MINISTERIO DE EDUCACIÓN SUPERIOR

UNIVERSIDAD DE HOLGUÍN "OSCAR LUCERO MOYA"

FACULTAD DE INFORMÁTICA Y MATEMÁTICA

DISEÑO DE UN MODELO DE REALIMENTACIÓN POR RELEVANCIA BASADO EN APRENDIZAJE ACTIVO PARA EL ENTORNO PERIODÍSTICO DIGITAL CUBANO

Tesis en opción al Título de Máster en Matemática Aplicada e Informática para la Administración

Autor: Ing. María del Carmen Rodríguez Hernández

Tutores: Dr.C. Luis Cuevas Rodríguez

M.Sc. Sergio Cleger Tamayo

Holguín

Septiembre, 2011

A mi familia,

a Yito,

a Sergio,

a Rosa,

y a todos los que creyeron en mí.

AGRADECIMIENTOS

Llegado este momento tan ansiado, sólo me queda agradecer:

En primer lugar, a mis tutores Sergio Cleger y Luis Cuevas, por haber confiado en mí desde el primer momento y enseñarme a crecer ante las dificultades. Ambos son grandes ejemplos a seguir en el ámbito profesional y científico. En especial, quiero agradecer a Sergio por aguantarme las 24 horas del día y dedicarme de ellas 25. Gracias por tus espléndidas terminaciones en la redacción de la investigación, por darme seguridad y emprender junto conmigo el camino de la Recuperación de Información. En todo este tiempo me has demostrado ser un gran amigo y para mí esto tiene un valor inmenso. Quiero que sepas que a pesar de tu ausencia el día de la defensa mi logro es también todo tuyo.

A Rosa Urquiza, por permitir que entrara a su vida como su “niñita insistente”. Me siento afortunada por tenerla siempre a mi lado y parecerme en muchos aspectos a usted. Gracias por ser parte de mi seguridad espiritual, laboral y científica. Ha sido testigo de todas mis lágrimas pero también de mis alegrías, porque sus consejos siempre me dan la fuerza necesaria para enfrentar los problemas con una sonrisa. Gracias por permitir que cumpla mis deseos profesionales y tener siempre un tiempo para dedicarme. Es muy importante en mi vida y un gran paradigma para mí.

A los Juanes, por introducirme y guiarme en el campo de la Recuperación de Información.

A Irma Lerma, por su buen carácter y dedicarme parte de sus vacaciones.

A mis compañeros de trabajo y amigos por su apoyo e interés.

A los miembros del tribunal por sus valiosas sugerencias.

A mis padres mi más sincero agradecimiento, por haberme apoyado y escuchado en todo momento, por ser parte de mis alegrías y tristeza, por la comprensión, amor y sobre todo por haberme guiado por el mejor camino. A mi hermanita, por estar al tanto de todo.

Mi más sentida gratitud va dirigida al sufridor incondicional de mi estrés y mis ansias obsesivas por trabajar en la computadora, a la persona que me ha esperado y aguantado durante todo este tiempo. No tengo palabras para agradecerle todo lo que ha hecho por mí, por regalarme nuestro tiempo. Gracias Yito, por ser tan especial y estar siempre ahí para mí.

RESUMEN

En la Casa Editora **¡*ahora*!** de Holguín se actualiza diariamente cuatro veces el periódico **¡*ahora*!** digital, con el fin de satisfacer las necesidades informativas del país, la provincia y el mundo. En muchas ocasiones, los periodistas y **Editora Web principal** requieren encontrar información oportuna en breve tiempo, haciendo del proceso de búsqueda una herramienta indispensable durante el proceso editorial. Sin embargo, el algoritmo de búsqueda actual no satisface en la mayoría de las veces las necesidades de información de los usuarios.

La presente investigación propone una solución a estas dificultades, a partir del diseño de un modelo de realimentación por relevancia basado en aprendizaje activo. El modelo que se presenta garantiza alta calidad en la búsqueda de información. Por tanto, beneficiará el funcionamiento interno de la entidad, al sintetizar determinados momentos en los que transitan los entes inmersos en el proceso editorial. La propuesta puede ser implantada definitivamente en el sitio Web del periódico **¡*ahora*!** e incluso incorporada en cualquier Sistema de Recuperación de Información que lo requiera.

El documento aborda en el marco teórico de la investigación los fundamentos del aprendizaje activo en la Recuperación de Información y los motivos que conllevan el diagnóstico del proceso de búsqueda de información durante el proceso editorial del periódico **¡*ahora*!** digital. Se describe el modelo propuesto, así como los resultados obtenidos en el estudio experimental para la evaluación de diversos Sistemas de Recuperación de Información sobre una colección noticiosa con medidas de **precisión** y **exhaustividad**.

ABSTRACT

In the publishing house **¡*ahora*!**, located in Holguin the digital newspaper **¡*ahora*!** is updated four times a day, in order to meet the information needs of the country, province and the world. In many occasions, the journalists and **main Web Editor** need to find timely information in a short period of time, making the search process an indispensable tool during the editing process. However, the current search algorithm mostly fails to meet the information needs of the users.

This research proposes a solution to these difficulties by designing a model of relevance feedback based on active learning. The presented model guarantees a high quality of the information search. Therefore, it will improve the internal functioning of the entity, by synthesizing determined moments in which the editing personnel is involved. The proposal can be definitively put into force in the Web site of **¡*ahora*!** and even incorporated in any Information Retrieval System that requires so.

The document addresses, in the theoretical framework, the foundations of active learning in the Information Retrieval and the reasons leading to the diagnosis of the process of information search during the editing process of digital newspaper **¡*ahora*!**. The proposed model is described, as well as the results derived from the experimental study for the assessment of different Information Retrieval Systems about a collection of news with measures of **precision** and **recall**.

ÍNDICE DE CONTENIDO

INTRODUCCIÓN 1

CAPÍTULO 1: APRENDIZAJE AUTOMÁTICO EN LA RECUPERACIÓN DE INFORMACIÓN PARA EL CONTEXTO PERIODÍSTICO CUBANO 10

1.1 Proceso de búsqueda de noticias en el periódico ¡*ahora*! digital 10

1.2 Introducción a la Recuperación de Información 16

1.2.1 Modelos de Recuperación de Información 17

1.2.2 Arquitectura de un Sistema de Recuperación de Información 19

1.2.3 Colecciones de pruebas 28

1.2.4 Motores de búsqueda de código abierto 29

1.3 Introducción al aprendizaje automático 30

1.3.1 Métodos de aprendizaje automático 31

1.3.1.1 Aprendizaje supervisado 31

1.3.1.2 Aprendizaje no supervisado 33

1.3.1.3 Aprendizaje semi-supervisado 33

1.3.2 Aprendizaje automático en la recuperación de la información 34

1.4 Introducción al aprendizaje activo 35

1.4.1 Protocolos 36

1.4.2 Estrategias de consultas 39

1.4.2.1 Consulta por comité 39

1.4.2.2 Muestra de incertidumbre 43

1.4.2.3 Reducción de la varianza 44

1.4.2.4 Cambio esperado del modelo 45

1.4.2.5 Reducción esperada del error 46

1.4.3 Análisis general de las estrategias de consulta 47

1.4.4 Aprendizaje activo en la recuperación de información 47

1.5 Conclusiones parciales 60

CAPÍTULO 2: DISEÑO DE UN MODELO DE REALIMENTACIÓN POR RELEVANCIA BASADO EN APRENDIZAJE ACTIVO 61

2.1 Descripción de la experimentación 61

2.1.1 Colección de prueba TIME 62

2.1.2 Medidas de evaluación utilizadas 63

2.1.3 Validación estadística 65

2.1.3.1 Resultados experimentales obtenidos 66

2.2 Incorporación de modelos de realimentación por relevancia basados en aprendizaje activo en un Sistema de Recuperación de Información 74

2.3 Conclusiones parciales 78

CONCLUSIONES 79

ANEXOS I

Anexo 1: Mapa de proceso del flujo editorial del periódico ¡ahora! digital I

Anexo 2: Guía para entrevista de diagnóstico de la satisfacción de los entes inmersos en el proceso editorial del periódico ¡ahora! digital II

Anexo 3: Interfaz visual para realizar el proceso de búsqueda de información en el periódico ¡ahora! digital III

Anexo 4: Algoritmo de la estrategia de consulta por comité para un protocolo de muestreo selectivo basado en flujo IV

Anexo 5: Algoritmo de la estrategia de consulta muestra de incertidumbre para un protocolo basado en fondo V

Anexo 6: Algoritmo de la estrategia de consulta reducción esperada del error para un protocolo basado en fondo VI

Anexo 7: Resultados experimentales del desempeño de los sistemas de recuperación de información VII

ÍNDICE DE TABLAS

Tabla 1: Principales características de las colecciones estándar de prueba.....29

Tabla 2: Primera iteración para una realimentación seudo. 64

Tabla 3: Segunda iteración para una realimentación seudo. 64

Tabla 4: Primera iteración para una realimentación explícita.....64

Tabla 5: Segunda iteración para una realimentación explícita. 64

ÍNDICE DE FIGURAS

Figura 1: Arquitectura de un sistema de recuperación de información. 20

Figura 2: Representación gráfica de la frecuencia de los términos ordenados según su posición en la ordenación: ley de Zipf. 22

Figura 3: Proceso de realimentación por relevancia (se resalta con líneas discontinuas) involucrado en un SRI. 26

Figura 4: Proceso de clasificación de documentos. 31

Figura 5: Los vectores de soporte y la separación de hiperplanos. 32

Figura 6: Proceso de realimentación por relevancia con el algoritmo de clasificación SVM. .35

Figura 7: Protocolos del aprendizaje activo. 38

Figura 8: El ciclo de aprendizaje activo basado en fondo. 38

Figura 9: Ejemplo de regiones del espacio de versión. Todas las hipótesis son consistente con los datos de entrenamiento etiquetado en T (como se indican en las formas de polígonos), pero cada una representa un modelo diferente en el espacio de versión. 40

Figura 10: Aplicación del aprendizaje activo en la recuperación de información en el transcurso de los años. 60

NOTACIÓN

D_{nr}	…	Conjunto de documentos considerados no relevantes por el usuario en la realimentación por relevancia.
D_r	…	Conjunto de documentos considerados relevantes por el usuario en la realimentación por relevancia.
D_i	…	Documento de una colección.
M	…	Número total de documentos de la colección.
N	…	Cantidad de términos de una colección.
n_k	…	Número de documentos donde se encuentra el término k-ésimo.
Q	…	Consulta realizada en un sistema de recuperación de información.
$Sim(D_i, Q)$	…	Función que devuelve la similitud entre un documento y una consulta.
T_k	…	Término de la colección.
w_k^i	…	Peso del término k-ésimo del vector del documento i-ésimo.
w_k^Q	…	Peso del término k-ésimo del vector de la consulta.
w_1	…	Acontecimiento correspondiente a un “documento es relevante”.
w_2	…	Acontecimiento correspondiente a un “documento es no relevante”.

INTRODUCCIÓN

La moderna tecnología de la información produce cada año computadoras más potentes que permiten almacenar enormes cantidades de información con muy bajo costo. Cada día aumenta de manera exponencial el volumen de documentos, imágenes, sonidos, videos, etc. Algunos investigadores plantean que existe un fenómeno denominado sobrecarga de información (Maes, 1994), debido a que el volumen y la disponibilidad hacen que los usuarios no cuenten con suficiente tiempo físico para procesar todo el cúmulo de medios a su alcance (Carlson, 2003).

En diversas situaciones, recuperar de manera exacta y rápida entre tanta información disponible electrónicamente resulta difícil, pero de vital importancia para diferentes personas (usuarios). Esto plantea la necesidad de beneficiar de alguna manera a los usuarios, a partir de volúmenes importantes de información con la posibilidad de acceder a ésta. Como consecuencia, alrededor del año 1950 surgen definiciones formales para instaurar soluciones a esa necesidad. De ahí que el área de Recuperación de Información (RI) es un campo relacionado con la estructura, análisis, organización, almacenamiento, búsqueda y recuperación de información (Salton, 1983). Se relaciona, entre otras áreas, con la de gestión de bases de datos, inteligencia artificial, procesamiento del lenguaje natural, aprendizaje automático, biblioteconomía y documentación.

Entre 1970 y 1980, muchas de las investigaciones realizadas en la RI fueron enfocadas hacia la recuperación de documentos (periódicos, libros, artículos científicos, etc.). El objetivo esencial que se persigue en esta área es la obtención de los documentos relevantes, dados una necesidad del usuario y un conjunto de documentos. Para la presentación al usuario, los documentos relevantes son ordenados a partir de su grado de relevancia. El término **relevancia** se refiere a la medida de cómo un documento se ajusta a una consulta, y esta última es la necesidad del usuario en cuestión. Los documentos **no relevantes** suelen denominarse **ruido** y se refieren a lo opuesto de la definición de **relevante**.

Cercano a dicho período se desarrollaron modelos de RI (booleano, vectorial, probabilístico, etc.), para la especificación de cómo representar los documentos y las consultas realizadas por el usuario, cómo comparar unos con otras. La implementación de un modelo para la obtención de un software originó el nacimiento de los Sistemas de Recuperación de Información (SRI).

Con el propósito de aumentar la calidad de respuesta de los SRI, se emplea la Realimentación[1] de Relevancia (RF, *Relevance Feedback*, por sus siglas en inglés) o refinamiento de la consulta, introducida por (Rocchio, 1971), que se responsabiliza de mejorar la consulta y reducir el número de iteraciones por el usuario. El objetivo tradicional radica en agregar términos a la consulta y generar una nueva. Con ésta, en una próxima operación de recuperación se espera obtener nuevos documentos relevantes.

El cumplimiento del objetivo se logra a partir de que un usuario seleccione conjuntos de documentos **relevantes** y **no relevantes**, el sistema los analice y extraiga aquellos términos significativos. Por tanto, se calculan nuevamente los pesos globales de todos los términos de la consulta de manera que ésta exprese más la necesidad de información del usuario y genere otra lista de documentos ordenados (Tolosa). Este proceso es iterativo, porque continúa hasta que el usuario lo determina, e interactivo, porque el usuario está constantemente intercambiando con el sistema.

Por otro lado, los avances actuales del Aprendizaje Automático (ML, *Machine Learning,* por sus siglas en inglés) introdujeron nuevas mejoras a los algoritmos de realimentación por relevancia y, por ende, a la recuperación en general. Entre otras, una fue la incorporación del aprendizaje supervisado. El surgimiento de este enfoque se originó por el interés de utilizar la información que aportaban también los documentos no relevantes, que hasta el momento en la realimentación original había sido ignorada, al tener en cuenta solamente los documentos juzgados relevantes por los usuarios.

En la combinación de la clasificación con la realimentación por relevancia, el usuario tiene la opción de etiquetar algunos de los documentos recuperados a partir de la consulta inicial, acorde a si ellos son **relevantes** o **no relevantes**. Los documentos etiquetados junto con la solicitud original son suministrados al procedimiento de aprendizaje supervisado, con el fin de generar un nuevo clasificador, que es usado para producir un mejor orden de relevancia. El empleo de este enfoque posibilita recuperar documentos de mayor utilidad para el usuario con respecto a la búsqueda original. El método original de realimentación se diferencia del nuevo enfoque en que el primero refina la consulta y no tiene en cuenta la información de los documentos no relevantes, en tanto el segundo refina el clasificador y soluciona la exclusión de los ejemplos no relevantes.

[1] En el contexto de la Recuperación de Información la traducción al idioma español del término *feedback* es realimentación en vez de retroalimentación, a pesar de referirse conceptualmente a éste último.

La desventaja de utilizar algoritmos de clasificación en la realimentación por relevancia está en que requiere de un conjunto de entrenamiento suficientemente amplio para garantizar una buena exactitud. Sin embargo, en el contexto de la RI se cuenta con un número pequeño de ejemplos etiquetados para entrenar el clasificador y un número amplio de ejemplos no etiquetados, que no se aprovechan en este sentido. Además, los métodos de realimentación que incorporan el aprendizaje supervisado escogen solamente los primeros documentos más relevantes de los ordenados para la realimentación, que no es necesariamente la mejor estrategia desde la perspectiva de aprendizaje. Por ejemplo, si los dos primeros documentos tienen contenido idéntico, el beneficio de aprendizaje de estos dos será casi igual a cualquiera de ellos por sí solo.

Aún el enfoque de selección para la realimentación por relevancia no ha sido bien direccionado. Motivo por el cual, recientemente el Aprendizaje Activo (AL, *Active Learning,* por sus siglas en inglés) ha ocupado un lugar importante en la Recuperación de Información, particularmente en la realimentación por relevancia. El problema de realimentación activa es esencialmente un problema de decisión, en el cual se escogen los mejores subconjuntos de documentos para juzgar la relevancia por el usuario. (Shen, 2005)

En el contexto periodístico cubano existen numerosos periódicos digitales que acumulan las noticias que se publican diariamente, para brindarles a lectores y periodistas la información deseada. Todos estos casos tienen un propósito social muy bien marcado, el de informar sobre la actualidad, ya sea de temas nacionales como internacionales para persuadir, promover, formar opinión, educar y entretener.

En la presente investigación se eligió como caso de estudio la Casa Editora ¡***ahora***!, de Holguín, para caracterizar y describir el proceso editorial, así como el de búsqueda de información en el periódico **¡*ahora*!** digital. Esta entidad se destaca por el grado de informatización de los procesos relacionados y cumplir al igual que los restantes periódico digitales con los estándares y normativas nacionalmente orientadas.

La Casa Editora **¡*ahora*!** tiene a cargo, entre otros medios periodísticos, el periódico **¡*ahora*!** en su versión digital para difundir en todos los niveles sobre el acontecer noticioso de la provincia y el país en general. Es un orientador por excelencia, cuya misión y objetivo van dirigidos a informar, de manera que se cubran las necesidades de información en todos los ámbitos sociales.

Durante el proceso editorial del periódico **¡*ahora*!** digital (Anexo 1), la **Editora Web principal** planifica y asigna las tareas a desarrollar (días antes a sus publicaciones) a los periodistas y (o) fotógrafos disponibles. Luego, la **Editora Web principal** (o Editora Web, indistintamente) revisa, corrige (en caso de requerirlo) y publica los trabajos periodísticos en la Web.

En la fase de planificación y asignación, la **Editora Web principal** es la encargada de definir los temas noticiosos a tratar y decidir los periodistas y (o) fotógrafos que la elaborarán. En este último momento, la editora suele buscar en el sitio los periodistas que redactaron noticias análogas a las planificadas para realizar el proceso de asignación. El fin de realizar una búsqueda previa es garantizar una buena calidad en el desempeño de la tarea a ejecutar por el periodista a cargo y minimizar el tiempo de entrega debido a su experiencia en el tema.

Los periodistas, antes de iniciar el desarrollo de la tarea asignada, tienden a buscar en el sitio trabajos similares (ya sean escritos por él o no) al propuesto por la editora, para realizar la tarea en menor tiempo, en consonancia con la línea editorial y con la calidad deseada por la **Editora Web Principal**. Esto conlleva a agilizar la fase de revisión y corrección por las editoras.

En el momento de la publicación, los directivos o la **Editora Web principal** pueden decidir no publicar un artículo que había sido planificado con anterioridad, al surgir la necesidad u orientación de publicar otro en su lugar. De acuerdo con la inmediatez que requiere la elaboración de una nueva noticia (ya sea por la necesidad de reemplazo o cubrir espacios noticiosos libres) y el número reducido de periodistas disponibles, la editora generalmente busca de manera manual la información adecuada entre las almacenadas (pendientes), así como las publicadas recientemente en el periódico impreso y fuentes periodísticas externas.

En las entrevistas realizadas (Anexo 2) a trabajadores inmersos en el proceso editorial del periódico **¡*ahora*!**, destaca que una de las labores más complicada durante el proceso en general es encontrar la información oportuna en un tiempo breve, principalmente en las fases de asignación de tareas y publicación de noticias (anteriormente explicados). Aseguran que es inútil utilizar el buscador que contiene el sitio, por su pobre capacidad de respuesta. La mayoría de las veces, los periodistas no encuentran lo que buscan, aun estando seguros de su publicación. De manera general, coinciden y sugieren tanto los periodistas como la **Editora Web Principal**, en la necesidad de sintetizar la búsqueda de información oportuna, con el fin de agilizar el proceso editorial del periódico **¡*ahora*!** digital.

A raíz de los criterios emitidos en las entrevistas efectuadas a periodistas, fotógrafos y editores se pudo obtener información de las principales limitantes del proceso editorial del periódico **¡*ahora*!** digital. Los aportes arrojados por las entrevistas dieron lugar al diagnóstico del proceso de búsqueda de noticias en el sitio Web, con la intención de proponer una solución a dichas dificultades y garantizar un correcto funcionamiento del proceso editorial en general.

En particular, el buscador está destinado al usuario (periodistas, editores, directivos y lectores) que tenga necesidad de consultar información oportuna en un tiempo breve. Dicho proceso inicia una vez que el usuario ingresa en la interfaz la necesidad de información. Luego, el algoritmo de búsqueda realiza consultas a la base de datos a partir de filtros, que incluye a todas las palabras introducidas por el usuario, cualquier palabra, o una frase exacta. El resultado de la búsqueda contiene los títulos de las noticias y fragmentos de textos correspondientes, que coinciden con la consulta efectuada. El orden en que se visualizan las noticias en la interfaz del sistema puede ser de manera alfabética, por la popularidad, actualidad o antigüedad de las mismas.

El proceso de búsqueda concebido para el sitio Web del periódico **¡*ahora*!** actualmente no beneficia a los lectores o entes inmersos en el proceso editorial del mismo. Una de las causas es el no poder realizarla por autor o rango de fecha. No obstante, la deficiencia principal radica en que la mayoría de las veces el resultado de la búsqueda no concuerda con lo que anhelan los usuarios. Además, en ocasiones el sistema tiende a mostrar la mayor cantidad de noticias que están relacionadas con la consulta efectuada. De ahí que provoque en el usuario un alto agotamiento visual, al tener que revisar entre tantas noticias para encontrar la deseada o desistir de la búsqueda.

Otro de los problemas que posee es que admite consultas en el rango de 3 a 20 caracteres. Además, al realizar una búsqueda exacta asume que una misma palabra con tilde o sin tilde es completamente diferente, aunque semánticamente conserven el mismo significado. De manera análoga ocurre con las palabras que empiezan con mayúscula o minúscula. Por otra parte, la búsqueda de documentos que contengan todos los términos expresados en la consulta no suele ser exitosa, al mostrar en ocasiones documentos que no tienen ninguno de estos.

De manera general, no existe una feliz comprensión del sistema respecto a la necesidad del usuario, cuando los entes inmersos en el proceso editorial requieren de alta calidad en los

resultados de la búsqueda. Como consecuencia, provoca que el usuario itere un número amplio de veces con el sistema hasta que encuentre la noticia que busca o hasta que el agotamiento visual o físico incida en terminar la búsqueda.

El proceso de búsqueda de información oportuna durante el proceso editorial del periódico constituye un ejemplo donde se evidencia la necesidad de la aplicación de un SRI. No obstante, para propiciar un aumento en la calidad de respuesta de estos sistemas (como elemento fundamental tanto para el contexto que se explica como para cualquier otro análogo), se hace imprescindible el mejorar la comprensión automática de la necesidad de información del usuario.

Como se explicó previamente, la incorporación de técnicas del aprendizaje automático a la realimentación por relevancia es una alternativa que incide en el aumento progresivo de la calidad de respuesta de los SRI, lo cual ha sido experimentalmente demostrado por los investigadores (Hong, 2000, Zhang, 2001, Drucker, 2001, Chen, 2001, Hoi, 2004) de este ámbito. De manera general, aplican a la realimentación por relevancia técnicas del aprendizaje supervisado con el activo pero todavía quedan brechas por indagar para propiciar aún más estas mejoras.

A partir de lo anteriormente mencionado, el estudio de la recuperación de información y el proceso de búsqueda del periódico **¡*ahora*!** digital, surge el siguiente **problema científico**: ¿Cómo elevar la calidad de la búsqueda de información dentro de los sistemas de recuperación de información en el entorno periodístico cubano?

A partir del problema se delimita el **objeto de investigación**: La recuperación de información en el entorno periodístico.

Para solucionar el problema se persigue el siguiente **objetivo**: Diseñar un modelo de realimentación por relevancia que emplee técnicas de aprendizaje activo para incrementar la calidad de la búsqueda de información dentro de los sistemas de recuperación de información en el entorno periodístico.

El objetivo de la investigación delimita el **campo de acción**: Realimentación por relevancia en el proceso de recuperación de información en las publicaciones digitales cubanas.

Para guiar la investigación, se trazaron las siguientes **preguntas científicas**:

1. ¿Cuáles son los fundamentos teóricos en cuanto al empleo de técnicas del aprendizaje automático en la recuperación de información para elevar la calidad de la

búsqueda de información dentro de los SRI en el entorno periodístico y en especial en el periódico **¡*ahora*!** digital?

2. ¿Cómo diseñar un modelo de realimentación por relevancia que emplee técnicas de aprendizaje activo para incrementar la calidad de la búsqueda de información dentro de los SRI en el entorno periodístico?
3. ¿Cómo evaluar el modelo como solución propuesta para la validación de los resultados, particularizando dominios noticiosos?

Para darles respuesta a las preguntas científicas y cumplir el objetivo trazado, se realizaron las siguientes **tareas científicas**:

1. Diagnosticar el estado actual del proceso de búsqueda de noticias en el periódico **¡*ahora*!** digital.
2. Elaborar los fundamentos teóricos en cuanto al empleo de técnicas del aprendizaje automático en la recuperación de información para elevar la calidad de la búsqueda de información dentro de los SRI en el entorno periodístico.
3. Analizar, diseñar e implementar un modelo de realimentación por relevancia que emplee técnicas de aprendizaje activo para incrementar la calidad de la búsqueda de información dentro de los SRI en el entorno periodístico.
4. Evaluar el modelo como solución propuesta para la validación de los resultados, particularizando dominios noticiosos a partir de colecciones estándares de prueba.

Para dar solución a las tareas planteadas, se utilizó una combinación de métodos de trabajo científico, entre los que destacan los siguientes:

Métodos Teóricos. El método de análisis y síntesis permitió descomponer mentalmente el problema de investigación en sus partes y profundizar en el estudio de cada una de éstas, para luego sintetizarlas en una solución que las integre. El método inducción y deducción se utilizó durante el proceso de desarrollo del conocimiento científico y como vía de la comprobación teórica durante el progreso de la tesis. El histórico y lógico, para el estudio crítico de investigaciones previas de manera cronológica y para utilizar éstos como elemento de referencia.

Métodos empíricos. El método de estudio de la documentación permitió realizar una revisión bibliográfica profunda para plasmar y referenciar el conocimiento adquirido

relacionado con el objeto de estudio, y la fundamentación de la solución propuesta. El método experimental se empleó para constatar la utilidad de los resultados obtenidos a partir de las propuestas definidas, así como para comparar la calidad de recuperación de varios SRI, que emplean realimentación por relevancia con diferentes técnicas de aprendizaje activo y uno clásico sin realimentación.

Métodos estadísticos. Los métodos estadísticos, para evaluar experimentalmente el diseño del modelo como solución propuesta para la validación de los resultados.

Significación práctica: Con la realización de esta investigación se dispone de un modelo de realimentación por relevancia que emplea técnicas de aprendizaje activo para incrementar la calidad de la búsqueda de información dentro de los SRI en el entorno periodístico, el cual podrá ser incorporado al SRI que se aplique en la Casa Editora **¡*ahora*!** para beneficiar el proceso editorial o en cualquier entidad que realice procesos análogos de búsquedas y requiera elevar la calidad de respuesta del sistema.

La memoria que se presenta se encuentra estructurada, además de la introducción, en dos capítulos que recogen los principales argumentos de la investigación, las conclusiones generales arribadas, las recomendaciones para futuras investigaciones y aquellos materiales que sirvieron de referencia y sustento para este trabajo. Otros elementos que enriquecen y permiten entender mucho mejor todo lo expuesto, pueden ser consultados en la sección de anexos.

El **Capítulo 1** permitirá plasmar el marco teórico relacionado con el proceso actual de búsqueda de información oportuna durante el proceso editorial del periódico **¡*ahora*!** digital; introducir las bases de la recuperación de información, así como del aprendizaje automático y activo. Además de la exposición de algunos tipos de aprendizaje automático, se resalta su relación con la recuperación de información. Por otra parte, se centra la atención en la vinculación de la realimentación por relevancia con el aprendizaje activo, que a su vez contempla al automático para beneficiar en términos de calidad la respuesta de los sistemas de recuperación de información. De ahí, la necesidad de profundizar con anterioridad los diversos protocolos y estrategias de consulta propuestas en la literatura, y sus aplicaciones a la recuperación de información. Los elementos teóricos anteriormente mencionados favorecerán la comprensión del siguiente capítulo y de la investigación en su totalidad.

En el **Capítulo 2** se presenta el diseño del modelo de realimentación por relevancia con aprendizaje activo como solución propuesta al problema inicial de la investigación. Inicialmente se describe la colección de prueba TIME utilizada para llevar a cabo la validación estadística, las herramientas utilizadas para implementar los diversos SRI, así como las medidas de evaluación de **precisión** y **exhaustividad**, empleadas para determinar la calidad de respuesta de los seis sistemas de recuperación de información descritos en el propio capítulo. Por otra parte, se comentan los resultados obtenidos de los experimentos realizados a los diversos sistemas a partir de las pruebas estadísticas no paramétricas de Friedman y Wilcoxon. Por último, se esbozan los pasos generales a seguir para incorporar el modelo de realimentación por relevancia propuesto a un SRI tan específico como el del entorno periodístico, u otro análogo a éste.

