockholm, Stockholm University.

Sarwar, B., G. Karypis, et al. (2000). Application of dimensionality reduction in recommender system-a case study, DTIC Document.

Schein, A. I., A. Popescul, et al. (2002). Methods and metrics for cold-start recommendations. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM.

Semeraro, G., Basile, P., de Gemmis, M., Lops, P (2009). Handbook of Research on Digital Libraries: Design, Development and Impact,, Y.L. Theng, S. Foo, D.G.H. Lian, J.C. Na (eds.).

Shani, G. and A. Gunawardana (2011). Evaluating recommendation systems. Recommender systems handbook, Springer**:** 257-297.

Shardanand, U. and P. Maes (1995). Social information filtering: algorithms for automating "word of mouth". Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press/Addison-Wesley Publishing Co.

Sheth, B. and P. Maes (1993). "Evolving Agents for Personalized Information Filtering." Proceedings of the Ninth Conference on Artificial Intelligence for Applications: 7.

Su, X. and T. M. Khoshgoftaar (2009). "A survey of collaborative filtering techniques." Advances in artificial intelligence **2009**: 4.

Zhang, Y., J. Callan, et al. (2002). Novelty and redundancy detection in adaptive filtering. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM.