Universidad de las Ciencias Informáticas

Facultad 10

Título: “Implementación de algoritmos de recomendaciones de imágenes para textos noticiosos.”

Trabajo de Diploma en opción al título de Ingeniero Informático

Autor(es): Yoannia Castillo Duvergel
Raidel Arenas Pérez.

Tutor(es): Joel Armada Herrera

Ciudad de la Habana, julio, 2009

“Año 50 de la Revolución”

"...Hay en el mundo un lenguaje que todos comprenden: es el lenguaje del entusiasmo, de las cosas hechas con amor y con voluntad, en busca de aquello que se desea o en lo que se cree."

Paulo Coelho

Agradecimientos

No hubiésemos podido concluir y presentar este trabajo de diploma más que en nuestras mentes, si no fuera por la Revolución Cubana y principalmente por nuestro Fidel Castro, por ser el creador de la Universidad de las Ciencias Informáticas, además de la incondicional entrega de nuestros amigos y compañeros, que sin escatimar su tiempo de descanso o en medio de un arduo trabajo, siempre estuvieron prestos a ayudarnos, imprimirnos fuerzas, entusiasmos e ideas para esta posible realidad.

A nuestro profesor, jefe de tribunal Manuel Vázquez, sin cuyos consejos, ayuda y aprobación, esta tesis no estuviera terminada.

A nuestro tutor que siempre pudimos contar con su ayuda y estuvo al tanto de este trabajo todo el tiempo.

A nuestros compañeros de estudios con los que pasamos 5 maravillosos años de universidad.

A nuestros padres y familiares que de una forma u otra han colaborado y apoyado incondicionalmente el desarrollo y culminación de esta tesis.

A todos muchas gracias, por haber confiado en nosotros.

Dedicatoria

Pudiera escribir los nombres de las personas que ayudaron a hacer mi sueño realidad, con el temor de olvidar mencionarlos a todos.

Primeramente agradecer a mis padres por darme esta educación, por tomar como propio este desempeño y confiar incondicionalmente en mí, por mis triunfos y éxitos, por ser mi clave, la que me abre puertas, me orienta caminos y me hacen amar tanto la vida. Quiero que sepan que son las personas que más quiero en la vida y espero que estén muy orgullosos de mí.

A mi hermano que fue mi apoyo, mi protector y mi ejemplo a seguir en todos los momentos de mi vida.

A mi primita querida Idelsis por su atención y por acogerme desde primer momento en la UCI como su hermana más pequeña.

A pachy por ser mi amigo, mi hermano y siempre confiar en mí.

A familiares y amigos en general que estuvieron al tanto de mis estudios y aportaron lo mejor de sí, para que pudiera culminarlos exitosamente.

A mis amigas: Marlen, Sahily, Norbelis, Epe, Gise e Isis porque me brindaron fuerzas, afecto y atención cuando las necesité.

A Delmis por su ayuda incondicional en todo momento.

A mi tesoro Manuel por sus consejos, críticas, por estimular en cada momento mi trabajo, por confiar en mí y darme fuerzas para continuar.

A todos, gracias por existir

Yoannia Castillo Duvergel

A mis padres, que sin su apoyo y confianza no hubiera podido hacer realidad este sueño.

A mi hermanita linda, Tata: que todos estos años de estudio y esfuerzos te sirvan de ejemplo.

A mi novia Puchy, que ha sabido comprenderme y darme mucho amor y cariño, Mimi te quiero.

A mi familia, que la quiero y se que ellos también a mí.

A mis amigos (mi piquete), que han estado conmigo en las buenas y en las malas.

Raidel Arenas Pérez

Declaración de autoría

Declaramos que somos los únicos autores del trabajo titulado:

Implementación de algoritmos de recomendaciones de imágenes para texto noticiosos.

y autorizamos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmamos la presente a los ____ días del mes de _____ del año _____.

_____ _____

Yoannia Castillo Duvergel Raidel Arenas Pérez

Autor Autor

_____

Joel Armada Herrera

Tutor

Resumen

En este trabajo de diploma se hace una investigación de los algoritmos de recomendación de imágenes. Se implementan tres de ellos adaptándolos a la recomendación de imágenes para textos noticiosos para perfeccionar la selección de gráficas relevantes a un texto dado. La investigación puede extenderse para que los métodos expuestos puedan ser usados en sitios web, cms u otras aplicaciones. Se expresan aspectos necesarios y determinantes para el objetivo: la importancia de la recuperación de información, indización de documentos, sistemas de recomendación y algoritmos paralelos.

Palabras Claves: Sistemas de recomendación, Sistemas de recuperación de información y algoritmos de recomendación.

Índice

Índice de Figuras

Introducción

La relevancia de una imagen (fotográfica o gráfica) en correspondencia a un texto noticioso en los medios de prensa queda a juicio de un operador. La presión de publicación implica que muchos textos aparezcan sin evidencia fotográfica o que no haya tiempo para realizar una correcta identificación de las ilustraciones a usar. Las preguntas pertinentes a esta tarea, entre otras, son: ¿Cuáles son las imágenes registradas que mejor se corresponden al texto dado?, ¿Cuántas veces se ha utilizado para otros textos?

Con la incorporación de nuevas herramientas y algoritmos independientes y fiables que faciliten el trabajo se gana en tiempo y calidad para que estos sitios de prensa puedan lograr un buen posicionamiento en la hoy vertiginosa “Autopista de la información”.

Los sistemas de recomendación han evolucionado y es posible encontrarlos en diversos ámbitos de aplicación, donde se han convertido en una herramienta fundamental para los proveedores y los servicios de información (1). En cada dominio se presentan diversos problemas a los que hay que dar diferentes soluciones. La especialización ha implicado que los sistemas de recomendación se hayan diversificado. (1)

Los sistemas de recomendación basados en contenido y de filtrado colaborativos que se analizarán en el presente trabajo utilizan antes para el pre-procesamiento de textos técnicas de recuperación de información donde una imagen es asociada a un texto dado, basada en la similitud existente entre un documento de la colección y el artículo.

La información visual desempeña un rol importante para los órganos de prensa escrita, no se limita únicamente a una función estética en la noticia, sino que constituye un recurso informativo que contextualiza, amplía y apoya la información textual. Además, encierra cualidades propias que son distintivas al resto de las fuentes de información. Pero en un sistema de prensa, prima siempre la variable de la velocidad. Es muy importante para un medio de prensa el publicar primero, pero las noticias se preparan por un grupo de trabajo que muchas veces no cuenta con las herramientas adecuadas para realizar este trabajo con la rapidez necesaria.

Esta es la oportunidad por la cual nos proponemos desarrollar un trabajo de diploma que contribuya a superar las debilidades descritas; implementando algoritmos de recomendación de imágenes para textos noticiosos.

De la situación antes expuesta se deriva el siguiente **problema a resolver**: ¿Qué algoritmos facilitar para encontrar imágenes relevantes para un texto, dado una colección de imágenes y sus asociaciones con otros textos?

Para lograr una óptima selección de los algoritmos es necesario estudiar a profundidad los implementados hasta el momento. Los sistemas de recomendación son el **objeto de estudio** central de este trabajo, tomando como **campo de acción** la implementación de algoritmos de recomendación de imágenes para textos noticiosos.

El **objetivo general de la presente investigación** es la selección e implementación informática de algoritmos de recomendación de imágenes para hacer más efectivo el trabajo de redacción en la prensa cubana.

Para cumplir con el objetivo propuesto se han definido los siguientes **objetivos específicos**:

- Realizar un estudio del estado del arte de los algoritmos de recomendación (incluyendo los algoritmos paralelos)
- Identificar las características relevantes de los textos para la selección de las imágenes asociadas.
- Identificar las características de las imágenes que son relevantes para la selección en relación a un texto dado
- Seleccionar e implementar los algoritmos más adecuados para sugerir imágenes a un texto noticioso según la clasificación del texto que se está escribiendo.
- Evaluar los algoritmos en cuanto a tiempo de procesamiento y recursos consumidos.

Surgen las siguientes **preguntas científicas:**

- ¿Qué tipos de métodos y algoritmos existen para clasificar, recomendar y cuáles son sus características?
- ¿Qué ventajas y desventajas presenta cada uno de estos algoritmos respecto a los demás?
- ¿Cuáles son los criterios más adecuados para clasificar textos noticiosos, que sean factibles también para clasificar imágenes?

- ¿Cuáles son los rasgos más representativos de las imágenes que se puedan utilizar como criterios para clasificar los textos?

- ¿Cuál será la vía para obtener la imagen en todo momento o tener actualizado en tiempo real un ranking de estas?

Las **principales tareas investigativas** que se proponen para dar respuesta a las preguntas antes expuestas se listan a continuación:

- Análisis de los principales problemas de la prensa cubana a la hora de seleccionar una imagen determinada para un texto y estudio de los algoritmos de recomendación de imágenes.

- Caracterización de cada uno de los algoritmos analizados en cuanto a ventajas y desventajas para implementar los mejores.

- Definición de los criterios más importantes de los algoritmos ya seleccionados para la clasificación de textos que sirven de igual forma para las imágenes, así como el análisis de cuáles son paralelos o paralelizables.

- Colección de una serie de noticias e imágenes para poder realizarle pruebas a los algoritmos que se confeccionen.

- Implementación de los algoritmos en lenguaje Python incluyendo uno paralelo y realización de pruebas a los mismos.

Para el desarrollo de este trabajo se emplearon los siguientes **métodos científicos de investigación**:

Los métodos teóricos utilizados por el autor en esta investigación son:

- **Analítico-Sintético**: Se buscaron y analizaron diferentes documentos, algoritmos y teorías, acerca de la recomendación y reconocimiento de imágenes, extrayendo los elementos fundamentales relacionados con nuestra temática.

- **Inductivo-Deductivo**: Le permitió al autor obtener conocimientos generalizados de reconocimiento de imágenes a partir del análisis de lo particular a lo general.

Los métodos empíricos utilizados por el autor en esta investigación son:

- **Observación**: Para ver y constatar los problemas que existen en la prensa por falta de una aplicación que agilice todo el trabajo en la prensa cubana.

Estructuración del contenido con una breve explicación de sus partes.

El presente trabajo consta de una introducción, tres capítulos, conclusiones generales, recomendaciones, referencias bibliográficas utilizadas durante el desarrollo del mismo y por último, los anexos que complementan el cuerpo del trabajo.

Capítulo 1: Fundamentación teórica. Se analizarán conceptualmente qué son los sistemas de recomendación, recuperación de información e indexación de textos o documentos y se caracterizarán y confrontarán los modelos actuales. Se incluirá una breve descripción de las técnicas, herramientas utilizadas, procesamiento de imágenes y algoritmos paralelos para la implementación, así como una reseña histórica, principales tendencias, lenguaje de programación y metodología a utilizar.

Capítulo 2: Descripción y Análisis de la solución propuesta. En este capítulo se presentarán las propuestas de soluciones para este trabajo basada en la situación, se describirán todo lo referente a la implementación con sus funcionalidades esenciales, se demostrarán las dependencias del compilador y del runtime entre los componentes del compilador del software con su fundamentación.

Capítulo 1

Fundamentación Teórica

Resumen

En este capítulo se brinda información conceptual referente del tema a tratar, se explican los fundamentos teóricos los sistemas de recomendación, la recuperación de información, indización de textos o documentos, procesamiento de imágenes y algoritmos paralelos, así como una reseña histórica, principales tendencias, herramienta, lenguaje de programación y metodología a utilizar.

1.1 Sistemas de recomendación.

Los Sistemas de Recomendación constituyen uno de los segmentos de más rápido crecimiento de Internet. Estos sistemas ayudan a aminorar la sobrecarga de información focalizando al usuario en sus necesidades y proporcionando en cada caso información personalizada.

La recomendación es un sistema vista de dos modos (por lo general, tales como personas y artefactos, películas y libros) y se ha estudiado en ámbitos que se centran en el aprovechamiento de recursos de información en línea, agregación de información y planes sociales para la toma de decisiones.

En los sistemas de recomendación es muy importante el volumen de la información a evaluar, que, para una dada velocidad de respuesta, determina la precisión de las recomendaciones. Factores como el tiempo de vida (del elemento a evaluar), el tipo de elemento (películas, personas, artículos, imágenes, etc.) y la cantidad de opciones generadas influyen de manera directa en la calidad de la recomendación.

1.1.1 Evolución e historia.

El comienzo de este tipo de sistemas se remonta a principios de la década de los 90 del pasado siglo, cuando comienzan a surgir dentro de los servicios de newsgroups (grupos de noticias) servicios de filtrado de noticias que permitían a su comunidad de usuarios acceder exclusivamente a aquellas potencialmente de su interés. (2)

Los sistemas de recomendación han evolucionado y es posible encontrarlos en diversos ámbitos de aplicación tales como el comercio electrónico, donde se han convertido en una herramienta fundamental para los proveedores en línea (1) y los servicios de información científica.

1.1.2 Categorías de los sistemas de recomendación.

Los sistemas de recomendación se han organizado en tres categorías principales de acuerdo a la habilidad que utilizan para realizar el filtrado: (1)

- Sistemas de recomendación sociales,
- Sistemas de recomendación basados en contenido.
- Sistemas de recomendación basados en factores económicos

Conozcan cada uno de ellos con mayor detalle.

Los sistemas de filtrado social o filtrado colaborativos

Los sistemas de filtrado social, usualmente conocidos por el anglicismo "filtrado en colaboración o colaborativos" (2), utilizan la información histórica de recomendaciones a usuarios de características similares o semejantes para generar nuevas recomendaciones, obviando el contenido de los recursos (se basa exclusivamente en las valoraciones que éstos reciben por parte de los usuarios o documentos). En este tipo de sistemas es habitual agrupar a los usuarios o textos en categorías específicas o estereotipos que los caracterizan a través de una serie de valores de preferencia definidos por defecto y que representan las necesidades de información y hábitos de búsqueda más comunes del grupo. Para cada elemento "usuario" se crea un conjunto de "vecinos cercanos" (3), en función de sus semejanzas. Los resultados para los elementos no calificados se predicen en base a la combinación de puntos (*scores*) conocidos de los vecinos cercanos (3).

En el filtrado colaborativo, el sistema no analiza los elementos evaluados, sino que las recomendaciones se basan solamente en la similitud de elementos analizados. Esto trae consigo algunas limitaciones:

- No es posible hacer recomendaciones hasta que el elemento a analizar no esté bien configurado o esté lo suficientemente completo para definir su grupo de vecinos cercanos

- Este tipo de sistemas tiende a ofrecer resultados pobres cuando se dispone de poca información de las cosas a analizar o las preferencias a considerar son muy heterogéneas.

Los sistemas de filtrado basados en contenido.

Los sistemas de filtrado basados en contenido generan recomendaciones comparando las preferencias del usuario o contenido analizado (expresadas por éste de forma implícita o explícita) (1) con los metadatos o características utilizados en la representación de los recursos o productos, ignorando de esta forma la información relativa a otros usuarios. Estos sistemas, al igual que los sociales, son poco fiables cuando se dispone de poca información sobre las cosas a analizar.

Estos sistemas emplean técnicas de recuperación de información (3). Por ejemplo, un documento de texto es recomendado basado en una comparación entre su contenido y los metadatos de la imagen.

Típicamente, el documento facilitará una lista de palabras claves y sus pesos correspondientes (3).

Para identificar el tema del documento se hace un análisis de frecuencia para extraer las palabras claves. Si una imagen corresponde a un documento, los pesos de las palabras extraídas se añaden a los pesos de los metadatos de las imágenes correspondientes. Este proceso es conocido como retroalimentación de relevancia.

Este método de recomendación presenta algunas limitaciones tales como la sobre-especialización; el sistema sólo muestra al usuario elementos similares a los que ya ha visto anteriormente y eventualmente oculta recomendaciones potencialmente relevantes. Algunas veces este problema es resuelto agregando a la búsqueda aleatoriedad (mediante algoritmos genéticos). Otra limitante se presenta al encontrar información multimedios, (con frecuencia presente en páginas Web) puesto que cuando las recomendaciones son hechas sobre documentos de texto, está información es ignorada. (3)

1.2 Sistemas de recuperación de información.

La recuperación de información es una disciplina de creciente interés, teniendo en cuenta el aumento de la disponibilidad de documentos en soporte electrónico y de la necesidad de obtener en cada momento aquellos que responden a una necesidad informativa dada. (4) O sea se define como el problema de la selección de información, depositada en un medio de almacenamiento, en respuesta a consultas realizadas por un usuario. Las semejanzas que existen entre la recuperación de información y otros

campos vinculados al procesamiento de la información, propician que se trasladen hacia el área de los sistemas encargados de realizar esta tarea.

Salton opina que “la recuperación de información se entiende mejor cuando uno recuerda que la información que se procesa consiste en documentos”, queriendo diferenciar a los sistemas encargados de su gestión con otro tipo de sistemas, como los gestores de base de datos. “Cualquier SRI puede ser descrito como un conjunto de ítem de información (DOCS), un conjunto de peticiones (REQS) y algún mecanismo (SIMILAR) que le determine que ítem satisfacen las necesidades de información expresadas por el usuario en la petición”. (5)

SIMILAR es el proceso de determinación de la similitud existente entre la presentación de la pregunta y la representación de los ítems de información.

Los ***Sistemas de Recuperación de Información (SRI)*** son una clase de sistemas de información que tratan con bases de datos compuestas por documentos y procesan las consultas de los usuarios permitiéndoles acceder a la información relevante en un intervalo de tiempo apropiado. Estas consultas son sentencias formales mediante las cuales el usuario expresa sus necesidades de información, formuladas usando un lenguaje de consulta. (5)

Para mayor entendimiento del concepto, se presenta la siguiente figura:

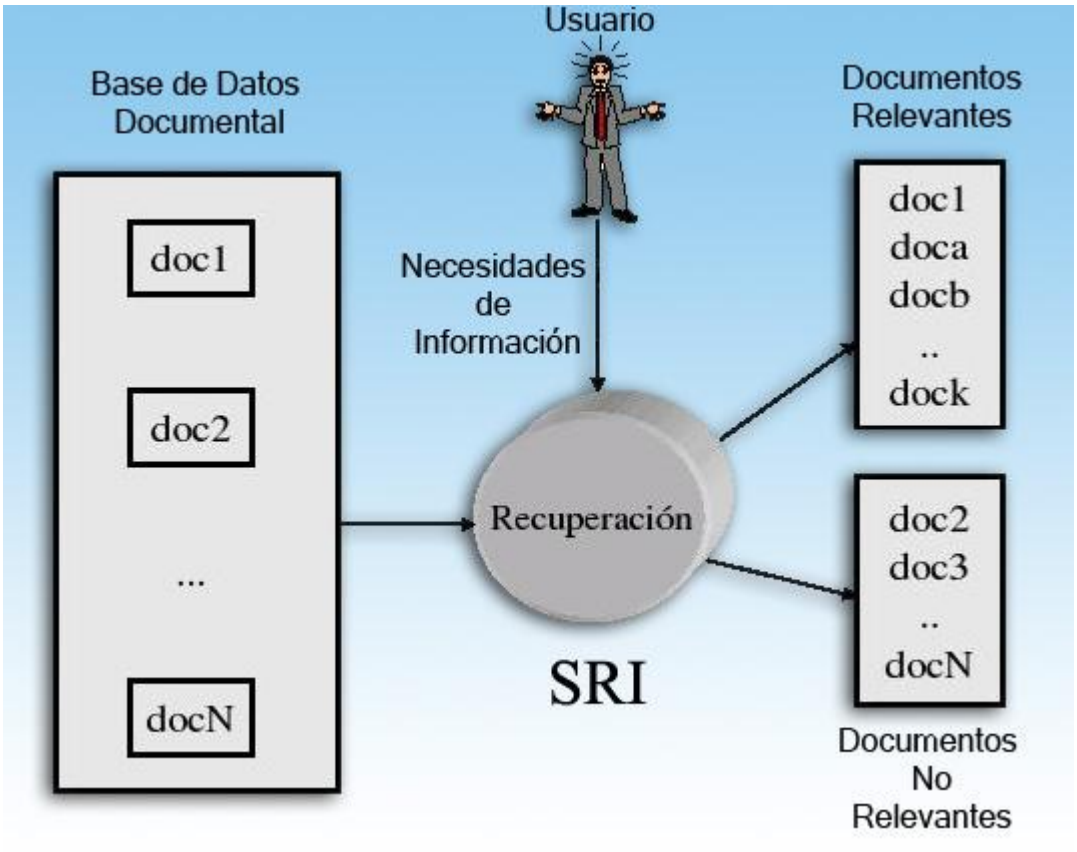

Figura 1: Proceso de recuperación de información.

Chowdhury identifica el siguiente conjunto de funciones principales en un SRI.

1. Identificar las fuentes de información relevantes a las áreas de interés de las solicitudes de los usuarios.
2. Analizar los contenidos de los documentos.
3. Representar los contenidos de las fuentes analizadas de una manera adecuada para compararlas con las preguntas de los usuarios.
4. Analizar las preguntas de los usuarios y representarlas de una forma que sea adecuada para compararlas con las representaciones de los documentos de la base de datos.
5. Realizar la correspondencia entre la representación de la búsqueda y los documentos almacenados en la base de datos.
6. Recuperar la información relevante.

7. Realizar los ajustes necesarios en el sistema basados en la retroalimentación con los usuarios.

1.2.1 Evolución de los SRI.

La simplificación de este proceso según Baesa-Yates:

Desarrollos iniciales. El autor refleja la existencia de métodos de recuperación de información con las antiguas colecciones de papiros. Una sería la tabla de contenidos de un libro, la cual es sustituida por estructuras más complejas a medida que ha crecido el volumen de información a gestionar.

Recuperación de la información en las bibliotecas. Estas fueron unas de las primeras instituciones en utilizar estos sistemas. Al principio eran originados por las propias bibliotecas y luego se crea un mercado de aplicaciones informáticas especializadas en este sector. Aparecen varias generaciones: mecanización de los catálogos manuales, aumento de las posibilidades de búsqueda y una tercera generación se encuentra vinculada al desarrollo de interfaces gráficas, características de hipertextos, arquitecturas de sistemas abiertos y automatización de procesos.

La World Wide Web. Los SRI y su evolución han tomado un camino lógico hacia la web, donde se encuentra una gran aplicación práctica y un aumento significativo del número de usuarios en el campo de los directorios y motores de búsqueda. (6)

Este proceso de analizar el contenido de los documentos (recursos, imágenes, textos, etc.) y asignarle términos descriptivos, generando un índice de puntos de acceso a través del cual poder recuperar dichos documentos es denominado indización social.

La distinción entre tipos de indización se suele realizar en función de cómo se realiza el proceso, dando pie a dos grandes categorías: indización humana e indización automática.

La indización humana es un proceso intelectual, donde es una persona (tradicionalmente un indizador profesional) quien, tras analizar el contenido del documento o parte de este, le asigna aquellos términos de indización automática, por el contrario, es realizada por algoritmos que mediante diversas técnicas o métodos determinan cuál es el peso con el que cada uno de los términos que aparecen en el documento representa su contenido temático. (7)

La indización de textos o documentos es el empleo eficiente de la información contenida en un texto T que requiere la disponibilidad de una representación compacta de tal información, simplemente llamada representación o indización de T.

1.2.2 Modelos de recuperación de información.

Los **modelos de recuperación** tienen como propósito proveer el proceso de comparación entre una consulta establecida y un conjunto de textos sobre los que se realiza la consulta, para esto definen distintas formas de representar los documentos. Estos **modelos de recuperación** están pensados únicamente para documentos de contenido textual.

Su funcionamiento es sencillo, para cada documento se construye un índice determinado en función del texto contenido en el documento. Derivado de esto tenemos el concepto de índice invertido que equivale a decir que la relación de los documentos en los que aparece en una determinada palabra.

Los índices de los documentos tienen en cuenta la frecuencia de aparición de las palabras. Cada documento se representa a través de un vector como los que se muestran a continuación:

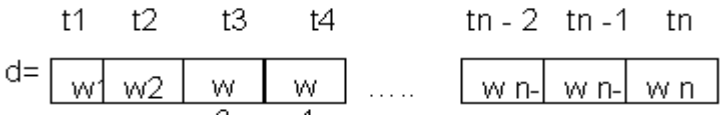

Figura 2: Representación de vectores

Donde W_i indica la importancia del índice t_i en el documento d. Suele tomar valor en el intervalo [0,1]. A las distintas formas de obtener el valor W_i se les denomina esquemas de asignación de pesos.

Entre los distintos esquemas de asignación de pesos cabe distinguir:

Esquema Binario: Se asigna peso1 (wi = 1) si la palabra aparece en el documento y peso 0 (wi = 0) en caso contrario.

Frecuencia Inversa de Documento (IDF):

IDF = log2 N / Ni + 1

Wij =IDF i * F ij

N = Nº total de documentos

Ni = Nº de documentos en los que aparece el término i

Fij = Frecuencia interna del término i en el documento j (8)

1.2.3 Clasificación de los modelos de recuperación.

- Modelos clásicos: Se fundamentan en la referencia al contexto temático expresado a partir de términos de indización, (9) entre los que se encuentran los modelos probabilístico, booleano y vectorial. Donde nos centraremos en este último.

- Modelos estructurales**:** Se generalizan en la estructura del texto, se destacan las listas no sobrepuestas y el método de los nodos proximales. (9)

Modelo de Recuperación Booleano:

El modelo de recuperación booleano es uno de los métodos más utilizados para la recuperación de información. Este modelo se basa en la agrupación de documentos, los cuales están compuestos por conjuntos de términos y en la concepción de las preguntas como expresiones booleanas, de ahí deriva el nombre de *modelo de recuperación booleano.* La principal característica es la consideración de la relevancia como un carácter puramente binario. Dentro del modelo, se presenta el lenguaje de consulta, y el mecanismo de indización utilizando los denominados índices inversos o archivos fantasmas.

Características Principales

Es un modelo de recuperación simple, basado en la teoría de conjuntos y el álgebra booleana. Dada su inherente simplicidad y su pulcro formalismo ha recibido gran atención y ha sido adoptado por muchos de los primeros sistemas bibliográficos comerciales. Su estrategia de recuperación está basada en un criterio de decisión binario (pertinente o no pertinente) sin ninguna noción de escala de medida, sin noción de un emparejamiento parcial en las condiciones de la pregunta. Para este modelo las variables de peso de los términos índice son todas binarias.

Indización de Documentos en el Modelo Booleano

Dentro de un sistema Booleano, los documentos se encuentran representados por conjuntos de palabras clave (términos). La indización se realiza asociando un peso binario a cada término del índice: 0 si el término no aparece en el documento y 1 si aparece aunque sea una sola vez. Las búsquedas consisten en expresiones de palabras claves conectadas con algunos operadores lógicos (AND, OR y NOT). El grado de similitud entre un documento y una consulta sería también binario y un documento sería relevante cuando su grado de similitud sea igual a 1, de lo contrario el documento no tendría ninguna

relevancia en cuanto a la consulta. Por tanto, en el caso de los SRI Booleanos, la función de indización quedaría así: (10)

F: D * T-----> {0,1}

Modelo de Recuperación Vectorial.

El modelo de recuperación vectorial o de espacio vectorial propone un marco en el que es posible el emparejamiento parcial asignando pesos no binarios a los términos índice de las preguntas y de los documentos. Estos pesos de los términos se usan para computar el grado de similitud entre cada documento guardado en el sistema y la pregunta del usuario.

Características Principales.

Ordenando los documentos recuperados en orden decreciente a este grado de similitud, el modelo de recuperación vectorial toma en consideración documentos que sólo se emparejan parcialmente con la pregunta, así el conjunto de la respuesta con los documentos alineados es mucho más preciso (en el sentido que empareja mejor la necesidad de información del usuario). La mayoría de los motores de búsqueda lo implementan como estructura de datos y que el alineamiento suele realizarse en función del parecido (o similitud) de la pregunta con los documentos almacenados.

Indización de Documentos en el Modelo Vectorial

Sea *D* el conjunto de documentos y *T* el conjunto de términos índice. El mecanismo de indización de este modelo se presentaría de la siguiente forma:

F: D * T------>I

Lo más habitual sería trabajar con una función de evaluación normalizada donde los vectores tengan los pesos reales, donde *I* = [0*;* 1]. Como hemos dicho anteriormente, una de las múltiples formas de definir la función *F* es la frecuencia inversa del documento (*idf)*. La bondad de la indización idf está en que pondera la importancia de los términos en función de su aparición en el resto de los documentos de la base documental además de su frecuencia de aparición en el documento actual. (10)

Modelo de Recuperación Probabilístico

El modelo de recuperación probabilístico, tiene como base principal de su funcionamiento el cálculo de la probabilidad de un documento de ser relevante a una pregunta dada. Los modelos anteriores están basados en la equiparación en la forma más dura. En el booleano es o no coincidente, y en el vectorial el umbral de similitud es un conjunto, y si un documento no está no es similar y, por lo tanto, no recuperable.

2.- Características Principales.

Dentro de la recuperación probabilística, se utiliza el modelo de recuperación probabilístico de independencia de términos binarios donde:

La probabilidad de los términos es independiente (un término es independiente de los otros).

Los pesos asignados a los términos son binarios

La equiparación probabilística se basa en que, dados un documento y una pregunta, es posible calcular la probabilidad de que ese documento sea relevante para esa pregunta. (9)

1.2.4 Comparación de los modelos clásicos.

Los beneficios de utilizar el método booleano es que es un modelo de recuperación sencillo. Mientras que la problemática es que básicamente se tiene que considerar la relevancia como un aspecto puramente binario a diferencia del modelo del espacio vectorial donde los términos individuales considerados en la consulta no están conectados por ningún operador (ni conjunción, ni disyunción, ni negación), la consulta se considera como un todo. La ventaja del modelo vectorial es que permite hacer correspondencias parciales, es decir, ordena los resultados por grado de relevancia. Su principal inconveniente es que no incorpora la noción de correlación entre términos (problema de todos los modelos clásicos). Por otra parte los modelos de recuperación probabilísticos envuelven muchos cálculos y premisas, todos los documentos seleccionados no son realmente relevantes. Se debe considerar la posibilidad de que un documento sea relevante o no, dado que haya sido ya seleccionado. Se obtienen algunos buenos resultados pero no son mucho mejores que los obtenidos en el modelo booleano y en el vectorial. Haciendo énfasis en lo antes mencionado este trabajo se concentrará en el modelo del espacio vectorial.

1.2.5 Herramientas.

Existen disímiles implementaciones de los modelos de recuperación, las más populares se presentan a continuación.

1- Lemur:

Como sistema de Recuperación de Información, Lemur permite todas las etapas desde la indización a la búsqueda de documentos.

Lemur aporta una poderosa API implementada en C++ y está diseñada para trabajar en todos los sistemas operativos, permite la indización incremental e indiza atributos de los documentos.

2- Xapian:

Xapian es una biblioteca de funciones OpenSource de Recuperación de Información, para crear motores de búsqueda está escrita en C++, pero también se encuentra disponible en otros lenguajes como Perl, Python, PHP, Java, C# y Ruby hasta el momento.

Xapian contiene una API potente y adaptable que le facilita al programador los procesos de indización y búsqueda.

3- Terrier

Terrier está implementado en Java, facilita mucho el trabajo a los programadores ya que permite indizar una colección de documentos de forma que se sabe cuántos documentos contienen un término determinado, permite la búsqueda.

4- Lucene

Algoritmo de búsqueda potente, fiable y eficiente. Permite ordenar resultados por relevancia. Lenguaje de consulta muy potente. Búsqueda por campos y rango de fechas. Búsqueda multi-índice y combinación de resultados. Permite búsquedas mientras se actualiza el índice (11)

Comparación de las herramientas.

- Lucene es multiplataforma, al igual que Lemur y xapian, las demás tecnologías no.
- Terrier y Lucene son implementados en Java, Lemur y Xapian en C++, aunque todas tienen soporte para otros lenguajes de programación.

- Todas indizan diferentes formatos de texto como: PDF, WORD, HTML, HTM, TXT, XML, RTF, entre otras.
- Todas admiten Stemming para varios idiomas.
- Todas las tecnologías son OpenSource.
- Lucene permite búsqueda mientras se actualiza el índice.
- Lemur y Lucene permiten la indización incremental, Xapian y Terrier no. (11)
- Todas trabajan con modelos probabilísticos, excepto Lucene que trabaja con el modelo de espacio vectorial, pero es muy inestable porque está en desarrollo.

Cabe mencionar que es de suma importancia conocer que existen otras tecnologías de Recuperación de información por eso en este apartado se muestran algunas de las ventajas que una u otra pueden tener.

1.3 Algoritmos paralelos.

Un algoritmo paralelo, al contrario de los algoritmos clásicos o algoritmos secuenciales, es un algoritmo que puede ser ejecutado por partes en distintas unidades de procesamiento al mismo tiempo, y finalmente unir todas las partes y obtener el resultado esperado.

Un ejemplo de algunos algoritmos que son fácilmente divisibles en partes lo constituye un algoritmo que calcule todos los números primos entre 1 y 100, donde se dividiría los números originales en subconjuntos y calcular los primos para cada uno de los subconjuntos de los números originales; al final, vincularíamos los resultados y tendríamos la solución final.

Existen algunos problemas que son fáciles de paralelizar, de ahí que se conozcan como problemas inherentemente secuenciales. Tanto los métodos numéricos iterativos (12) como el método de Newton (13) o el problema de los tres cuerpos (14) constituyen un ejemplo de estos métodos. Por otro lado, algunos problemas son difícilmente paralelizables, aunque tengan una estructura recursiva. Como es la búsqueda primero en profundidad en un grafo.

La importancia de estos algoritmos paralelos radica en la rapidez a tratar grandes tareas de computación mediante la paralelización que no por medio de técnicas secuenciales. Esta es la forma en que se trabaja en el desarrollo de los procesadores modernos, porque resulta más difícil incrementar la capacidad de procesamiento con un único procesador que aumentar su capacidad de cómputo mediante la inclusión de

unidades en paralelo, obteniéndose así la ejecución de varios flujos de instrucciones dentro del procesador. Pero hay que tener cuidado con la excesiva paralelización de los algoritmos ya que cada algoritmo paralelo tiene una parte secuencial y debido a esto, los algoritmos paralelos puedes llegar a un punto de saturación (*Ley de Amdahl*) (15). Por todo esto, a partir de cierto nivel de paralelismo, añadir más unidades de procesamiento puede sólo incrementar el coste y la disipación de calor.

La complejidad o el coste de los algoritmos secuenciales se estima en términos del espacio (memoria) y tiempo (ciclos de procesador) que requiera. Los algoritmos paralelos también necesitan optimizar la comunicación entre diferentes unidades de procesamiento. Esto se obtiene mediante la aplicación de dos paradigmas de programación y diseño de procesadores distintos: memoria compartida o paso de mensajes.

La técnica memoria compartida necesita del uso de pasadores (cerrojos) en los datos para impedir que se modifique simultáneamente por dos procesadores, por lo que se produce un coste extra en ciclos de CPU desperdiciados y ciclos de bus. Exigiendo a serializar alguna parte del algoritmo.

La técnica paso de mensajes usa conductos y mensajes pero esta comunicación añade un coste al bus, memoria adicional para los mensajes y latencia en el mensaje. Los diseñadores de procesadores paralelos usan buses especiales para que el coste de la comunicación sea pequeño pero siendo el algoritmo paralelo el que decide el volumen del tráfico.

Concluyendo, una subclase de los algoritmos paralelos, los algoritmos distribuidos son algoritmos diseñados para trabajar en entornos tipo clústeres y de computación distribuida, donde se usan otras técnicas, fuera del alcance de los algoritmos paralelos clásicos.

1.4 Tratamiento de imágenes.

Definición y funciones de la fotografía

La fotografía es un mensaje encargado de mostrar la imagen real de lo acontecido. Las funciones de la fotografía en la prensa diaria son varias:

- Informativa. Tanto mejor será una fotografía cuanto menos texto exija para su explicación.

- Documental. En este sentido, puede ser descriptiva (si muestra detalles) o un modo de autentificación de lo que se dice en el texto.

- Simbólica. La fotografía se convierte en símbolo de algo.

- Ilustrativa. Habituales en los reportajes. Rompen la monotonía del texto.

- Estética. No es conveniente su uso con este único fin, salvo en revistas ilustradas u otras excepciones.

- De entretenimiento. Suelen ser fotografías que captan el lado humorístico de algo. (22)

1.4.1 Evolución e historia.

Antes de la introducción de la fotografía, un comandante dependía de observadores que exploraran un área y desde un terreno alto reconociera o vigilara la actividad enemiga, informando sobre la base de su vista y memoria. Una vez introducida la fotografía, esta información quedaba grabada con detalles en un medio físico, mejorando así la calidad de información obtenida.

- Primera Guerra Mundial

La Primera Guerra Mundial vio el primer uso global de la fotografía terrestre y aérea. Por primera vez, un general podía acceder a información reciente y exacta. El valor de esta información provocó que fueran atacados los observadores ubicados en aerostatos y aviones, primero con armas primitivas y luego con ametralladoras, desarrollándose así los primeros aviones caza. Frank Luke, piloto estadounidense, comenzó a utilizar municiones incendiarias y fueron tantos los aerostatos incendiados y derribados, que ganó el apodo de rompe-globos.

- Entre guerras

Al terminar la guerra, disminuyó la utilización que se venía haciendo de esta tecnología, lo que ocasionó un estancamiento en su desarrollo. Al presentarse luego la amenaza militar de Japón y Alemania, resurgió la capacidad técnica de las grandes potencias, ayudando a los planificadores como Eisenhower a prepararse adecuadamente en este aspecto para el caso de una futura guerra.

En los años 1930, se introdujeron las imágenes infrarrojas. Una las primeras aplicaciones fue la de investigar los fraudes en pinturas.

- Segunda Guerra Mundial y guerra fría

En la Segunda Guerra Mundial se introdujo el radar con su capacidad para detectar aviones enemigos. En los primeros años de la guerra fría, los soviéticos, al detectar aviones de vigilancia, hacían uso de

transmisores direccionales para atraerlos a su espacio aéreo con la intención de derribarlos. Las pantallas circulares ya se usaban en los aviones de mayor tamaño. Haciendo uso de la pantalla, era posible navegar estas áreas peligrosas. De hecho, hay fotos publicadas por la Fuerza Aérea de los Estados Unidos donde se muestra esta práctica.

- Época post Vietnam

La época post Vietnam vio la introducción de sensores infrarrojos aéreos. Las diferencias en temperatura de un objeto y el terreno cercano permitían la detección de actividad militar. Estos sensores ya eran capaces de grabar su información en medios magnéticos, a la cual se accedía cuando el avión regresaba a su base.

Esta época también vio la introducción del ultrasonido y la tomografía computarizada para usos médicos. La introducción del ultrasonido permitió ver diferencias en el tejido humano, lo que hizo posible detectar anomalías en su fisiología. También tecnologías derivadas se comenzaron a usar para detectar fallos materiales en productos manufacturados.

En los últimos años de la guerra fría, se introdujo la tecnología del radar de apertura sintética donde la apertura de una cámara optical sería emulada por las ondas de radar. La calidad de la imagen sería guardada como secreta, más un indicio se puede ver en un imagen duplicada por la NASA en el cual se puede ver un arroyo aun cuando está cubierta por arena. (23)

1.4.2 Análisis e información de las imágenes.

La optimización y análisis de las imágenes emplean técnicas que permiten modificar los datos originales de las imágenes de tal manera que se facilita la extracción de información derivada:

Imágenes que aportan datos:

Los metadatos o más comúnmente conocidos como Exif (*Exchangeable image file format*), no son más que una especificación de formato de imagen para archivos TIFF y JPEG creada en Japón en su versión 2.1 el 12 de Junio de 1998 y posteriormente en su versión 2.2 en Abril de 2002. Aunque la especificación no la mantiene ni la Industria ni ninguna organización casi todas las cámaras digitales y los sistemas operativos y programas de tratamiento de imagen utilizan el mismo estándar. (24)

Los metadatos son "datos sobre los datos" que incluyen muchos dispositivos y que los fabricantes justifican alegando que su inclusión mejora la edición, el visionado, el archivado y la recuperación de

documentos. Con el Exif podemos además añadir datos adicionales al archivo de imagen. Existe una gran cantidad de software capaz de leer estos metadatos en las fotografías digitales.

Teniendo en cuenta que los metadatos son incluidos a la imagen, pueden solucionar verdaderos problemas. Sólo conocer de su existencia, un poco de edición y precaución antes de hacer público un archivo pueden dar inconscientemente, mucho más información de lo que se analiza. (25)

1.5 Xapian.

Xapian es una biblioteca de recuperación de información, liberado bajo la licencia GPL. Se trata de un motor de búsqueda de texto completo para programadores.

Está escrito en C++, con enlaces para permitir el uso de Perl, Python, PHP, Java, Tcl, C# y Ruby. Xapian es muy portable y corre sobre Linux, Mac OS X, FreeBSD, NetBSD, OpenBSD, Solaris, HP-UX, Tru64, IRIX, y Microsoft Windows.

Xapian tiene gran capacidad de adaptación que permite a los desarrolladores agregar fácilmente indexación avanzada y búsqueda para sus propias aplicaciones. Apoya el modelo probabilístico de recuperación de información y un rico conjunto de operadores de búsqueda booleanos.

1.5.1 Características.

- Soporta Unicode y almacena los datos indexados en UTF-8.
- Búsqueda de palabras importantes asignándole mayor peso que las palabras sin importancia y así obtener una lista de documentos más similares
- Retroalimentación relevante dando uno o más documentos, xapian puede sugerir el término indexado más relevante a la consulta, sugerir documentos relacionados y clasificar los documentos.
- Frase y búsqueda aproximada, los usuarios pueden buscar palabras que ocurren en una frase exacta o dentro de un determinado número de palabras, ya sea en un determinado orden, o en cualquier orden.

- Posee una amplia gama de operadores de búsqueda booleanos estructurados. Los resultados de la búsqueda booleana se clasifican por el peso probabilístico. Los filtros booleanos también pueden aplicarse para restringir una búsqueda probabilística.

- Apoya las derivadas de los términos de búsqueda (por ejemplo, una búsqueda de "fútbol" coincide con los documentos que mencionan "fútbol" o "futbolista"). Esto ayuda a encontrar documentos pertinentes que de otro modo podrían perderse. Los stemmers están incluidos para Danés, Holandés, Inglés, finlandés, francés, alemán, húngaro, italiano, noruego, portugués, rumano, ruso, español, sueco y turco.

- Tiene comodín de búsqueda con apoyo (por ejemplo, "* XAP").

- Los sinónimos son compatibles, tanto explícitamente como una forma automática de la ampliación de consultas.

- Xapian puede sugerir correcciones ortográficas para el usuario suministrada por las consultas. Esto se basa en las palabras que se producen en los datos indexados, así que funciona incluso para decir que no se encuentran en un diccionario (por ejemplo, "xapian" se propuso como una palabra correcta para "xapain").

- Soporta archivos de base de datos mayores a 2 GB esenciales para la expansión de las grandes colecciones de documentos.

- Posee formatos de datos independientes de la plataforma, se puede construir una base de datos en un ordenador y buscar en otro.

- Permite la actualización y la búsqueda simultánea. Agrega nuevos documentos en la búsqueda de inmediato.

1.5.2 Evolución e Historia.

Xapian es una moderna biblioteca de clases, pero ha evolucionado por más de 25 años de experiencia académica y comercial.

Xapian es derivada en parte por el motor de la *Open Mascate*, desarrollado por *BrightStation* PLC y liberado bajo la GPL. Mascate fue construido para ser un sustituto para el propietario Mascate versión 3,6

de los sistemas de recuperación de información, que fue escrito casi en su totalidad en BCPL, y así extenderlo en las formas que se querían.

Muscat 3.6 fue escrita originalmente por el Dr. Martin Porter en la Universidad de Cambridge. En 1984, Martin y John Snyder fundaron CD Editorial Cambridge creada para explotar comercialmente la tecnología, la empresa pronto se reagrupó. La compañía fue nombrada Muscat Ltd cuando el foco pasó de los CDs a la web. Muscat Ltd fue comprado por Maid PLC, a la que renombraron primero Corporación de Diálogo y luego *BrightStation PLC* (cuando se vendió la marca y el contenido a Thomson).

A principios de 2001, la administración *BrightStation* fue renombrada de *Open Muscat* a *Omsee*, y poco después anunció que lamentablemente estaban tomando el desarrollo de fuente cerrada.

Un número de desarrolladores (ambos ex *BrightStation* y otros empleados de la incipiente comunidad *Open Muscat*) ocupa la última versión GPL y han continuado el desarrollo bajo la GPL. Este proyecto fue inicialmente conocido como *OmSeek*, pero se quejó de esta *BrightStation* era demasiado cerca de su nombre *Omsee*. Es más fácil de cambiar que argumentar, y el nombre fue elegido Xapian.

La última noticia fue *BrightStation* numeradas al *Open Muscat* 0.4.1, si bien hubo un amplio desarrollo en el CVS después de eso. La primera noticia oficial fue Xapian 0.5.0.

Open Muscat (utilizando el *muscat36 backend*) que fue el motor que potencia la recuperación *BrightStation* del *Webtop*, motor de búsqueda, que ofrece una búsqueda más de alrededor de 500 millones de páginas web (alrededor de 1,5 terabytes de archivos de base de datos), donde las búsquedas toman menos de un segundo. Todo el motor de búsqueda, incluidos los rastreadores web y el índice de construcción incorporan esta tecnología. *Webtop* también incorpora la tecnología de *Muscat Ltd's EuroFerret* que es un motor de búsqueda.

1.6 Lenguaje utilizado.

El lenguaje de programación **Python** creado e interpretado por Guido van Rossum en CWI en los Países Bajos como un sucesor del lenguaje de programación ABC, capaz de manejar excepciones e interactuar con el sistema operativo Amoeba (12) .en el año 1991.

Usualmente este lenguaje es comparado con otros lenguajes como son TCL, Perl, Scheme, Java y Ruby. En la actualidad Python se desarrolla como un proyecto de código abierto, administrado por la Python Software Foundation.

Python es calificado como la "oposición leal" a Perl, lenguaje con el cual mantiene una rivalidad amistosa. Los usuarios de Python consideran a éste mucho más limpio y elegante para programar.

Python permite fragmentar el programa en módulos reutilizables desde otros programas Python. Aparece con una gran colección de módulos estándar que se pueden manipular como base de los programas (o como ejemplos para empezar a aprender Python). También hay módulos incluidos que proporcionan E/S de ficheros, llamadas al sistema, sockets y hasta interfaces a GUI (interfaz gráfica con el usuario) como Tk, GTK, Qt.

Como es un lenguaje de programación interpretado, ahorra un tiempo considerable en el desarrollo del programa, puesto que no es necesario compilar ni enlazar. El intérprete se puede utilizar de modo interactivo, facilitando experimentar con características del lenguaje, escribir programas desechables o probar funciones durante el progreso del programa.

1.6.1 Características y paradigmas del lenguaje.

Python constituye un lenguaje de programación multiparadigma. Esto significa que más que exigir a los programadores a adoptar un estilo particular de programación, permite varios estilos: programación orientada a objetos, programación estructurada y programación funcional. Muchos paradigmas están soportados mediante la utilización de extensiones. Python usa tipo de dato dinámico y referencia contenida para el manejo de memoria. Una característica importante de Python es la resolución dinámica de nombres, lo que enlaza un método y un nombre de variable en el momento de la ejecución del programa.

Otro punto del diseño del lenguaje era la facilidad de extensión. Nuevos módulos se pueden escribir fácilmente en C o C++. Python puede usarse como un lenguaje de extensión para módulos y aplicaciones que necesitan de una interfaz programable. Aunque el diseño de Python es de alguna manera discrepante a la programación funcional tradicional del Lisp, existen bastantes analogías entre Python y los lenguajes minimalistas de la familia del Lisp como puede ser *Scheme*.

1.7 Metodología de desarrollo.

Las metodologías de desarrollo del software son una serie de procedimientos, técnicas y ayudas a la documentación para el proceso de desarrollo de productos de software. La utilización de metodologías ayuda a los desarrolladores, indicando los pasos a seguir para lograr el producto final deseado, describiendo además que papel debe desempeñar cada miembro del equipo en el desarrollo de las actividades del proyecto. Además cada metodología define la información que se debe tener en cuenta antes de comenzar cada actividad y los entregables que se deben producir como resultado de dichas actividades.

Hasta hace muy poco tiempo el proceso de desarrollo del software brindaba vital importancia al control del proceso mediante rigurosas definiciones de roles, actividades y artefactos incluyendo modelado y documentación detallada. Este esquema tradicional ha demostrado ser necesario para el desarrollo de grandes proyectos. Sin embargo, en muchos de los proyectos actuales, más pequeños en cuanto a recursos y tiempo, en los que se sigue exigiendo una alta calidad, estas metodologías tradicionales no siempre son las más adecuadas. Es ante esta situación que surgen las metodologías ágiles, orientadas a proyectos pequeños sin renunciar a las prácticas esenciales para asegurar la calidad del producto.

Dadas las características del proyecto “Informatización de la prensa”, se decidió, que la más adecuada para dirigir el proceso de desarrollo de los algoritmos era Scrum (Ver Anexo1)

Capítulo 2

Descripción y Análisis de la solución propuesta

Resumen.

En este capítulo se presentarán las propuestas de soluciones para este trabajo basada en la situación, se describirán todo lo referente a la implementación con sus funcionalidades esenciales, se demostrarán las dependencias del compilador y del runtime entre los componentes del compilador del software con su fundamentación.

2.2 Solución propuesta.

Dado el problema expuesto anteriormente, nuestro trabajo presenta algunas propuestas de solución. (**Figura: 3**)

Estas soluciones están conformadas por 4 variantes, en dos de ellas se puede asumir las imágenes que aportan datos e imágenes que no aportan datos. Estos datos están basados no en el contenido visual de la imagen sino internamente (metadatos). Pero, como la mayoría de las imágenes no presentan estos metadatos no podemos trabajar directamente con estas imágenes para la recomendación. Nos apoyaremos de esta forma tomando todas las imágenes como si no tuvieran datos.

Y las otras dos variantes, la primera es en la que interviene un usuario el cual es el encargado de recomendar imágenes para un documento dado o sea un usuario establece la relación entre el documento y la imagen, pero en los medios de prensa para el proceso de publicación intervienen más de un usuario y puede ocurrir que a dos documentos similares le otorguen imágenes que no se correspondan unas con otras, proporcionando de esta forma distintos criterios visuales y un aumento en el volumen de datos almacenados; la segunda variante y fue la que escogimos para el desarrollo de los algoritmos de recomendación, donde a un documento se le recomienda imágenes, para esto nos apoyamos en la colección de documentos almacenados con sus imágenes asociadas, o sea, buscar la similitud entre el documento entrante (noticia) y los documentos de la colección, obteniendo para el documento noticia la

imagen del documento más similar a este, y de esta manera no existen estas diferencias de criterios visuales ni duplicados de imágenes en la base de datos.

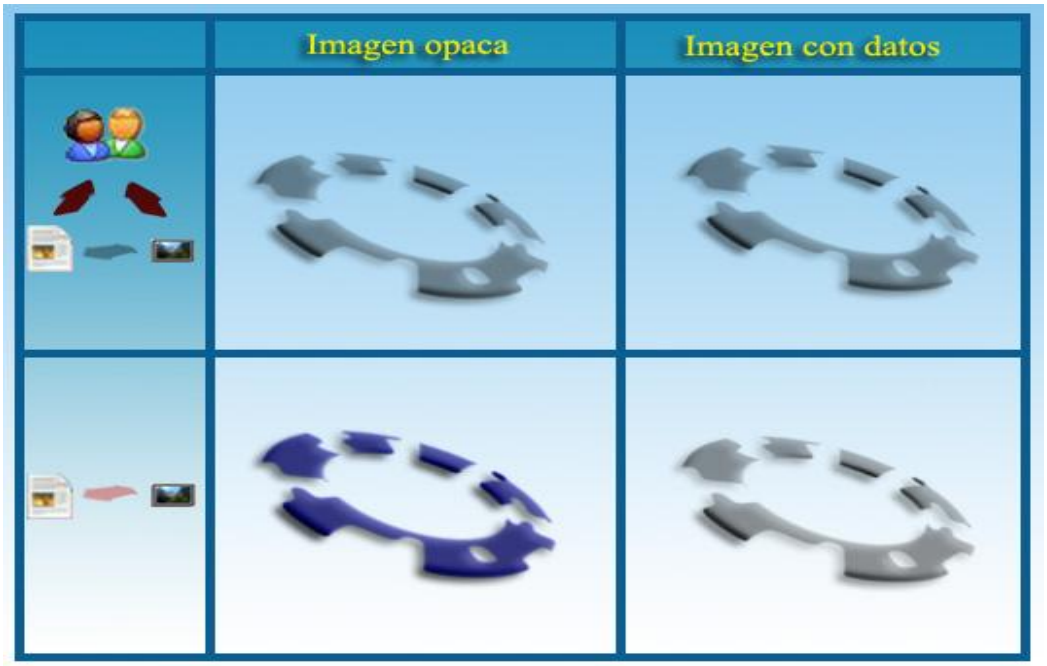

Figura 3: Propuesta de solución.

Ya fijados en este campo, se procede a explicar por pasos cómo dar solución al problema propuesto. Figura 4.

Este es un proceso realizado en línea que tiene como objetivo fundamental recomendar imágenes teniendo como entrada documentos nuevos (noticias). Este proceso se divide en tres tareas fundamentales: Recuperación de información (Fase de representación de cada documento en vectores a partir de los pesos asociados a cada palabra), Cálculo de Similitud (basada en cosenos y puntos de correlación) y Recomendación (Fase de recomendar la imagen más afín al documento analizado).

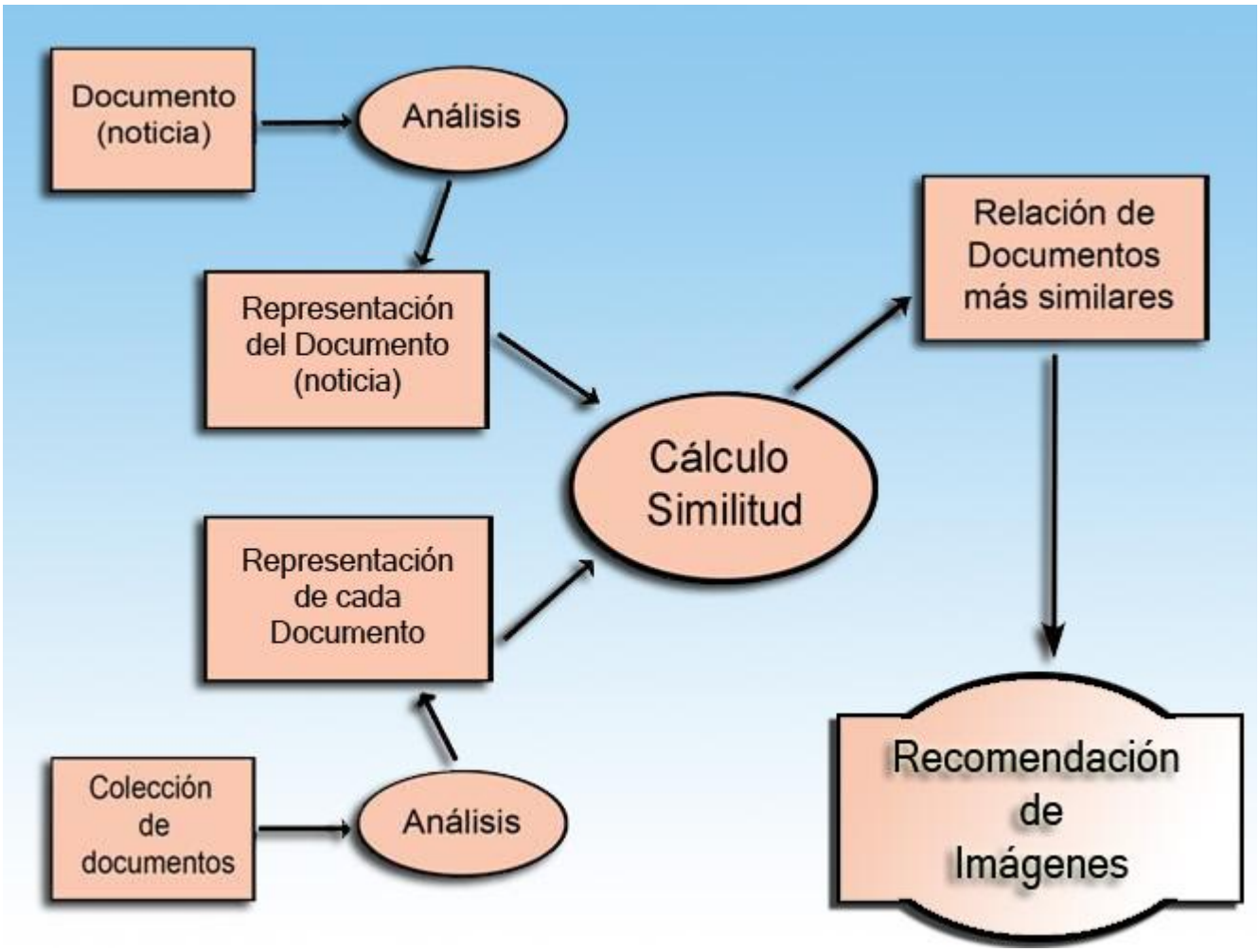

Figura 4: Proceso para la recomendación de imágenes.

Análisis y representación de un documento.

El indizado indica la actividad de cómo hacer el mapeo de un texto en una forma reducida de su contenido. La representación más frecuentemente empleada para representar esto es a partir de un vector con términos, aplicado en el modelo del espacio vectorial, donde más adelante se explica su funcionamiento. En este proceso se reduce la dimensión del vector haciendo un filtrado de palabras respecto a una lista de palabras vacías (palabras frecuentes que no transmiten información ej.: contracciones, conjunciones, preposiciones, artículos, etc.). Luego se obtiene cada una de las frecuencias que dan una visión más clara de la importancia de cada término para el texto y para la colección de documentos ya almacenada en la base de datos, de ahí el peso asociado a cada palabra.

Con respecto al peso se tienen distintas formas de calcularlo, entre las más aplicadas en la comunidad científica se tienen el ponderado Booleano, ponderado por frecuencia de término y el producto de la frecuencia de término y la frecuencia inversa, esta última es empleada en este trabajo de diploma para cuantificar la relevancia de cada palabra o término.

- *Ponderado Booleano*: Consiste en asignar el peso de 1 si la palabra ocurre en el documento y 0 en otro caso
- *Ponderado por frecuencia de término*: Es asignar el número de veces que un término ocurre en un documento.
- *Ponderado tf * idf:* Es el producto de la importancia de frecuencia de cada término en el documento por la importancia de cada término para la colección de documentos.

2.2.1 Indización de textos o documentos.

La indización de textos o documentos es el empleo eficiente de la información contenida en un texto T que requiere la disponibilidad de una representación compacta de tal información, simplemente llamada representación o indización de T. (16)

Dada la comparación de los modelos clásicos en el Capítulo1, se llegó a la conclusión de utilizar el modelo del espacio vectorial. Se elaborará una descripción del mismo que define el espacio de las representaciones de documentos como un espacio vectorial.

Veamos una pequeña síntesis de su funcionamiento:

En primer lugar se seleccionan por cada documento entrante los vocablos mejor afines con su semántica o significado. El conjunto unión de los términos seleccionados por cada uno de estos textos se denomina conjunto de términos de indización.

No todos los elementos de términos son igualmente relevantes para describir el contenido de un texto. Este hecho es capturado con la asignación de un peso numérico a cada término de indización en la representación de un documento, cuantificándose dicha relevancia.

Es indispensable señalar que en este modelo de indización que se analizará se asume que ya se han llevado a cabo los requeridos procesos de filtrado de los textos, en particular la segmentación.

Segmentación del texto: Que transforma las zonas en segmentos apropiados, usualmente en oraciones. (17)

Filtrado del texto: Encargado de seleccionar los segmentos relevantes y de eliminar información irrelevante como ciertas etiquetas o marcas de formato. (17)

2.2.2 Modelo Vectorial (MV).

Este modelo es muy popular en la *Recuperación de Información.*

Se basa en la construcción de una matriz o tabla de términos y documentos, en la cual las filas son estos documentos y las columnas los términos incluidos en ellos. Siendo las filas de esta matriz equivalentes a los documentos que se expresarían en función de las frecuencias de cada término. De esta forma un documento podría expresarse de la manera d1= (1, 2, 0, 0,0,... ... ..., 1, 3) siendo esto el número de veces que aparece cada término en el documento. La longitud del vector de documentos sería igual al total de términos de la matriz (el número de columnas).

La segunda idea asociada a este modelo es calcular la similitud entre la pregunta (convirtiéndose en el vector pregunta, expresado en función de los n términos en la expresión de búsqueda) y los m vectores de documentos almacenados. Siendo los más similares aquellos que se coloquen en los primeros lugares de las respuestas.

Para calcular esta similitud disponemos de varias fórmulas.

La más conocida es la Función del Coseno: calcular el producto escalar de dos vectores de documentos (A y B) y dividirlo por la raíz cuadrada del sumatorio de los componentes del vector A multiplicada por la raíz cuadrada del sumatorio de los componentes del vector B.

Estas similitudes se explican explícitamente más adelante.

Términos seleccionados de un documento.

Son las palabras la fuente de obtención de los términos, motivo por el cual tiene lugar el siguiente pre procesamiento del texto en cuestión:

- Análisis lexicográfico: Permite la obtención de las palabras del texto. Es preciso por tanto el tratamiento de los símbolos alfanuméricos, los guiones y los signos de puntuación:

  Dígitos: Las palabras que contengan secuencias de dígitos son eliminadas.

  Guiones: Constituye un delimitador de palabras.
  Signos de puntuación: Son eliminados de la secuencia de símbolos.
  Letras: Componen las palabras. Todas son llevadas a minúscula.

- Eliminación de términos superfluos (*stopwords*): Las palabras de uso común son eliminadas ya que presumiblemente éstas tendrán elevadas frecuencias de ocurrencia en todos los textos, lo cual, bajo el Enfoque Vectorial, las hace de poca utilidad.

- Extracción de la raíz de las palabras: Permite desechar posibles variaciones sintácticas tomando en consideración en lugar de una palabra la raíz de la misma, o sea, aquella porción que resulta de eliminar sus afijos (prefijos y sufijos).

Un término resultante del pre-procesamiento anterior constituye un término seleccionado del texto en cuestión si su número de ocurrencias en el texto es mayor que F Є N.

Asignación de pesos.

En analogía con técnicas de agrupamiento (*clustering*), se tienen en cuenta dos aspectos:

- Similitud intra-cluster: Valoración de la importancia de un término en el texto bajo análisis con respecto a otros términos del propio texto. (16)

- Disimilitud inter-clúster: Valoración de la importancia de un término para distinguir al texto bajo análisis del resto de los documentos entrantes. (16)

La clave del modelo está en lograr un balance entre estos dos aspectos. Veamos una forma de lograr este balance, enmarcada dentro de los llamados esquemas tf-idf. Sea ni el número de documentos entrantes

en los que ti es un término seleccionado. La frecuencia normalizada de ti en el documento dj está dada por:

$$Tfi, j = \{freqi, jmaxk\ freqk, j\} \quad \text{si ti es un término seleccionado de dj}$$

Donde freqi, j es la frecuencia de ocurrencia de ti en dj y el máximo es computado sobre todos los términos seleccionados de dj. Por otro lado, la frecuencia inversa de ti en los documentos entrantes está dada por: (16)

$$idf_i = \log \left| D \right| / \left| \{d_j : t_i \in d_j\} \right|$$

Finalmente, el peso de t_i en d_j viene dado por:

$$p_{i,j} = tf_{i,j} \times idf_i$$

Como es posible apreciar, el factor tf, representando la similitud intra-cluster, está basado en la frecuencia del término dentro del texto. Por su parte, el factor idf, en representación de la disimilitud inter-cluster, se basa en el hecho de que un término no es distintivo si aparece en muchos documentos entrantes.

2.2.3 Medidas de similitud o distancia.

La representación gráfica de los datos está basada en distancias (similitudes) y algoritmos que permiten dividir los datos en grupos. Son (en general) medidas subjetivas del parecido entre elementos de una base de datos compleja. Para agrupar objetos se utiliza algún tipo de distancia y para agrupar variables se utilizan coeficientes de correlación o medidas similares de asociación. (18)

Tipos:

- **Correlación:** Se traslada el concepto tradicional de covariación, de conexión entre variables, de "pautas" de transición (por ejemplo, el cálculo de un coeficiente de correlación) aplicándolo a las observaciones de los documentos como si fuesen observaciones de variables. (19)

- **Medidas de similitud/Distancia:** Definen proximidad, no Covariación, y su elección (tipos) viene determinada por la escala de medida de las variables: binaria u ordinal o de intervalo/razón. (19)

  - ✓ Medidas de distancia: para escalas ordinales, de intervalo o razón; amplia variedad

    1- Distancia Euclidiana

$$d(i,j)=\sqrt{\left|x_{i1}-x_{j1}\right|^2+\left|x_{i1}-x_{j1}\right|^2+\cdots+\left|x_{in}-x_{jn}\right|^2}$$

Problemas con las unidades de medida**:** normalización previa de variables recomendable (19)

2- Distancia Manhattan

$$d(i,j)=\left|x_{i1}-x_{j1}\right|+\left|x_{i1}-x_{j1}\right|+\cdots+\left|x_{in}-x_{jn}\right|$$

Problemas con la colinealidad (19)

- ✓ Medidas de similitud para variables nominales binarias: reciben el nombre de medidas de asociación. (20)

  1- *Coeficiente de Dice*

  2- *Coeficiente de Jaccard*

  3- *Medida Cosenoidal*

Similitud basada en el coseno.

Esta similitud da una buena medida del "parecido" de dos vectores en un espacio multidimensional, el espacio puede describir características de usuarios o de ítems, tales como palabras claves. La similitud entre ítems es medida computando el coseno entre el ángulo de estos dos vectores. Es la más utilizada por los desarrolladores.

La Función del Coseno equivale a calcular el producto escalar de dos vectores de documentos (A y B) y dividirlo por la raíz cuadrada del sumatorio de los componentes del vector A multiplicada por la raíz cuadrada del sumatorio de los componentes del vector B.

De esta manera se calcula este valor de similitud. Como es obvio, si no hay coincidencia alguna entre los componentes, la similitud de los vectores será cero ya que el producto escalar será cero (circunstancia muy frecuente en la realidad ya que los vectores llegan a tener miles de componentes y se da el caso de la no coincidencia con mayor frecuencia de lo que cabría pensar).

También es lógico deducir que la similitud máxima sólo se da cuando todos los componentes de los vectores son iguales, en este caso la función del coseno obtiene su máximo valor, la unidad. Lo normal es que los términos de las columnas de la matriz hayan sido filtrados (supresión de palabras vacías).

En la figura 5 se considera cada elemento como un vector dentro de un espacio vectorial de m dimensiones y se calcula la similitud con el coseno del ángulo que forman. Es decir dos vectores x1, x2 serán más similares mientras más pequeño sea el ángulo que se forma entre ellos y viceversa.

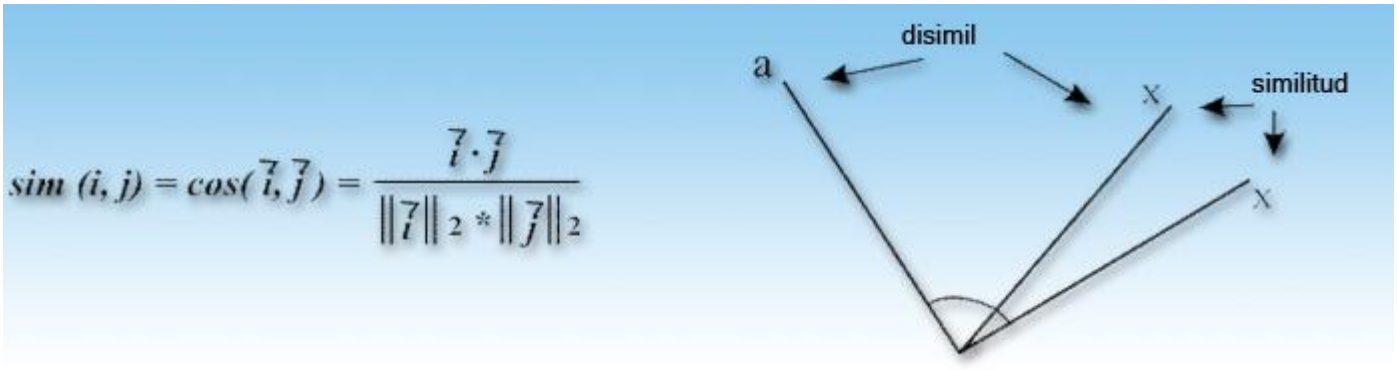

Figura 5: Similitud basada en el Coseno

Similitud basada en Correlación de Pearson.

Para medir la similitud entre variables se suelen emplear el coeficiente de correlación. Este es un índice estadístico que mide la relación lineal entre dos variables cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

Es una medida de que tan bien dos conjuntos de datos caben en una línea recta. Esta fórmula es una de las más complicadas que tiende a dar mejores resultados cuando los datos no están bien normalizados.

En la figura 6 se muestra una línea recta llamada línea de mejor ajuste, nos da una mejor visión de la cercanía de dos grupos de datos, en nuestro caso se toman los términos comunes de cada documento y se calcula el coeficiente para analizar que tan relacionados están los documentos.

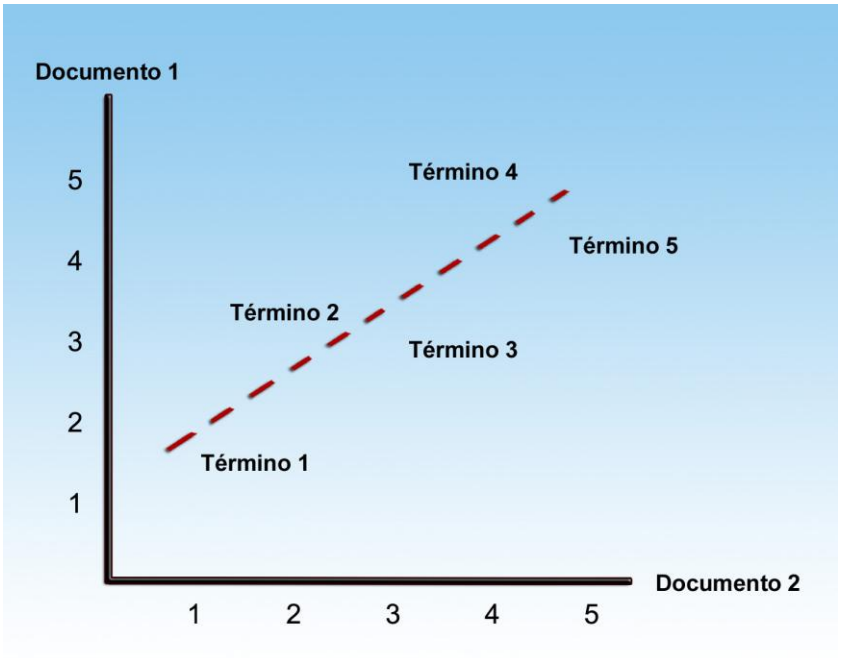

Figura 6: Similitud de documentos por Pearson.

El cálculo del coeficiente de correlación lineal se realiza dividiendo la covarianza por el producto de las desviaciones estándar de ambas variables: (21)

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y},$$

Siendo:

cov (X, Y): la covarianza de (*X*, *Y*)

σ_X y σ_Y: las desviaciones típicas de las distribuciones marginales.

Finalmente queda:

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\,\sqrt{E(Y^2) - E^2(Y)}}.$$

El valor del índice de correlación varía en el intervalo [-1, +1]:

- Si r = 0, no existe relación lineal. Pero esto no necesariamente implica una independencia total entre las dos variables, es decir, que la variación de una de ellas puede influir en el valor que pueda tomar la otra. Pudiendo haber relaciones no lineales entre las dos variables. Estas pueden calcularse con la razón de correlación.

- Si r = 1, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en idéntica proporción.

- Si 0 < r < 1, existe una correlación positiva.

- Si r = -1, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en idéntica proporción.

- Si -1 < r < 0, existe una correlación negativa.

2.2.4 Recomendación de imágenes.

Para recomendar imágenes se hace con Sistemas de recomendación, utilizando algoritmos basados en filtrado colaborativo, este método tiene algunas limitaciones que fueron antes mencionadas en el capítulo 1, pero en esta tesis se trabaja con él puesto que si es posible hacer recomendaciones porque los elementos a analizar están bien configurados y lo suficientemente completos para encontrarle su grupo de documentos más similares en la colección. Además estos sistemas tienden a ofrecer resultados pobres cuando se dispone de poca información o las preferencias a considerar son muy heterogéneas y en este caso ocurre lo contrario, se está trabajando con grandes cantidades de documentos almacenados, lo que hace que sus limitaciones no provoquen afectaciones en los algoritmos que se han implementado.

En la figura 7 se representa la recomendación de imágenes basadas en filtrado colaborativo ya que este proceso recomienda basándose en características similares de los datos que se tienen almacenados y no se tiene en cuenta las características del analizado solamente, en este caso se tiene una colección de documentos con sus respectivas imágenes, cuando aparece un texto nuevo se es necesario recomendar imágenes para el antes mencionado, se busca los textos más similares a este nuevo documento y se devuelve una lista de imágenes afines al documento.

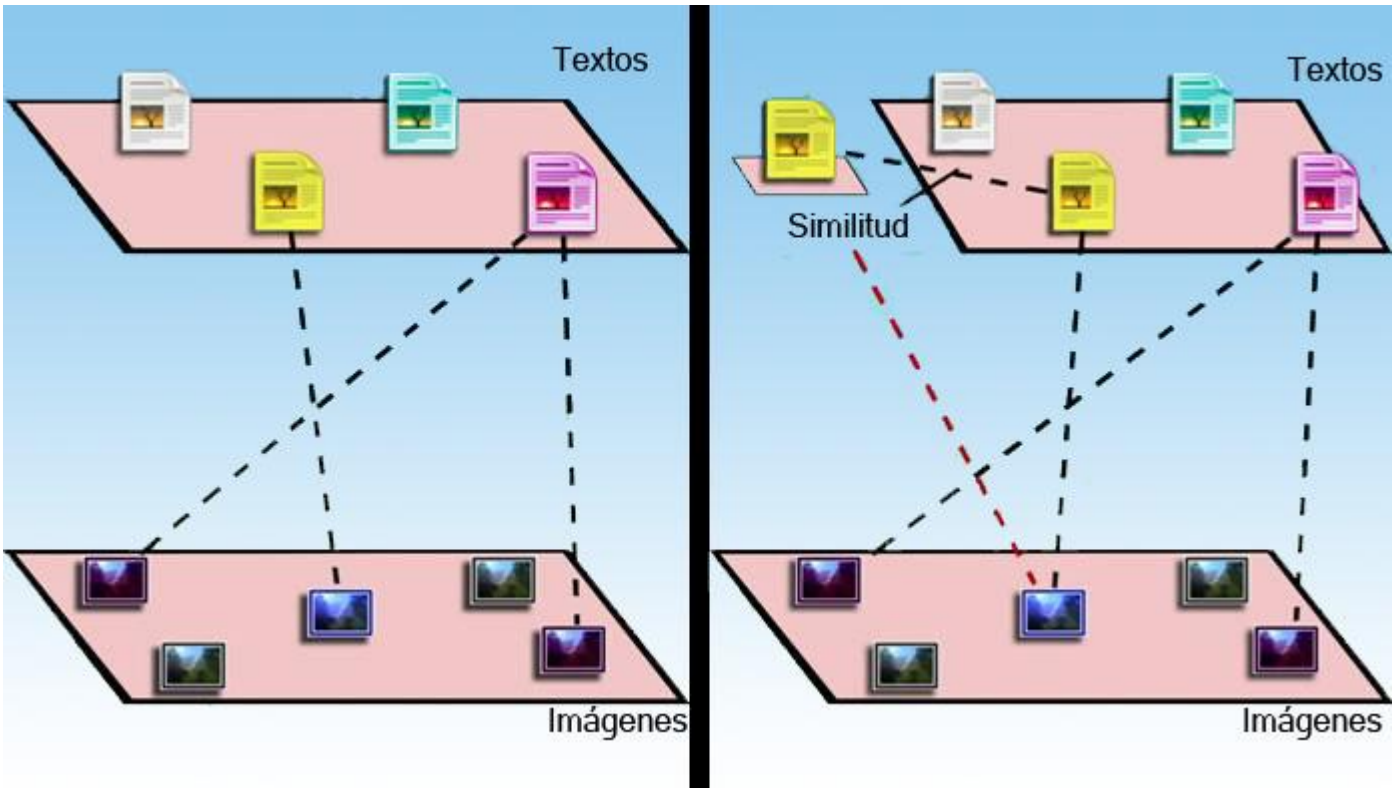

Figura 7: Recomendación de imagen basada en filtrado colaborativo.

2.3 Implementación.

Para la implementación de este trabajo se tienen los ficheros cosine.py y pearson.py con sus clases *CosineRecommendationSystem* y *PearsonRecommendationSystem* respectivamente, las cuales son las contenedoras de los métodos para el cálculo de similitud, distributed.py contiene la clase *DistributedRecommendationSystem* para la implementación del algoritmo en paralelo, utilizando algunos módulos de Pyxel como es el Pyxel.launch permitiendo el trabajo con distintos nodos (unidades de procesamiento). También se cuenta con el fichero *_common.py*, el cual contiene todos los métodos comunes entre todos los ficheros para evitar la reutilización del código, en la clase *AbstracRecommendationSystem*.

2.3.1 Trabajo con Xapian.

Xapian es una librería, a pesar que cuenta con el modelo probabilístico, resulta de mucha ayuda y confiabilidad para este trabajo. Para la recomendación de imágenes que se realizan estos algoritmos es utilizado el metodo vectorial, es por eso que se realizan algunos procesos al Xapian para el cálculo del Peso (TF * IDF) a cada término. Del Xapian se utlizan varias clases:

- Xapian::Documento (): encargada del procesamiento del documento a vector de tipo Documento().

- Xapian::Database (): para el trabajo con la base de datos.

2.3.2 Trabajo con Pyro.

Para el desarrollo de este trabajo, se utiliza Pyro, el cual es una abreviatura para Python de objetos a distancia. Se trata de un avanzado y potente sistema de tecnología de objetos distribuidos escritos totalmente en Python, que está diseñado para ser muy fácil de usar. No hay que preocuparse acerca de cómo escribir código de red de comunicación. Con sólo unas pocas líneas de código adicional, Pyro se encarga de la red de comunicación entre sus objetos una vez que ellos dividido en diferentes máquinas de la red. Todos los detalles de la programación de los sockets son atendidos. Se puede llamar a un método de un objeto remoto como si fuera un objeto.

Pyro proporciona una forma orientada a objetos de RPC (Llamada a Procedimiento Remoto. Para aquellos que estén familiarizados con Java, Java Pyro asemeja al Método de Invocación Remoto (RMI).

2.3.2 Diagrama de componentes.

Muestra las dependencias del compilador y del runtime entre los componentes del compilador del software: por ejemplo los archivos del código fuente y las bibliotecas

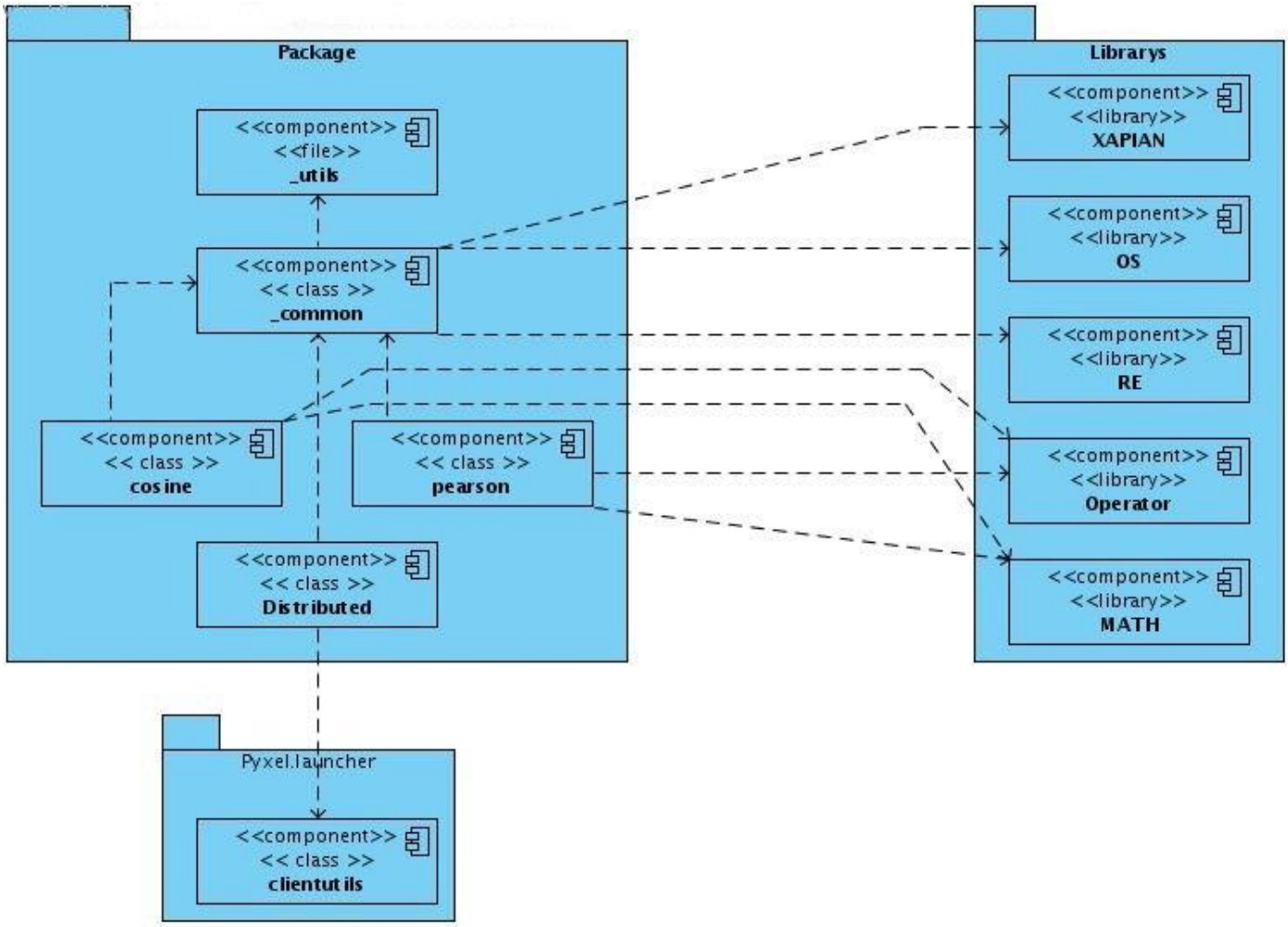

Figura 8: Diagrama de componentes

2.4 Pruebas realizadas.

La comunidad internacional, interesada en la optimización de los sistemas de recuperación de la información (RI), ha desarrollado a lo largo de los años una serie de metodologías extremadamente exhaustivas para la completa evaluación de la recuperación de documentos en dichos sistemas.

El objetivo de cualquier sistema de recuperación de la información es proporcionar a pedido, con máxima utilidad (Recall) y precisión y con un mínimo gasto, información pertinente, en respuesta a cualquier pregunta. (Ver Anexo 2)

Para verificar el correcto funcionamiento de los algoritmos y la probabilidad de acierto se realizó pruebas teniendo en cuenta las métricas explicadas anteriormente a una colección de 100 documentos. (Ver Anexo 3)

Se tendrán en cuenta las siguientes variables:

D, conjunto de documentos

RC, conjunto de documentos relevantes

$\mathbf{R_N}$ = **D − RC**, conjunto de documentos no relevantes

TR, conjunto de documentos recuperados

RR, conjunto de documentos relevantes recuperados

1- *Prueba para el documento #1*: Este documento es una noticia que enuncia una reflexión de nuestro Comandante Fidel Castro Ruz:

Precisión: **RR/TR** Recall: **RR/RC**

Algoritmo	Tiempo	RC	RR	TR	Precisión	Recall
Coseno	2min-40s	1	1	1	1	1
Pearson	2min-45s	1	1	2	0.5	1

2- *Prueba para el documento # 19*: Este documento es una noticia que habla del tanto del argentino Messi y que una primera parte excepcional bastó al conjunto de Barcelona para llevarse el triunfo esperado por todos:

Precisión: **RR/TR** Recall: **RR/RC**

Algoritmo	Tiempo	RC	RR	TR	Precisión	Recall
Coseno	14s	3	3	3	1	1
Pearson	13s	3	3	26	0.12	1

3- *Prueba para el documento # 38:* Este documento es una noticia que refleja al mundo la celebración en Cuba del Primero de Mayo con un desfile multitudinario que presidió el presidente, Raúl Castro, en la Plaza de la Revolución de La Habana, y en el que se reiteró el compromiso de nuestra isla con el socialismo en el año en que se conmemora el 50 aniversario del triunfo de la revolución:

Precisión: **RR/TR** Recall: **RR/RC**

Algoritmo	Tiempo	RC	RR	TR	Precisión	Recall
Coseno	42s	8	1	1	1	0.125
Pearson	41s	8	2	5	0.4	0.625

4- *Prueba para el documento#70:* Este documento es una noticia que habla del esfuerzo que realiza el Rey de España animando a Estonia a responder a la crisis financiera con proyectos económicos hispano-estonios en sectores de vanguardia como las tecnologías de la información y la comunicación o las energías renovables, donde ambos países disponen de "empresas punteras a escala mundial":

Precisión: **RR/TR** Recall: **RR/RC**

Algoritmo	Tiempo	RC	RR	TR	Precisión	Recall
Coseno	40s	1	1	1	1	1
Pearson	40s	1	1	18	0.06	1

Para todas las pruebas se alcanzó el resultado esperado con un rango de estimación de tiempo aceptable a la colección tratada, que, visiblemente se puede inferir que en cada uno de los casos para el método del coseno la precisión es igual o mayor que el recall y en el método de Pearson el recall es mayor que la precisión. Se puede concluir que ambos métodos analizando su funcionamiento devuelven sus resultados pertinentes, pero el método del coseno es mucho más preciso para este trabajo en la prensa que el Pearson porque todos sus resultados de precisión superan el 0.8%.

Con respecto al algoritmo distribuido, las pruebas realizadas arrojaron los mismos resultados lo que ganando en tiempo por el trabajo con diferentes nodos y de forma sincronizada, a medida que haya más nodos más rápido se obtienen las recomendaciones. (Ver Anexo 3)

Conclusiones

Con la culminación del presente trabajo de diploma se cumple con el objetivo trazado en el mismo,

- Se seleccionaron los algoritmos que han demostrado ser más efectivos.
- Se definió el procesamiento de los textos para extraer la información relevante.
- Se codificaron y ensamblaron las partes para llegar a un resultado final
- Se evaluaron experimentalmente los resultados en velocidad de procesamiento, recursos necesarios y pertinencia de la recomendación

Recomendaciones

- Se recomienda capacitar a los miembros del proyecto de Informatización de la Prensa con el objetivo de seguir mejorando la rapidez y eficiencia en el desarrollo de futuros portales web utilizando este servicio.
- Vincular esta herramienta a Ecumene para que pueda ser empleada por los medios de prensa.

Trabajos citados

1. **Documentación, la Sección Científica de Ciencias.** hipertext.net. *Sistemas de Recomendación Semánticos. Un análisis del estado de la cuestión.* [En línea] [Citado el: 20 de octubre de 2008.] http://www.hipertext.net/web/pag286.htm. 1.

2. **Castillo, José Manuel Morales del.** *Modelo de servicio semántico-difuso de difusión selectiva de información.* [En línea] [Citado el: 17 de noviembre de 2008.] http://hera.ugr.es/tesisugr/17620727.pdf. 2.

3. **Fernández Ramírez, M.L.** *Ágora: Creación de grupos virtuales en bibliotecas digitales.* [En línea] [Citado el: 19 de noviembre de 2008.] http://ict.udlap.mx/people/lulu/documento/capitulo4.html. 3.

4. **Carlos G. Figuerola, José Luis Alonso Berrocal, Ángel Francisco Zazo Rodríguez.** *Diseño de un motor de recuperación de la información para uso experimental y educativo.* [En línea] [Citado el: 07 de diciembre de 2008.] http://www.ub.edu/bid/04figue2.htm. 4.

5. **CARMEN, ZARCO FERNÁNDEZ.** *APLICACIÓN DE LOS ALGORITMOS EVOLUTIVOS AL APRENDIZAJE AUTOMÁTICO DE CONSULTAS BOOLEANAS EXTENDIDAS PARA SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN DIFUSOS.* [En línea] 2002. [Citado el: 07 de enero de 2009.] http://www.cibernetia.com/tesis_es/LINGUISTICA/LINGUISTICA_APLICADA/DOCUMENTACION_AUTOMATIZADA/1. 5.

6. *La recuperación y los sistemas de Recuperación de información.* [En línea] [Citado el: 10 de enero de 2009.] http://descargas.cervantesvirtual.com/servlet/SirveObras/02472741989036164198835/010010_3.pdf. 6.

7. **Montero, Yusef Hassan.** *Indización Social y Recuperación de Información.* [En línea] 2006. [Citado el: 20 de enero de 2009.] http://www.nosolousabilidad.com/articulos/indizacion_social.htm. 7.

8. **Sánchez, Iván Bernabé.** *Recuperación y Organización de la información.* [En línea] 25 de abril de 2006. [Citado el: 10 de febrero de 2009.] http://modelos-recuperacion.galeon.com. 8.

9. **Broncano, Rubén García.** *Recuperación y organización de la información.* [En línea] Página realizada para la universidad Carlos III de Madrid , 9 de abril de 2006. [Citado el: 13 de febrero de 2009.] http://modelosrecuperacion.tripod.com/booleano.html. 9.

10. **Rodríguez, María Luque.** *Modelos de Recuperación de la información basados en Información Linguística Difusa y Algoritmos evolutivos.Mejorando la representación de las necesidades de información.* [En línea] [Citado el: 18 de febrero de 2009.] http://hera.ugr.es/tesisugr/15350605.pdf. 10.

11. **Pabloe.** [En línea] 18 de mayo de 2008. [Citado el: 24 de febrero de 2009.] http://www.scribd.com/doc/3013167/Indizacion-y-Busqueda-a-traves-de-Lucene#document_metadata. 11.

12. **Palacios, Francisco.** *Métodos Numéricos: Guía de estudio.* [En línea] Abril de 2008. [Citado el: 28 de febrero de 2009.] www.eupm.upc.edu/~fpq/numerico/orientaciones/2007-08/sist-orien.pdf. 12.

13. **M., G. Figueroa.** Tecnologías de internet en la Enseñanza de la Matemática. *Resolviendo ecuaciones con Excel.* [En línea] [Citado el: 28 de febrero de 2009.] www.cidse.itcr.ac.cr/revistamate/HERRAmInternet/ecuaexecl/node5.html. 13.

14. **anónimo.** MalaCiencia . *El problema de los tres cuerpos .* [En línea] 28 de marzo de 2006. [Citado el: 28 de febrero de 2009.] http://www.malaciencia.info/2006/03/el-problema-de-los-tres-cuerpos.html. 14.

15. —. [En línea] [Citado el: 28 de febrero de 2009.] www.di.ujaen.es/asignaturas/ArqTel/teoria/Tema1.pdf.

16. **González, Carlos Luis Medina.** *AutoClassifier: un sistema de clasificación de textos en lenguaje natural.* Habana : s.n., 2006. 16.

17. **Valero, Alberto Téllez.** *Extracción de Información con Algoritmos de Clasificación.* [En línea] 2005. [Citado el: 15 de marzo de 2009.] http://ccc.inaoep.mx/~mmontesg/tesis%20estudiantes/TesisMaestria-AlbertoTellez.pdf. 17.

18. **anónimo.** [En línea] [Citado el: 20 de marzo de 2009.] http://www.escet.urjc.es/~matemati/TBC/CLUSTERDOC/sld034.htm. 18.

19. —. [En línea] [Citado el: 28 de abril de 2009.] http://www.uam.es/departamentos/economicas/econapli/fse03/cluster.doc. 19.

20. **Manuel Montes y Gómez, Luis Villaseñor Pineda.** *Agrupamiento de Documentos.* [En línea] [Citado el: 25 de marzo de 2009.] http://ccc.inaoep.mx/~mmontesg/cursos/Tratamiento%20Automatico%20de%20Textos/Curso-TAT1-DC.ppt. 20.

21. **anónimo.** *COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON.* [En línea] [Citado el: 02 de abril de 2009.] http://personal.us.es/vararey/adatos2/correlacion.pdf. 21.

22. **Julieth, Mayerin.** *¿Qué es la fotografía? .* [En línea] 28 de abril de 2008. [Citado el: 20 de abril de 2009.] http://www.wikilearning.com/apuntes/la_fotografia-que_es_la_fotografia/11077-1. 22.

23. **anónimo.** [En línea] 26 de abril de 2009. [Citado el: 23 de abril de 2009.] http://es.wikipedia.org/wiki/An%C3%A1lisis_de_im%C3%A1genes. 23.

24. **Briones, Jose L.** *¿Que son los metadatos de una fotografía?* [En línea] 1 de abril de 2008 . [Citado el: 02 de mayo de 2009.] http://www.theblog.es/mt/archives/2008/04/que_son_los_metadatos_de_una_f.php. 24.

25. **Santos, Sergio de los.** *Ciertos metadatos en fotografías vuelven a revelar información sensible.* [En línea] 19 de junio de 2006. [Citado el: 02 de mayo de 2009.] http://www.hispasec.com/unaaldia/2795. 25.

Bibliografía

La recuperación y los sistemas de Recuperación de información. [En línea] [Citado el: 10 de enero de 2009.] http://descargas.cervantesvirtual.com/servlet/SirveObras/02472741989036164198835/010010_3.pdf. 6.

Abadal, E. Codina, L. 2005. Recuperación de Información. *En: Bases de Datos Documentales: Características, funciones y método. Capítulo 2. p. 29-92.* [En línea] 2005. [Citado el: 23 de enero de 2009.] http://www.lluiscodina.com/riv2.doc.

anónimo. [En línea] [Citado el: 28 de febrero de 2009.] www.di.ujaen.es/asignaturas/ArqTel/teoria/Tema1.pdf. 15.

—. [En línea] [Citado el: 20 de marzo de 2009.] http://www.escet.urjc.es/~matemati/TBC/CLUSTERDOC/sld034.htm. 18.

—. [En línea] [Citado el: 28 de abril de 2009.] http://www.uam.es/departamentos/economicas/econapli/fse03/cluster.doc. 19.

—. *COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON.* [En línea] [Citado el: 02 de abril de 2009.] http://personal.us.es/vararey/adatos2/correlacion.pdf. 21.

—. 2009. [En línea] 26 de abril de 2009. [Citado el: 23 de abril de 2009.] http://es.wikipedia.org/wiki/An%C3%A1lisis_de_im%C3%A1genes. 23.

—. 2006. MalaCiencia . *El problema de los tres cuerpos .* [En línea] 28 de marzo de 2006. [Citado el: 28 de febrero de 2009.] http://www.malaciencia.info/2006/03/el-problema-de-los-tres-cuerpos.html. 14.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. *Item-based Collaborative Filtering Recommendation.* [En línea] 1-5 de May de 2001. [Citado el: 15 de february de 2009.] http://citeseer.ist.psu.edu/sarwar01itembased.html.

Baeza-Yates, R. Ribeiro-Neto, B. 1999. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Information Processing and Management. *Modern Information Retrieval.* 1999.

Briones, Jose L. 2008 . *¿Que son los metadatos de una fotografía?* [En línea] 1 de abril de 2008 . [Citado el: 02 de mayo de 2009.] http://www.theblog.es/mt/archives/2008/04/que_son_los_metadatos_de_una_f.php. 24.

Broncano, Rubén García. 2006. *Recuperación y organización de la información.* [En línea] Página realizada para la universidad Carlos III de Madrid , 9 de abril de 2006. [Citado el: 13 de febrero de 2009.] http://modelosrecuperacion.tripod.com/booleano.html. 9.

Carlos G. Figuerola, José Luis Alonso Berrocal, Ángel Francisco Zazo Rodríguez. *Diseño de un motor de recuperación de la información para uso experimental y educativo.* [En línea] [Citado el: 07 de diciembre de 2008.] http://www.ub.edu/bid/04figue2.htm. 4.

CARMEN, ZARCO FERNÁNDEZ. 2002. *APLICACIÓN DE LOS ALGORITMOS EVOLUTIVOS AL APRENDIZAJE AUTOMÁTICO DE CONSULTAS BOOLEANAS EXTENDIDAS PARA SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN DIFUSOS.* [En línea] 2002. [Citado el: 07 de enero de 2009.] http://www.cibernetia.com/tesis_es/LINGUISTICA/LINGUISTICA_APLICADA/DOCUMENTACION_AUTOMATIZADA/1. 5.

Castillo, José Manuel Morales del. *Modelo de servicio semántico-difuso de difusión selectiva de información.* [En línea] [Citado el: 17 de noviembre de 2008.] http://hera.ugr.es/tesisugr/17620727.pdf. 2.

Documentación, la Sección Científica de Ciencias. hipertext.net. *Sistemas de Recomendación Semánticos. Un análisis del estado de la cuestión.* [En línea] [Citado el: 20 de octubre de 2008.] http://www.hipertext.net/web/pag286.htm. 1.

Fernández Ramírez, M.L. *Ágora: Creación de grupos virtuales en bibliotecas digitales.* [En línea] [Citado el: 19 de noviembre de 2008.] http://ict.udlap.mx/people/lulu/documento/capitulo4.html. 3.

González, Carlos Luis Medina. 2006. *AutoClassifier: un sistema de clasificación de textos en lenguaje natural.* Habana : s.n., 2006. 16.

He, Ming. 2007. *Feature Selection Based on Neighborhood Systems and Rough Set Theory.* [En línea] 2007. [Citado el: 14 de marzo de 2009.] http://portal.acm.org/citation.cfm?id=1136439.

Hofmann, Thomas and Puzicha, Jan. 1999. *Latent Class Models for Collaborative Filtering.* [En línea] 1999. [Citado el: 12 de abril de 2009.] http://eprints.kfupm.edu.sa/47879/.

Julieth, Mayerin. 2008. *¿Qué es la fotografía? .* [En línea] 28 de abril de 2008. [Citado el: 20 de abril de 2009.] http://www.wikilearning.com/apuntes/la_fotografia-que_es_la_fotografia/11077-1. 22.

M., G. Figueroa. Tecnologías de internet en la Enseñanza de la Matemática. *Resolviendo ecuaciones con Excel.* [En línea] [Citado el: 28 de febrero de 2009.] www.cidse.itcr.ac.cr/revistamate/HERRAmInternet/ecuaexecl/node5.html. 13.

Manuel Montes y Gómez, Luis Villaseñor Pineda. *Agrupamiento de Documentos.* [En línea] [Citado el: 25 de marzo de 2009.] http://ccc.inaoep.mx/~mmontesg/cursos/Tratamiento%20Automatico%20de%20Textos/Curso-TAT1-DC.ppt. 20.

Montero, Yusef Hassan. 2006. *Indización Social y Recuperación de Información.* [En línea] 2006. [Citado el: 20 de enero de 2009.] http://www.nosolousabilidad.com/articulos/indizacion_social.htm. 7.

Mount, David M. 2009. *Algorithms for Fast Vector Quantization.* [En línea] 19 de febrero de 2009. http://www.cs.umd.edu/~mount/Papers/DCC.pdf.

Pabloe. 2008. [En línea] 18 de mayo de 2008. [Citado el: 24 de febrero de 2009.] http://www.scribd.com/doc/3013167/Indizacion-y-Busqueda-a-traves-de-Lucene#document_metadata. 11.

Palacios, Francisco. 2008. *Métodos Numéricos: Guía de estudio.* [En línea] Abril de 2008. [Citado el: 28 de febrero de 2009.] www.eupm.upc.edu/~fpq/numerico/orientaciones/2007-08/sist-orien.pdf. 12.

Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins. *Recommendation systems: a probabilistic analysis.* [En línea] [Citado el: 12 de marzo de 2009.] http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.3495.

Rodríguez, María Luque. *Modelos de Recuperación de la información basados en Información Linguística Difusa y Algoritmos evolutivos.Mejorando la representación de las necesidades de información.* [En línea] [Citado el: 18 de febrero de 2009.] http://hera.ugr.es/tesisugr/15350605.pdf. 10.

Sánchez, Iván Bernabé. 2006. *Recuperación y Organización de la información.* [En línea] 25 de abril de 2006. [Citado el: 10 de febrero de 2009.] http://modelos-recuperacion.galeon.com. 8.

Santos, Sergio de los. 2006. *Ciertos metadatos en fotografías vuelven a revelar información sensible.* [En línea] 19 de junio de 2006. [Citado el: 02 de mayo de 2009.] http://www.hispasec.com/unaaldia/2795. 25.

Segaran, Toby. august, 2007. Programming. Colective Intelligence. United States of America : s.n., august, 2007.

Tieli Sun, Lijun Wang, Qinghe Guo. 2009. IEEE Xplore. *A Collaborative Filtering Recommendation Algorithm Based on Item Similarity of User Preference.* [En línea] 23-25 de Jan de 2009. [Citado el: 23 de marzo de 2009.] http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4771878.

Valero, Alberto Téllez. 2005. *Extracción de Información con Algoritmos de Clasificación.* [En línea] 2005. [Citado el: 15 de marzo de 2009.] http://ccc.inaoep.mx/~mmontesg/tesis%20estudiantes/TesisMaestria-AlbertoTellez.pdf. 17.

Anexos

1- Metodología Scrum [1]

1. **Ailyn Alfonso, Joel Armada. Trabajo de Diploma.** mayo del 2008. [Citado el: 23 de abril de 2009.]

La metodología *Scrum* define cada proceso como iterativo e incremental. El esqueleto de *Scrum* se muestra en la figura 3. El círculo inferior representa una iteración del desarrollo de las actividades que ocurren una tras otra. El producto de cada iteración es un incremento en el producto. El círculo superior representa la reunión diaria que ocurre durante la iteración, en la cual los miembros individualmente del grupo conocen, inspeccionan las actividades y hacen los cambios apropiados. Como resultado de la iteración queda una lista de requerimientos. Este ciclo se repite a lo largo de todo el proyecto.

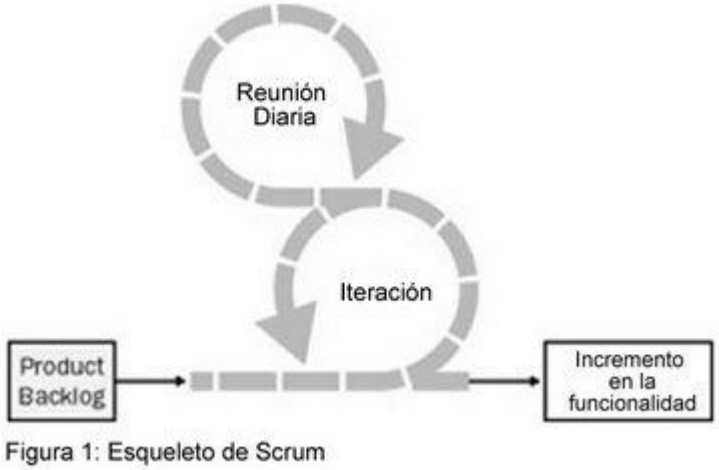

Figura 1: Esqueleto de Scrum

Figura 1.1: Esqueleto de Scrum

Este esqueleto opera de esta manera:

- Al comienzo de la iteración, el equipo revisa que es lo que debe hacer.
- Luego, selecciona lo que cree que puede hacer para tener un incremento y un potencial prototipo funcional al término de la iteración.

- El equipo se separa y hace su mejor esfuerzo por el resto de la iteración. Cuando ésta termina, el equipo presenta el incremento de la funcionalidad que construyó, de manera que los otros miembros del equipo puedan revisar las funcionalidades y hacer las modificaciones necesarias al proyecto.

Los roles en *Scrum*

Hay solo tres roles en *Scrum*:

• Product Owner.

• Team (el equipo).

• Scrum Master.

Todas las responsabilidades de manejo de un proyecto se dividen entre estos tres papeles.

Product Owner

Es el responsable de cuidar los intereses de cada uno de los participantes. El *Product Owner* estima el financiamiento inicial y el requerido en el curso del proyecto mediante la creación de *The project's initial overall requirements* (los requisitos totales e iníciales del proyecto), preocupándose de retornar los objetivos de inversión (ROI), y los planes de revisión.

La lista de requerimientos se llama *Product Backlog*. El *Product Owner* es el responsable de usar el *Product Backlog* para asegurarse que las funcionalidades de mayor valor sean producidas e implementadas. Frecuentemente se prioriza el primer elemento del *Product Backlog* ya que se van ordenando los requisitos de mayor a menor valor.

Team

El equipo de desarrollo tiene la responsabilidad, en cada iteración, de transformar el *Product Backlog* en un incremento funcional para el producto, con vistas a lograr esto deben planificar su propio trabajo. Los miembros del equipo son responsables en conjunto del éxito de cada iteración y del proyecto en su totalidad.

Scrum Master

Es responsable del proceso *Scrum*. Debe enseñar la metodología *Scrum* a cada integrante implicado en el proyecto, preocupándose de poner la metodología en práctica y hacerla parte de la cultura de trabajo de la organización, asegurándose de que cada miembro del equipo siga las reglas y prácticas de *Scrum*.

El flujo de *Scrum*

Un proyecto de *Scrum* comienza con una visión del sistema que se irá desarrollando a medida que este avance. El *Product Owner* es responsable de financiar el proyecto y además debe entregar sus ideas de manera que se maximice su ROI. También debe formular un plan que pueda ser incluido en el *Product Backlog*.

El *Product Backlog* es una lista de los requisitos funcionales y no funcionales donde los elementos de mayor valor funcional son priorizados.

Sprint

A las tareas del *Product Backlog* se les va dando cumplimiento trabajando en *Sprints*. Cada Sprint es una iteración de 30 días consecutivos (esto puede adaptarse a las necesidades del proyecto.).

Se inicia cada *Sprint* con un *Sprint Planning Meeting (Reunión de planeamiento del Sprint)*, donde el *Product Owner* y el *Team* idean juntos lo que se espera para el siguiente *Sprint*, seleccionando del *Product Backlo*g las tareas de más prioridad. De esta forma *Product Owner* le dice al *Team* qué desea, y el *Team* le dice al *Product Owner* cuánto de lo que quiere se puede transformar en funcionalidad durante el *Sprint*. Las reuniones de planeamiento del *Sprint* no pueden durar más de ocho horas. La meta es conseguir trabajar, no pensar en el trabajo:

- Las primeras cuatro horas se dedican al *Product Owner* que presenta la prioridad más alta del *Product Backlog* al equipo. El Team le pregunta sobre el contenido, el propósito, el significado, y las intenciones del *Product Backlog*. Posteriormente el equipo selecciona del *Product Backlog* lo que cree poder transformar en un incremento funcional para el final del Sprint.
- Durante las segundas cuatro horas el equipo planifica su propio *Sprint*. Las tareas que componen este plan se ponen en una lista llamada *Sprint Backlog*.

Diariamente, el equipo realiza una reunión minuciosa de 15 minutos llamada *Daily Scrum* en la cual cada miembro del equipo contesta a tres preguntas:

1. ¿Qué has terminado desde la última *Daily Scrum*?

2. ¿Qué planeas hacer a favor del proyecto entre esta y la próxima *Daily Scrum*?

3. ¿Tienes algún impedimento que te dificulte cumplir con el Sprint y con el proyecto?

El propósito de la reunión es sincronizar el trabajo de todos los miembros del equipo y programa cualquier reunión que sea necesaria para seguir avanzando.

Al final del Sprint, se realiza una Reunión de Revisión del Sprint. Ésta es una reunión de unas cuatro horas, en la cual el equipo presenta al *Product Owner* qué desarrolló durante el Sprint.

Después de la Reunión de Revisión del Sprint y antes de la Reunión de Planeamiento del próximo Sprint, el *Scrum Master* convoca a una *Sprint Retrospective* (Reunión Retrospectiva del Sprint) con el equipo. En esta reunión de tres horas, el *Scrum Master* hace que el equipo revise su proceso de desarrollo Scrum, tratando de hacerlo más eficaz y agradable para el próximo Sprint.

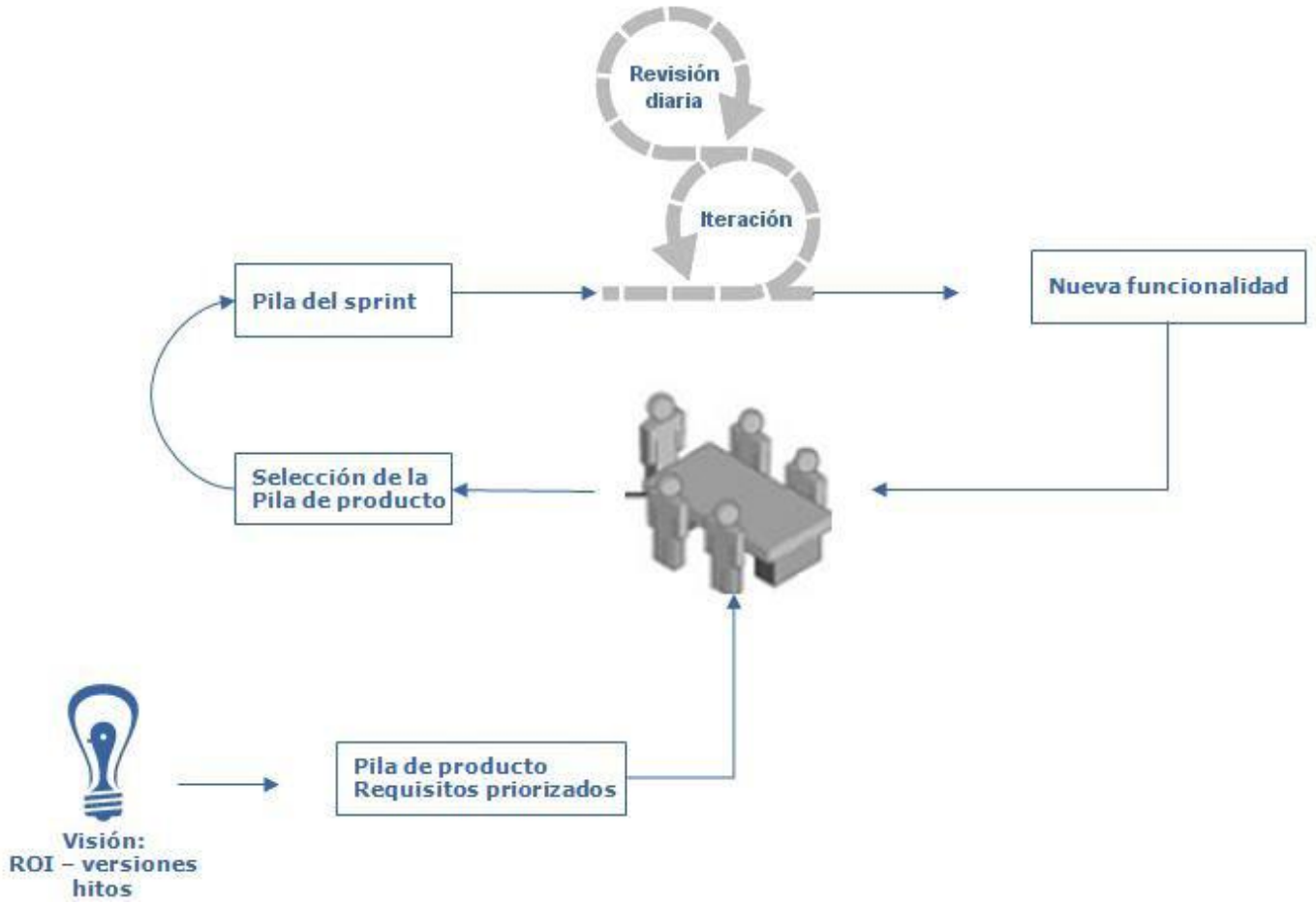

Figura 1.2: Diagrama del proceso Scrum

Artefactos de Scrum.

Product Backlog

Id	Módulo	Descripción	Est.	Por
Crítico				
1		Plataforma tecnológica	30	AR
2	Cliente	Interfaz de usuario	40	LR
3	Cliente	Un usuario se registra en el sistema	40	LR
4	Trastienda	El operador define el flujo y textos de un expediente	60	AR
5	Trastienda	Etc...	999.	XX
Necesario				
6	Cliente	El usuario modifica su ficha personal	30	AR
7	Cliente	El usuario consulta los expedientes asignados	15	LR
8	Cliente	El usuario tramita un expediente	35	LR

Figura 1.3: Product Backlog

Los requisitos del producto a desarrollar se enumeran en el *Product Backlog*. *El Product Owner* es responsable del contenido, priorización, y disponibilidad del *Product Backlog*. Es simplemente una estimación inicial de los requisitos. Se desarrolla paralelamente a medida que el producto y el ambiente en el cual se trabaja evoluciona. Es dinámico. Maneja constantemente los cambios para identificar que necesita el producto para ser: apropiado, competitivo, y útil. Mientras exista un producto, el *Product Backlog* también existe.

Sprint Backlog

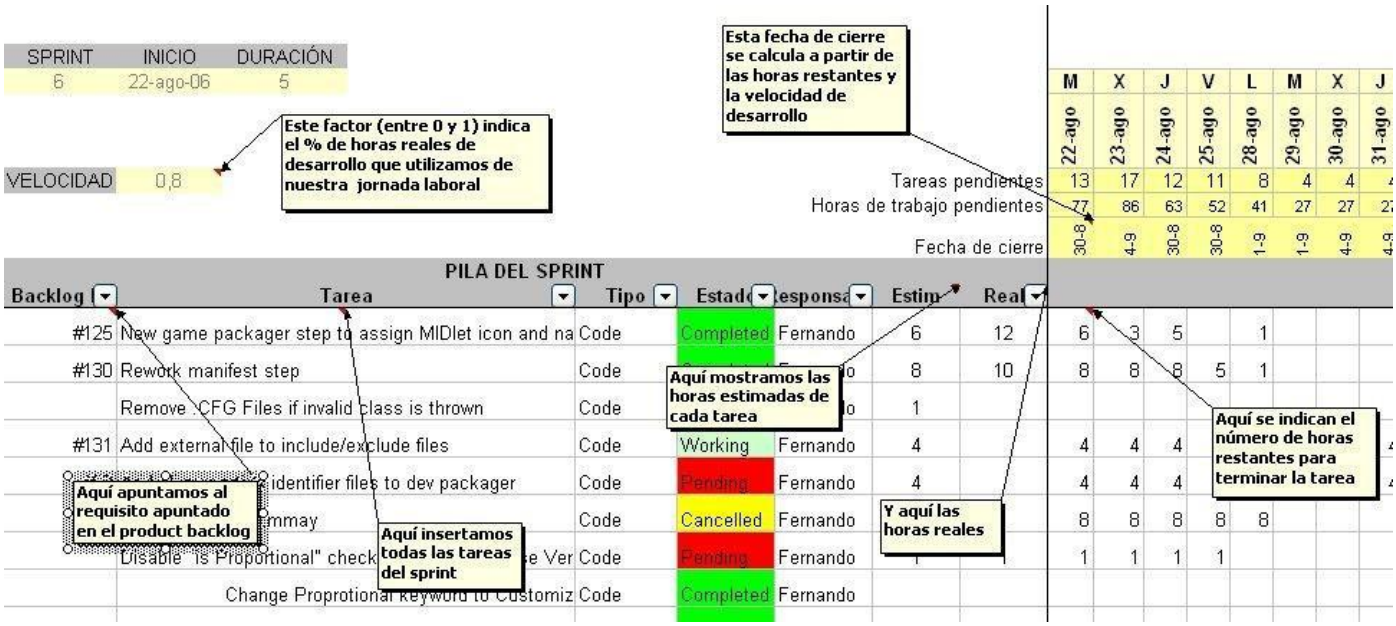

Figura 1.4: Sprint Backlog

El *Sprint Backlog* define las tareas seleccionadas del *Product Backlog* que el equipo desarrollará para lograr incrementos potencialmente funcionales del producto. El equipo crea una lista inicial de estas tareas en la segunda parte de la reunión de planificación del *Sprint*. Las tareas deben ser divididas de modo que cada una demore entre 1 a 16 horas finalizarlas. Las tareas de largo mayor de 16 horas se consideran secundarias, ya que todavía no se han definido apropiadamente. Solamente el equipo puede cambiar el *Sprint Backlog*.

2- Medidas de evaluación.

La comunidad internacional, interesada en la optimización de los sistemas de recuperación de la información (RI), ha desarrollado a lo largo de los años una serie de metodologías extremadamente exhaustivas para la completa evaluación de la recuperación de documentos en dichos sistemas.

El objetivo de cualquier sistema de recuperación de la información es proporcionar a pedido, con máxima utilidad (recall) y precisión y con un mínimo gasto, información pertinente, en respuesta a cualquier pregunta.

Recall.

La capacidad de un sistema de recuperación y organización de la información para proveer de documentos relevantes se mide con una métrica llamada Recall o Exhaustividad, que se define la proporción de material relevante recuperado, del total de los documentos que son relevantes en la colección de documentos, independientemente de que éstos, se recuperen o no.

$$Recall = \frac{Nº\ de\ documentos\ relevantes\ recuperados}{Nº\ de\ documentos\ relevantes\ en\ la\ colección}$$

Figura 2.1: Fórmula de Recall

Precisión.

Por supuesto, cualquier sistema de recuperación de información podría conseguir una *Exhaustividad* del 100% simplemente devolviendo todos los elementos de la colección. Por ello se utiliza también otra métrica, llamada Precisión, que se define como la proporción de material recuperado realmente relevante, del total de los documentos recuperados.

$$Precisión = \frac{Nº\ de\ documentos\ relevantes\ recuperados}{Nº\ total\ de\ documentos\ recuperados}$$

Figura 2.2: Fórmula de Precisión

A continuación podemos observar de manera gráfica un ejemplo de la representación de estas dos métricas:

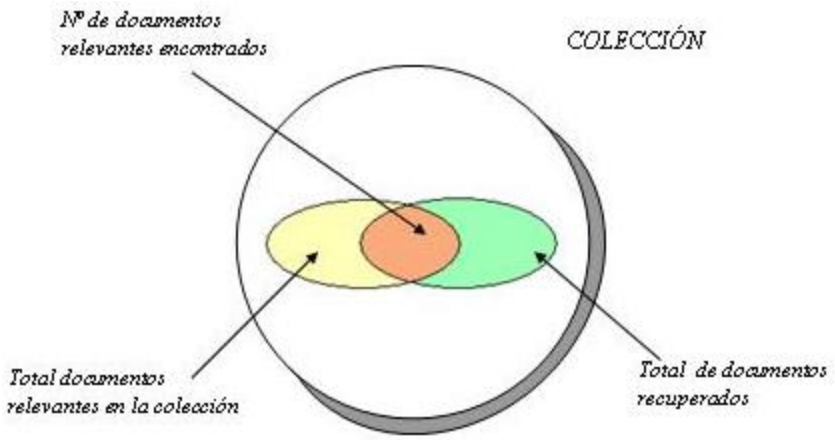

Figura 2.3: Representación de métricas

Muchos sistemas de recuperación y organización de la información no hacen afirmaciones explícitas sobre la relevancia o no de un documento, sino que ordenan la colección de mayor a menor relevancia respecto a una consulta. Para un mismo sistema y una misma consulta, Recall y Precisión son inversamente proporcionales. En efecto, si se recuperan los *n* documentos de mayor relevancia, tendremos una alta Precisión y bajo Recall para valores pequeños de *n,* así como una baja Precisión y alto Recall para valores grandes de *n.* Para reflejar este hecho se utilizan frecuentemente las curvas *Precisión-Recall*, que muestran el valor de precisión para distintos niveles de Exhaustividad o recall:

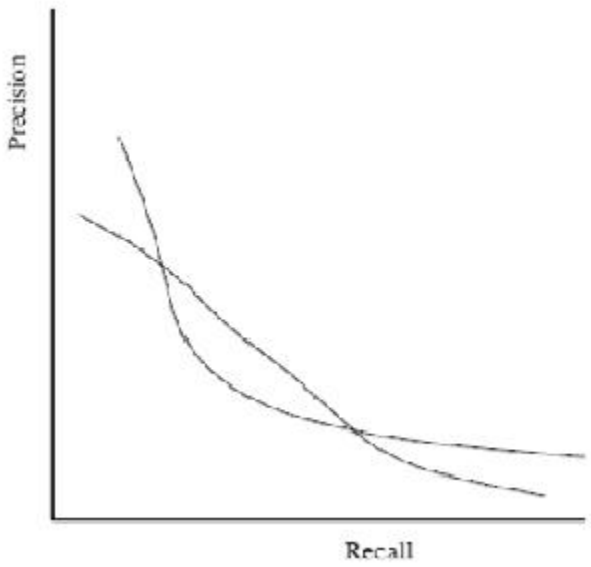

Figura 2.4: Curvas Precisión-Recall

En la figura, pueden observarse las curvas *Precisión-Recall* de dos sistemas de recuperación de información: Uno de ellos tiene mejor Precisión para valores muy pequeños o muy grandes de recall, mientras que el otro es mejor en los valores intermedios de Recall. No es fácil, a la vista de la gráfica, llegar a una conclusión sobre la superioridad de uno u otro. Por ello en ocasiones se utilizan métricas que resumen toda la curva para una consulta en un solo valor. Un ejemplo de estos valores es la llamada Precisión-R que se define de la siguiente manera: Si R es el número total de documentos relevantes para una consulta, llamamos Precisión-R a la Precisión cuando el número de documentos recuperados es R. Otra alternativa es combinar Recall y Precisión en una sola métrica, que consista en la media armónica de ambas.

3 Pruebas

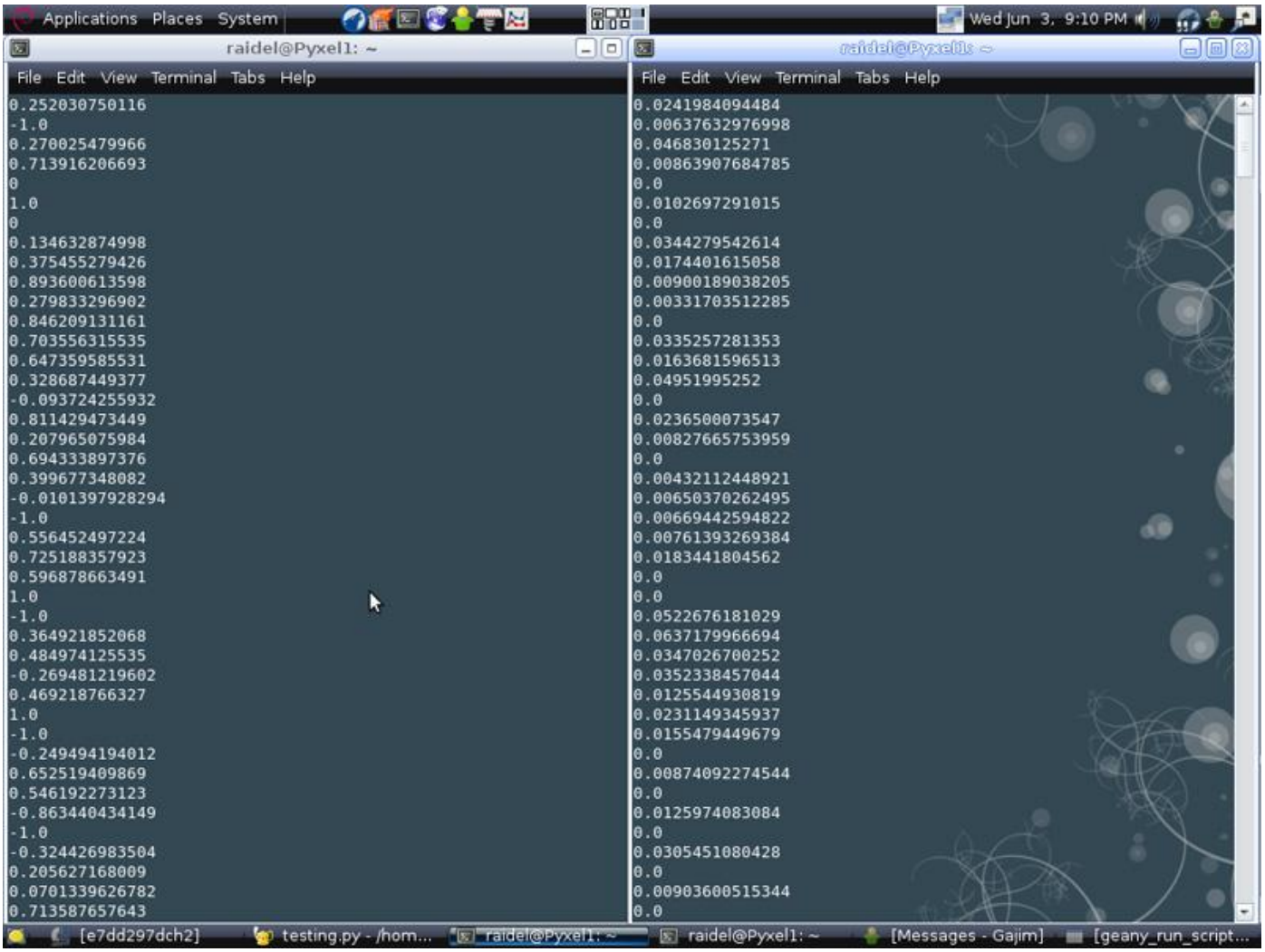

Figura 3.1 Grado de similitud por cada documento

En la figura se muestra el grado de similitud devuelto por cada documento en los nodos que se encuentran corriendo. En el nodo de la izquierda se le aplicó por el método de Correlación de Pearson y en el de la derecha por similitud basada en Coseno. Esta prueba es específicamente del algoritmo distribuido, la cual para una colección de 100 documentos se tomó solamente 0.30 segundos.

A continuación en la figura 3.2, se muestra lo que devuelve este algoritmo (la lista de ID de las imágenes recomendadas con su respectivo ranking).

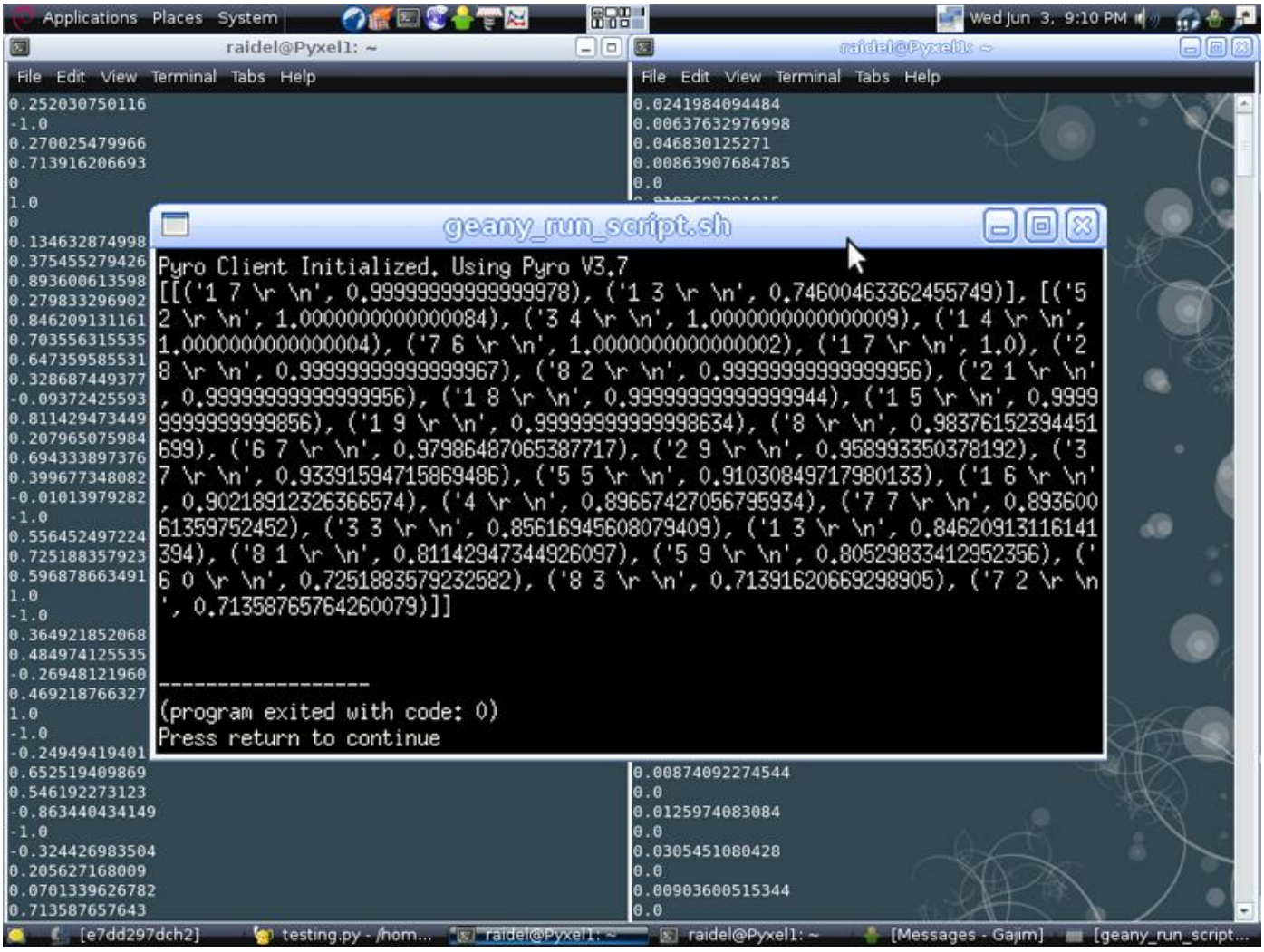

Figura 3.2 Lista de ID de imágenes recomendadas.