CLUSTERDOC, UN SISTEMA DE RECUPERACIÓN Y RECOMENDACIÓN DE DOCUMENTOS BASADO EN ALGORITMOS DE AGRUPAMIENTO

(ClusterDoc, a recommendation and retrieval system based document clustering algorithms)

Marylin Giugni O.
Universidad de Carabobo, Venezuela.
mgiugni@uc.edu.ve

Luis León G.
Universidad de Carabobo, Venezuela.
lleon@uc.edu.ve

Joaquín Fernández
Universidad Politécnica de Cataluña, España.
jfernandez@upc.edu

RESUMEN

El fenómeno de internet ha traído consigo una extensa gama de posibilidades de comunicación, y con ello un vertiginoso crecimiento de la información digitalizada. Cada día el usuario es abrumado con la inmensa información que obtiene durante los procesos de búsqueda, donde difícilmente puede identificar aquellos que posean mayor relevancia con respecto a su necesidad informativa; además examinar este ingente volumen de documentos se puede convertir en un problema mayor. En este sentido, las tecnologías de información y comunicación, adquieren un rol importante, no sólo para almacenar la información, sino también para proveer mecanismos adecuados destinados a extraer, de un conjunto de documentos, aquellos que sean pertinentes a una necesidad informativa dada. De ello deriva el objetivo de este trabajo al describir un sistema de recuperación y recomendación automática de documentos denominado ClusterDoc, dirigido a usuarios con necesidades de búsqueda de información, que a través de algoritmos de agrupamiento divide el conjunto de datos en pequeños grupos con características comunes, lo cual permite minimizar el espacio de búsqueda y proporcionar información adaptada a los intereses del usuario. Los resultados iniciales denotan la efectividad de la categorización y personalización del contenido administrado por ClusterDoc.

Palabras claves: Recuperación de información, Categorización de documentos, Algoritmos de agrupamiento, Recomendación de documentos.

ABSTRACT

The Internet phenomenon has resulted in a wide range of communication possibilities and thus a rapid growth of digitized information. Each day the user is overwhelmed with the vast information obtained during the search process, which can hardly identify those that have greater relevance to their information need; in addition, examining the huge volume of documents can become a major problem. In this sense, information and communication

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

technologies, acquire an important role not only store information but also to provide appropriate mechanisms to extract from a set of documents, those that are relevant to a given information need. It derives from the objective of this work to develop a recovery system and automatic recommendation of documents, known ClusterDoc aimed for users with information searching needs, through clustering algorithms divide the dataset into small groups with common characteristics which helps to minimize the search space and provide information tailored to user interests. Initial results denote the effectiveness of categorization and personalization of content managed by ClusterDoc.

Keywords: Information retrieval, Document clustering, Clustering algorithms, Recommendation of documents.

1. Introducción

Internet se ha convertido en un inmenso espacio de información, a menudo muy poco estructurado u organizado, el cual es utilizado por excelencia para la transmisión y generación de información. Es allí donde el usuario diariamente se enfrenta al proceso de encontrar información pertinente, lo cual conlleva una mayor dedicación en la búsqueda, recuperación, selección y síntesis del contenido realmente válido.

En el marco de lo antes expuesto, han surgido los Sistemas de Recuperación de Información (SRI), herramientas que permiten la recuperación de información de grandes colecciones de datos, de naturaleza no estructurada, para satisfacer los requerimientos del usuario (Manning et al., 2009); (López & Soffer, 2008); (Salton, 1989).

Básicamente, los SRI utilizan dos mecanismos, que no son excluyentes entre sí, uno es la búsqueda mediante palabras clave y el otro es la clasificación en categorías. El primero de ellos constituye un primer paso en el proceso de una búsqueda general y emplea diversas técnicas orientadas a superar la barrera semántica (Bloedorn & Mani, 1998).

En lo que se refiere al segundo mecanismo, búsquedas mediante categorización previa, éstas suelen efectuarse manualmente, y han demostrado ser ineficaces pues conlleva una mayor inversión de tiempo, ya que consisten en la elaboración manual de una descripción del contenido temático de cada uno de los documentos (Stubbs et al., 2000).

La motivación para el uso de categorización (clustering) como una forma de mejorar la eficacia de recuperación se basa en la hipótesis de cluster, que establece que los documentos que se agrupan en conjunto tienen una relevancia similar a una determinada consulta (Van Rijsbergen, 1979). La categorización automática tiene en cuenta el contenido de los documentos y se espera que documentos similares contengan términos similares.

Clustering es una técnica de minería de datos, también llamada segmentación, útil para el descubrimiento de las distribuciones de datos y los patrones en ellos. Se trata de

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

un proceso de agrupación de objetos en clases de objetos similares (Wanga & Shao, 2004); (Hsu, 2008).

Es por ello que se ha desarrollado ClusterDoc, un sistema orientado a filtrar los objetos recuperados (documentos), usando los intereses señalados por el usuario y las recomendaciones propuestas sobre dichos objetos; emplea el modelo de espacio vectorial (Baeza & Ribeiro, 1999) y algoritmos de agrupamiento que permiten organizar automáticamente la información, a partir de sus similitudes.

El enfoque de este trabajo se centra en mejorar la fase de búsqueda de información, restringiendo el espacio de exploración e identificando las intenciones del usuario.

Ahora bien, con el propósito de explicar el proceso de desarrollo de ClusterDoc, este artículo se estructura en tres secciones principales. Inicialmente la introducción, una segunda sección donde se abordan los antecedentes de la investigación, la metodología utilizada, se presenta una idea general del funcionamiento del sistema y por último se muestra la evaluación del sistema de recomendación. Finalmente, en la tercera sección se extraen las conclusiones más importantes del trabajo.

2. Trabajos relacionados

Los sistemas de recomendación tratan de ser un paso adelante en el contexto de la recuperación de información tradicional, los mismos se encargan de recomendar o sugerir a los usuarios ítems o productos concretos basándose en sus preferencias (Vélez & Santos, 2006).

Cuando se habla sobre proveer un servicio personalizado en sistemas de e-learning existen dos áreas principales de investigación: sistemas adaptativos y sistemas de recomendación, las cuales constituyen importantes campos de investigación para resolver problemas que involucren grandes volúmenes de datos.

En relación con los Sistemas Hipermedia Adaptativos (SHA), estos son capaces de diseñar información para los usuarios de manera individual, considerando sus objetivos, intereses, forma de estudio, conocimiento y preferencias, entre otros (Brennan & Macnutt, 2006); (Brusilovsky, 1996).

Estos sistemas tienen la capacidad de ajustar su funcionamiento a las metas y otras características de los usuarios (Brusilovsky & Maybury, 2002). Entre algunos SHA en el ámbito educativo se pueden nombrar los siguientes: ELM–ART (Brusilovsky et al., 1996); (Romero et al., 2009), KBS-Hyperbook (Henze & Nejdl, 2001); MLTutor (Essalmi et al., 2010) y MAS–PLANG (Schiaffino et al., 2008).

Los sistemas adaptativos han resultado útiles en diversas áreas de aplicación, donde el espacio de búsqueda de la información es razonablemente grande, y donde el sistema puede ser utilizado por personas con diferentes objetivos; (Durán et al., 2007). En este sentido, ClusterDoc puede ser considerado un sistema de recomendación que se basa en

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

la teoría de los sistemas adaptativos, para proveer planes de lectura personalizados, de acuerdo a los intereses del usuario.

Para seleccionar los contenidos se han empleado diversos enfoques, entre los que destaca el Espacio Vectorial (EV) (Salton & McGill, 1983), que tiene sus raíces en el área de Recuperación de Información (en inglés Information Retrieval) mediante el cual se representa el carácter temático de los objetos, como un vector de términos.

Otro enfoque hace uso de diferentes técnicas del área del Aprendizaje Automatizado (Machine Learning) como son las Redes Bayesianas (Fan et al., 2001), la Clasificación (Classification) y la Agrupación (Clustering) (Sun et al., 2001).

Algunas de estas técnicas utilizan TF-IDF (Salton & Buckley, 1988); (Salton, 1991) (del inglés Term Frequency Inverse Document Frequency), para determinar cuán relevante es un término dentro de un documento.

Mediante TF-IDF se mejora la exactitud con la que se clasifican los documentos al considerar la existencia de palabras claves dentro de ellos (Zhang et al., 2005); es importante recordar que dentro del campo de la investigación cualquier texto de carácter científico (papers, proceedings, entre otros) posee determinadas palabras claves que son las más relevantes dentro su contexto.

En lo que respecta a selección de contenidos, existen diversos algoritmos de clasificación dependiendo de las representaciones elegidas para el modelo de usuario y los documentos: fórmula del coseno, reglas asociadas a estereotipos, redes neuronales, vecino más cercano, clasificador bayesiano ingenuo, distancia euclidiana, entre otros.

ClusterDoc emplea la distancia euclidiana como medida de semejanza entre los documentos que conforman los cluster, la cual es fundamental para determinar el grado de pertinencia que tiene el documento en relación con el área de investigación (cluster) donde está ubicado.

3. Metodología

La investigación está guiada por la estrategia de investigación empírica denominada “Investigación Acción” (Baskerville, 1999). De igual manera utiliza una combinación de metodologías cuantitativas y cualitativas que se complementan y permiten compensar los sesgos metodológicos del estudio (Cataldi, 2005). A continuación se detallan las cinco fases de la investigación:

1. Fase de diagnóstico: está relacionada con la identificación y descripción de la situación actual.

2. Fase de planificación de la acción: define las acciones que deben ser ejecutadas para mejorar el problema.

3. Fase de implementación de la acción: se realiza la acción planificada.

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

4. Fase de evaluación: una vez culminadas las acciones, se evalúan las salidas obtenidas.

5. Fase de especificación del aprendizaje: corresponde a la reflexión sobre los resultados de la fase de evaluación, lo cual podría dar inicio a una nueva iteración.

Por otro lado, ya que la investigación comprende la fase de desarrollo de software en el área de minería de datos, se utiliza el modelo de referencia CRISP-DM (Hernández et al., 2004); (Venter et al., 2007) acrónimo de CRoss-Industry Standard Process for Data Mining, que consta de las fases: comprensión del negocio, comprensión de datos, preparación de los datos, modelado, evaluación y explotación.

4. Resultados

4.1 Fase de diagnóstico

El desarrollo de ClusterDoc comprendió diversas etapas, enfocadas en desarrollar un sistema capaz de proveerle al usuario un espacio para localizar y acceder a información adaptada a sus intereses.

La fase inicial de la investigación se enfocó en la comprensión de la problemática. Durante esta etapa se realizaron entrevistas y observaciones directas sobre los diversos procesos de búsqueda de información, empleados por los estudiantes que conforman la muestra.

Además, se aplicó una encuesta al 84% (21 estudiantes) de la población estudiantil del último año de la Licenciatura en Computación, de la Facultad de Ciencias y Tecnología (FaCyT) de la Universidad de Carabobo (UC), año lectivo 2009.

En dicha encuesta se observó que el 86,66% de los encuestados invierten de 18 a 22 horas semanales en la búsqueda de información digital, de las cuales el 20% (4 horas/semana), son utilizadas para su clasificación.

Por otra parte, el 92% de la muestra señaló que la información recuperada satisface entre un 25% y 35% sus requerimientos. Adicionalmente, el 100% de los encuestados expresó que debían realizar las búsquedas con diversos criterios, para obtener resultados acordes a sus intereses. Es por ello que resulta de especial interés el diseño de mecanismos que permitan automatizar y facilitar el proceso de búsqueda y recuperación de información en un entorno de investigación.

Además, proveer una herramienta a los usuarios, la cual a partir de una representación lógica de los documentos y la agrupación automática de los mismos, proporcione documentos relevantes a sus intereses.

4.2 Fase de planificación e implementación de la acción

Como se ha venido mencionando, la proliferación de fuentes de información, tanto en el ámbito científico, profesional e incluso doméstico, reflejan la importancia de

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

proporcionar aplicaciones que faciliten el acceso a la información, minimicen los tiempos de respuesta en su recuperación, mejoren la calidad de la información obtenida y contribuyan a su reutilización. Considerando estas necesidades se ha diseñado y desarrollado ClusterDoc, el cual se detalla en las próximas secciones del artículo.

5. Descripción del sistema

ClusterDoc se puede describir a través de sus tareas básicas: representación lógica de los documentos, representación de las necesidades de los usuarios, agrupación de los documentos, generación de los planes de lectura y retroalimentación del usuario.

a) Representación lógica de los documentos: el sistema provee un espacio de búsqueda y recuperación de información adaptada a las necesidades de los investigadores, pero estos documentos inicialmente deben ser depurados mediante una fase de limpieza, la cual consiste en la eliminación de acentos, imágenes y otros caracteres especiales.

Posteriormente, los documentos se representan numéricamente empleando el modelo de espacio vectorial (Salton, 1991), el cual permite representar cada documento mediante un vector de pesos.

Para el cálculo de los pesos ClusterDoc utiliza el esquema TF-IDF (Salton & Buckley, 1988); (Salton, 1991); (Roelleke & Wang, 2008), el cual determina cuán relevante es un término dentro de un documento.

Un documento d_j es representado por un conjunto de palabras (t_1, t_2,...,t_n), donde cada t_i corresponde a una palabra clave que representa al documento y n es el tamaño del conjunto; de esta manera cada palabra t_i tiene asociado un peso w_i.

De esta forma, el peso de la palabra t_i en el documento d_j, viene dada por la Ecuación 1 (Salton & Buckley, 1988); (Berger et al., 2000):

$$w_{ij} = \mathrm{f}_{ij} \cdot \log_2(n/\mathrm{df}_i) \quad (1)$$

En la Ecuación 1, el término $f_{i,j}$ representa el número de apariciones del término i en el documento j, el total de documentos viene dado por n; df_i se refiere al número de documentos en los que aparece el término i, y w_{ij} representa el peso del término i en el documento j. Por lo tanto, una vez que se ha obtenido el peso de cada palabra en el documento, éste se representa por su vector característico, el cual se observa en la Ecuación 2:

$$d_j = (w_1, w_2, w_3, \ldots, w_n) \quad (2)$$

Donde w_i corresponde al peso de la palabra i en el documento j. A partir de este momento se hará referencia al documento d_j como el vector característico, representado por el peso de sus palabras claves.

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

b) Representación de las necesidades de los investigadores: está relacionada con el perfil de usuario, el cual describe los intereses y metas del investigador. Este perfil inicialmente considera información de un dominio específico y se obtiene valorando el interés del usuario en una escala de uno a cinco, lo cual representa su interés en un dominio de investigación determinado.

El perfil inicial se utiliza para generar un plan de lectura acorde a las necesidades del usuario, de esta forma el sistema podrá recuperar y mostrar los documentos que conforman dicha recomendación. El perfil del usuario se actualizará a medida que los otros usuarios califiquen cada uno de los documentos suministrados por el sistema; señalando a través de elementos tipo Likert de nivel 5, el grado de contribución a su área de investigación.

c) Agrupación de los documentos: una vez que los documentos se han caracterizado, ClusterDoc los agrupa utilizando el algoritmo de agrupamiento k-means, propuesto por (McQueen, 2007), el cual es ampliamente utilizado por su facilidad, rapidez y efectividad (Garre et al., 2007); (Ayaquica, 2007); (Rizman, 2008).

ClusterDoc sigue una forma fácil y simple para dividir los documentos en k grupos de documentos similares de una misma área de investigación, donde k representa el número de áreas de investigación. La idea principal es definir k centroides (uno para cada grupo), estos k centroides representan los documentos más relevantes de cada área de investigación.

Ahora bien, una vez que se ha establecido el número de grupos, se eligen k documentos que representarán los centroides (centros) iniciales de cada cluster. En este paso es importante seleccionar documentos cuyo contenido tenga una alta correspondencia con el área de investigación, por lo tanto si los documentos no son los apropiados, la clasificación podría ser ineficiente.

Cada vez que un nuevo documento se agrega al sistema, es situado en el grupo más cercano, para ello se calcula la distancia entre su vector característico y cada uno de los k centros, utilizando la distancia euclidiana (Ecuación 3).

$$dis = \sqrt{\sum_{i=1}^{n} \left(c_{ki} - D_{ji}\right)^2} \quad (3)$$

En esta ecuación c_k corresponde a la componente i del vector característico del centro del grupo (cluster) k. D_{ji} es la componente i del vector característico del documento j, siendo ambos vectores de tamaño n, y dis representa la distancia euclidiana (Alsabti et al., 1998). Así pues cuando un documento es agregado a uno de estos grupos, se recalcula el centroide de todos los cluster y la distancia de todos los demás documentos respecto a estos, para verificar si alguno de los antiguos documentos debe ser reasignado.

d) Generación de los planes de lectura: una vez que se han conformado los grupos de documentos, el sistema genera el plan de lectura pl_j representado por un conjunto de

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

documentos (d_1, d_2,..., d_w), adaptado a los intereses del usuario j, donde w representa el número de documentos que conformarán el plan de lectura.

Para determinar el plan de lectura, se calcula la distancia euclidiana entre el vector característico del usuario j y los documentos del cluster k, es decir, los documentos correspondientes al área de investigación k, identificada en el perfil del usuario. El plan lo formarán los documentos que tengan mayor similitud con el vector que representa los intereses del usuario.

e) Retroalimentación del usuario: esta etapa corresponde a la retroalimentación explicita que realiza el usuario j a cada uno de los documentos que conforman su plan de lectura. Los valores de retroalimentación del sistema están entre cero y seis inclusive.

A medida que el usuario evalúa los documentos se modifica el perfil del usuario, generando un plan de lectura acorde a sus preferencias. El perfil se actualiza mediante el uso de la fórmula de Rocchio (Ecuación 4), (Rocchio, 1971); (Alonso et al., 2000), la cual se basa en la relevancia de los documentos evaluados por el usuario. Para ClusterDoc un documento es relevante si su valoración es mayor o igual a 4 puntos (en una escala del cero al seis).

$$C_1 = \alpha C_0 + \beta \sum_{i=1}^{n_r} \frac{R_i}{n_r} - \gamma \sum_{i=1}^{n_{nr}} \frac{NR_i}{n_{nr}} \quad (4)$$

En la Ecuación 4, C_0 representa el vector característico del perfil inicial del usuario. C_1 representa el vector característico del perfil del usuario luego de una iteración. R_i es el vector característico del documento relevante i, NR_i es el vector característico del documento no relevante i, n_r representa el número de documentos relevantes, n_{nr} el número de documentos no relevantes.

Las constantes α, β y γ permiten ajustar el impacto de los documentos relevantes y los no relevantes, dichos valores fueron ajustados siguiendo la sugerencia de Manning (Manning et al., 2009), por lo que α tiene el valor 1, β de 0,75 y γ de 0,25.

6. Evaluación del sistema

ClusterDoc puede ser utilizado en diversos dominios de investigación, su interfaz de configuración permite extender sus funcionalidades a poblaciones de estudio con diversos intereses. Aunque su limitación se encuentra en el formato de los documentos que utiliza en la categorización, uno de los principales beneficios se observa en la construcción de un espacio para la colaboración.

Ahora bien, ClusterDoc ha sido evaluado en un grupo de usuarios con necesidades de búsqueda de información, conformado por 12 estudiantes del último año de la Licenciatura en Computación de la UC, todos ellos en la fase de elaboración de Trabajo Especial de Grado. Como caso de estudio se realizó una prueba donde se alimentó al sistema con 100 documentos de tipo Portable Document Format (PDF), enmarcados dentro de cuatro áreas de conocimiento de las ciencias de la computación, en particular:

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

Interacción Humano Computador, Ingeniería del Software, Sistemas Operativos y Lenguajes de Programación (IEEE, 2008). La selección de este subconjunto de áreas obedece a los intereses de la muestra intencional, la cual se encontraba realizando investigaciones en los ámbitos mencionados.

En relación a la representación lógica de los documentos, estos se transformaron a un formato de texto plano y seguidamente fueron caracterizados según el modelo de espacio vectorial descrito previamente, calculando los valores TF-IDF para cada palabra clave dentro del documento. La Tabla 1 muestra los vectores característicos de 10 documentos del sistema

Tabla 1. Vector característico, representando el peso de las primeras 4 palabras claves de diez documentos

Documento	pesos			
	w_1	w_2	w_3	w_4
d_{73}	0,2290	0,9220	0,6657	0,0308
d_{21}	0,0131	0,9023	0,7311	0,0815
d_{65}	0,4738	0,9146	0,7402	0,0175
d_1	0,2242	0,3617	0,8079	0,1486
d_{29}	0,2300	0,7658	0,5539	0,0117
d_{37}	0,7860	0,6449	0,5366	0,0592
d_{93}	0,8701	0,2283	0,7656	0,1526
d_5	0,9252	0,2561	0,8991	0,0621
d_{25}	0,5352	0,0124	0,7357	0,0402
d_{69}	0,5604	0,0256	0,4780	0,1481

En la Tabla 1 se observan los pesos correspondientes a las cuatro primeras palabras claves de los vectores característicos de cada documento. Cabe destacar que cada área de conocimiento, en el experimento realizado, contiene cuatro descriptores o palabras claves que la identifican, para un total de 16 descriptores en general. Los documentos que se observan en la Tabla 1 están organizados ascendentemente, en función a la distancia entre cada documento y el perfil del usuario.

Después de configurado el sistema, indicando el número de grupos a conformar (k cluster), se ejecutó el algoritmo k-means, de acuerdo a lo descrito previamente (apartado 5c). De esta forma se generaron 4 grupos de documentos organizados por la similitud entre ellos. La Tabla 2, muestra el vector característico asociado al centro del tercer cluster.

Por otro lado, ClusterDoc genera el vector característico de cada uno de los usuarios que interactúan con el sistema, para ello utiliza un formulario de entrada que identifica el grado de interés del estudiante en un área determinada. Seguidamente, el sistema presenta al usuario los documentos ordenados por el nivel de relevancia, mostrándole los documentos cuyas evaluaciones sean más altas y que mejor se adapten a su perfil de búsqueda, esto último se logra determinando la cercanía del perfil del usuario a cada documento.

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

Tabla 2. Centro del cluster 3

cluster	pesos			
	w_1	w_2	w_3	w_4
c_3	0,4200	0,5069	0,4135	0,4296

Es importante destacar que los planes de lectura se van actualizando a medida que los documentos son evaluados por los otros usuarios del sistema. Evaluaciones positivas o negativas, afectan la relevancia de los documentos en un área de investigación determinada. En la Tabla 3 se observa el plan de lectura inicial recomendado al usuario i del experimento, los documentos se muestran ordenados ascendentemente de acuerdo a su similitud con el usuario (menor distancia euclidiana).

Tabla 3. Plan de lectura inicial recomendado al usuario i.

Documento	Distancia
d_{65}	0,7194
d_{73}	0,7422
d_{21}	0,7800
d_1	0,8098
d_{93}	0,8295
d_5	0,8500
d_{37}	0,8614
d_{29}	0,8693
d_{25}	0,9040
d_{69}	0,9301

En un principio, al usuario se le recomiendan los documentos más cercanos a sus intereses, ordenados por distancia, ya que inicialmente todos los documentos presentan la misma valoración. Aquellos documentos cuya distancia se aproxime a cero, indican una mayor similitud con el perfil del usuario, por el contrario valores distantes de cero tienen menor relación con las necesidades del usuario. A medida que el usuario interactúa con el sistema, los documentos varían su posición en los planes de lectura, según la retroalimentación del usuario (apartado 5e).

Efectividad del proceso de personalización

Para determinar la efectividad el proceso de personalización se han utilizado dos tipos de evaluación, una evaluación cualitativa que viene dada por la satisfacción del usuario y otra evaluación de tipo cuantitativa, en función de la precisión del plan de lectura.

Con respecto a la satisfacción de los usuarios en relación al uso del sistema y en particular de la recomendación de los planes de lectura, se aplicó un cuestionario a los 12 usuarios que conforman la muestra, obteniendo los datos que se observan en la Tabla 4.

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

Tabla 4. Resultados del cuestionario aplicado a la muestra (Datos expresados en porcentaje)

Aspectos evaluados	(1)	(2)	(3)	(4)	(5)
Contenidos ajustados al perfil	0,00	0,00	8,33	25,00	66,67
Utilidad de los contenidos	0,00	0,00	8,33	33,33	58,33
Las categorías generadas resultan adecuadas a las necesidades de información	0,00	0,00	25,00	25,00	50,00
Contribuye a resolver las necesidades de información	0,00	0,00	8,33	16,67	75,00
Planes de lectura acorde a las necesidades	0,00	0,00	8,33	25,00	66,67
Facilidad de uso	0,00	0,00	0,00	16,67	83,33
Relevancia de los documentos recomendados por el plan de lectura	0,00	0,00	8,33	33,33	58,33
Planes evolucionan de acuerdo al perfil del usuario	0,00	0,00	8,33	33,33	58,33
El contenido de los documentos recuperados corresponden al dominio de estudio	0,00	0,00	0,00	25,00	75,00
Los documentos recuperados tienen alta pertinencia con la categoría en la que se encuentran	0,00	0,00	8,33	25,00	66,67

Leyenda: (1) totalmente en desacuerdo, (2) en desacuerdo, (3) indiferente, (4) de acuerdo, (5) totalmente de acuerdo.

En relación a los resultados obtenidos, se observa que al sumar los datos de las columnas “totalmente de acuerdo” y “de acuerdo”, de los ítems 1, 5, 8 y 9 el 93,75% de la muestra considera que ClusterDoc genera planes de lectura adaptados a los intereses del usuario. Aunque en promedio el 8,33% de los encuestados consideró indiferente los resultados generados por el sistema, el otro 91,66% estuvo “de acuerdo” y “totalmente de acuerdo” con los documentos recomendados por el sistema. Asimismo, el 100% de los usuarios señaló que el sistema es fácil de utilizar.

Por otro lado, considerando la importancia que tiene el usuario en este tipo de sistemas, se realizaron entrevistas a los mismos, de esta forma se observó la contribución de los documentos de acuerdo a los intereses del usuario.

Las entrevistas fueron fundamentales para identificar fortalezas y debilidades del sistema, ya que en ellas se evidenció la importancia de la taxonomía de las palabras claves. Además, se observaron los criterios de cada usuario para valorar como relevante o no relevante un documento.

Para evaluar de manera cuantitativa la calidad del agrupamiento de los documentos dentro del cluster, se utilizó la similitud promedio (Yolis et al., 2003) (Steinbach et al., 2000) (Ver ecuación 5), la cual es una medida que permite determinar el grado de semejanza entre los documentos dentro de cada conjunto (cluster).

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

$$SP_j = \frac{1}{n_j^2}\left(\sum_{i=1}^{n_j}\sum_{j=1}^{n_j} sim(d_i, d_j)\right) \quad (5)$$

Como se puede observar en la ecuación de similitud promedio, inicialmente es necesario calcular la similitud entre cada par de documentos dentro de un mismo cluster, para ello se utiliza el coeficiente coseno extendido (Ver Ecuación 6) (Steinbach et al., 2000).

Esta función mide el ángulo que forman dos vectores, donde cada vector representa un documento; un resultado muy cercano a uno indica que el ángulo formado entre los vectores es muy pequeño, es decir, los documentos están muy próximos en el espacio y son muy semejantes; un resultado cercano a cero, indica que el ángulo formado por los vectores es amplio, por lo que los documentos no se relacionan entre sí.

$$sim(d_i, d_j) = \cos(\phi) = \frac{\vec{d}_i \bullet \vec{d}_j}{\|d_i\| \cdot \|d_j\|} = \frac{\sum_{k=1}^{n} w_{ki} \cdot w_{kj}}{\sqrt{\sum_{k=1}^{t} w_{ki}^2} \cdot \sqrt{\sum_{k=1}^{t} w_{kj}^2}} \quad (6)$$

De esta forma, cuando la similitud promedio retorna valores cercanos a uno, significa que los documentos guardan relación entre sí, y el cluster es muy homogéneo, por el contrario un valor cercano a 0 indica que hay una gran diferencia entre los documentos que forman el cluster.

Una vez realizada la evaluación de los 4 cluster, se observa en la Tabla 5 que la similitud promedio de cada uno de ellos es muy cercana a uno (superior a 0,70), lo cual indica que los documentos en cada grupo son semejantes.

Tabla 5. Similitud promedio de los 4 cluster

SP_1	SP_2	SP_3	SP_4
0,7317	0,7946	0,8171	0,7069

7. Conclusiones

En este artículo se ha descrito ClusterDoc como un sistema de apoyo a los investigadores, el cual a través de algoritmos de categorización contribuye en el proceso de generación de planes de lectura, adaptados a los intereses de los usuarios.

La investigación y los resultados obtenidos confirman que los algoritmos de agrupamiento son una poderosa herramienta para la resolución de problemas en los cuales el espacio de soluciones es amplio. La generación de cluster ha creado una configuración de los documentos acorde a los resultados esperados, obteniéndose "vectores patrón" representativos de cada categoría o clase.

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

Con el desarrollo de ClusterDoc se ha obtenido una primera aproximación para ayudar a los usuarios de una comunidad científica. Se ha generado un sistema de recuperación que primero pre procesa los documentos y los agrupa por similitud, facilitando las búsquedas por tópicos y proporcionando al usuario planes de lectura, conformados por los documentos que se asemejan a su perfil, es decir, agrupados por su relevancia con el tema y sus necesidades de información.

El estudio experimental ha comprobado que los planes de lectura generados por el sistema son acordes a los intereses del usuario y han sido calificados de manera positiva por el mismo. Aunque las evaluaciones cuantitativas realizadas al sistema, son satisfactorias, es necesario realizar una evaluación más exhaustiva en otros dominios de investigación, que permita validar estos resultados.

Finalmente, cabe señalar que el desarrollo de sistemas de recuperación y de categorización de documentos, considera dos importantes líneas de trabajo futuras, e investigaciones en estas áreas pueden contribuir en el diseño de herramientas de ayuda a los usuarios que faciliten la búsqueda de información en diversos dominios de investigación.

8. Referencias Bibliográficas

Alonso, J., Figuerola, C. & Zazo A. (2000). Categorización automática de documentos en español: algunos resultados experimentales. I Jornadas de Bibliotecas Digitales, JBIDI 2000, Valladolid, España, 149-160.

Alsabti, K., Ranka, S. & Singh, V. (1998). An Efficient K-Means Clustering Algorithm. Proceedings of IPPS/SPDP Workshop on High Performance Data Mining.

Ayaquica, I., Martínez., J., Carrasco, J. (2007). Restricted Conceptual Clustering Algorithms based on Seeds, Computación y Sistemas, Vol.11 (2), México.

Baeza-Yates. R. & Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison-Wesley. New York.

Baskerville, R. (1999). Investigating Information Systems with Action Research. Communications of the Association for Information Systems, Vol. 2, Art. 19.

Berger, A., Caruana R., Cohn D., Freitag D. & Mittal, V. (2000). Bridging the Lexical Chasm: Statistical Approaches to Answer Finding. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 192-199.

Bloedorn, E. & Mani, I. (1998). Using NLP for machine learning of user profiles, Intelligent Data Analysis. Vol. 2(3), 3-18.

Brennan, M. & Macnutt, L. (2006). Learning styles and learning to program: an experiment in adapting online resources to match a student´s learning style. International

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

Conference on Innovation, Good Practice and Research (EE 2006). University of Liverpool. England, 177-182.

Brusilovsky, P. & Maybury, M. (2002). From adaptive hypermedia to the adaptive web. Communications of the ACM. Vol. 45(5), 30-33.

Brusilovsky, P. (1996). Adaptive Hypermedia: An attempt to analyze and generalize. Multimedia, Hypermedia, and Virtual Reality Models, Systems, and Applications. Lecture Notes in Computer Science. Vol. 1077, 288-304.

Brusilovsky, P., Schwarz, E. & Weber, G. (1996). ELM–ART: An intelligent tutoring system on World Wide Web. Intelligent Tutoring Systems. Lecture Notes in Computer Science. Vol. 1086, 261–269.

Cataldi, Z. (2005). El aporte de la tecnología informática al aprendizaje basado en problemas usando modelos de trabajo interactivos. Universidad de Sevilla. Tesis doctoral no publicada.

Durán, E., Costaguta, R., Maldonado, M. & Unzaga S. (2007). Sistemas Adaptativos Inteligentes. IX Workshop de Investigadores en Ciencias de la Computación. Argentina. Mayo, 2007. pp. 143-146.

Essalmi, F., Jemni, L., Jemni, M., Kinshuk, Graf, S. (2010). A fully personalization strategy of e-learning scenarios. Computers in Human Behavior, Vol. 26, 581–591.

Fan, Y., Zheng, C., Wang, Q.Y., Cai, Q.S., & Liu, J. (2001). Using naive bayes to coordinate the classification of web pages. Journal of Software. Vol. 12. No, 9, 1386-1392.

Garre, M., Cuadrado, J., Sicilia, M., Rodríguez, D., Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software, REICIS, Revista Española de Innovación, Calidad e Ingeniería del Software, Vol. 3 (1), 1885-4486.

Henze, N., & Nejdl, W. (2001). Adaptation in open corpus hypermedia. International Journal of Artificial Intelligence in Education. Vol 12, 325–350.

Hernández J., Ramírez M. & Ferri C. (2004). Introducción a la minería de datos. Editorial Pearson.

Hsu, M. (2008). A personalized English learning recommender system for ESL students. Expert Systems with Applications, 34, 683–688.

IEEE (2008). CS2008 Review Taskforce. Computer Science Curriculum 2008: An Interim Revision of CS 2001. Association for Computing Machinery and IEEE Computer Society. Documento en línea. Disponible en: http://www.acm.org//education/curricula/ComputerScience2008.pdf. Consulta: diciembre de 2008.

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

López, M. & Soffer, M. (2008). Automatic Text Processing for Spanish Texts. CERMA, Electronics, Robotics and Automotive Mechanics Conference. 74-79.

Manning, C., Raghavan, P. & Schütze, H. (2009). An Introduction to Information Retrieval. Cambridge University Press, 181-183.

McQueen, J. (2007). Some methods for classification and analysis of multivariate observations, 5-th Berkeley Symposium on mathematics, Statistics and Probability, 1, 281-297.

Rizman, K. (2008). An efficient k-means clustering algorithm. Pattern Recognition Letters, Vol. 29, 1385–1391.

Rocchio, J. (1971). Relevance Feedback in Information Retrieval. In Salton, G. (Ed.), The SMART Retrieval System: Experiments in Automatic Document Processing, 313-323. Prentice Hall.

Roelleke T., Wang J. (2008). TF-IDF uncovered: a study of theories and probabilities. Retrieval Models.

Romero, C., Ventura, S., Zafra, A., De Bra, P. (2009). Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational Systems. Computers & Education. Vol. 53, 828–840.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management. Vol. 24, No. 5, 513-523.

Salton, G. & McGill, M.J. (1983). Introduction to modern information retrieval, McGraw-Hill. New York, USA.

Salton, G. (1989). Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer. Addison-Wesley.

Salton, G. (1991). Developments in automatic text retrieval. Science, Vol. 253, 974-979.

Schiaffino, S., García, P., Amandi., A. (2008). eTeacher: Providing personalized assistance to e-learning students. Computers & Education, 51, 1744–1754.

Steinbach, M., Karypis, G., Kumar, V. (2000). A comparison of document clustering techniques.

Stubbs, E., Mangiaterra, N. & Martínez, A. (2000). Internal quality audit of indexing: a new application of interindexer consistency, Cataloguing & Classification Quaterly. Vol. 28(4) 53-70.

Sun, J., Wang, W., & Zhong, Y.X. (2001). Automatic text categorization based on k-nearest neighbor. Journal of Beijing University of Posts & Telecomms. Vol. 24, No. 1, 42-46.

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010

Van Rijsbergen, C. (1979). .

Vélez, O. & Santos, C. (2006). Sistemas Recomendadores: Un enfoque desde los algoritmos genéticos. Revista Industrial Data, Perú, Vol. 9(1), 23-31.

Venter, J., Waal, A. & Willers, C. (2007). Specializing CRISP-DM for Evidence Mining. Advances in Digital Forensics III. Springer Boston. Vol. 242.

Wanga, F. and Shao, H. (2004). Effective personalized recommendation based on time-framed navigation clustering and association mining, Expert Systems with Applications 27, 365–377.

Yolis, E., Britos, P., Perichisky, G. & García-Martínez R. (2003). Algoritmos Genéticos aplicados a la Categorización Automática de Documentos. Proceedings del VIII Congreso Argentino de Ciencias de la Computación. 1468-1479.

Zhang, Y., Gong, L & Wang, Y. (2005). An improved TF-IDF approach for text classification. Journal of Zhejiang University - Science A. Vol. 6, No. 1, 49-55.

Depósito Legal: PPX200002ZU2142 / ISSN:1856-4194. Volumen 9 Edición No 2 – Año 2010