CAPÍTULO 1: APRENDIZAJE AUTOMÁTICO EN LA RECUPERACIÓN DE INFORMACIÓN PARA EL CONTEXTO PERIODÍSTICO CUBANO

En este capítulo se explica el proceso de búsqueda de noticias inmerso en el proceso editorial del periódico **¡*ahora*!** digital, así como los conceptos básicos sobre la recuperación de información y el aprendizaje automático. La fundamentación teórica expuesta favorecerá la comprensión del próximo capítulo y de la tesis en su totalidad. Inicialmente, se expondrán las bases de la Recuperación de Información, describiendo brevemente los modelos clásicos, arquitectura y mecanismos de evaluación a partir de colecciones de prueba. Además, se explica la realimentación por relevancia como manera de favorecer la recuperación de información. El estudio se centra en la incorporación de técnicas del aprendizaje automático y activo a la realimentación del usuario para beneficiar en término de calidad la respuesta de los SRI. Del aprendizaje activo, se exponen los principales protocolos y estrategias de consulta propuestas en la literatura, así como sus aplicaciones en la recuperación de información.

1.1 Proceso de búsqueda de noticias en el periódico ¡*ahora*! digital

Los medios de difusión masiva en Cuba tienen un propósito social muy bien marcado: informar de la actualidad, ya sea sobre temas nacionales como internacionales para persuadir, promover, formar opinión, educar y entretener. Entre otros medios, los periódicos impresos regionales y nacionales tienen una versión digital con el fin de dar a conocer la verdad cubana a cualquier nivel, ya sea nacional o internacional con una frecuencia diaria y un alto grado de certeza y credibilidad.

Existen periódicos digitales que se centran en informar sobre el acontecer nacional e internacional, tales son los casos del periódico **Granma**[2], **Prensa latina**[3], **Agencia de Información Nacional** (AIN)[4], así como el **Juventud Rebelde**[5] destinado a los jóvenes y el **Trabajadores**[6] en función de la mayoría trabajadora. Con un enfoque provincial se editan los periódicos digitales **Girón**[7], en Matanzas; **Adelante**[8], en Camagüey; **Venceremos**[9], en

[2] http://www.granma.cu/
[3] http://www.prensa-latina.cu/
[4] http://www.ain.cu/
[5] http://www.juventudrebelde.cu/
[6] http://www.trabajadores.cu/
[7] http://www.giron.co.cu/
[8] http://www.adelante.cu/
[9] http://www.venceremos.cu/

Guantánamo; **Vanguardia**[10], en Villa Clara; **Victoria**[11], en la Isla de la Juventud; **Guerrillero**[12], en Pinar del Río; **La Demajagua**[13], en Granma; **26**[14], en Las Tunas; **Escambray**[15], en Sancti Spiritus; **Invasor**[16], en Ciego de Ávila; **5 de Septiembre**[17], en Cienfuegos; **Tribuna de La Habana**[18], en Ciudad de La Habana; **El Habanero**[19], en provincia de La Habana. Órganos todos ellos del Partido Comunista de Cuba (PCC) y respaldados por la Unión de Periodistas de Cuba (UPEC). Las noticias que se publican en los medios de prensa anteriormente mencionados se almacenan (desde su concepción hasta hoy) en un servidor de base de datos, que radica en el Comité Central por política de seguridad.

En la presente investigación se eligió como caso de estudio la Casa Editora ¡***ahora***!, de Holguín, para caracterizar y describir el proceso editorial, así como el de búsqueda de información en el periódico **¡*ahora*!** digital. Entidad que se destaca por el grado de informatización de los procesos relacionados y cumplir, al igual que los restantes periódicos digitales, con los estándares y normativas nacionalmente orientadas.

La Casa Editora **¡*ahora*!** tiene a cargo la revista **Ámbito**, destinada al intercambio cultural; **Serranía**, dirigida a las zonas montañosas del Plan Turquino; el periódico **¡*ahora*!** impreso, con el fin de difundir sobre el acontecer noticioso de la provincia, y el periódico **¡*ahora*!** en su versión digital para diariamente publicar las noticias a nivel internacional. Este último, en particular, es un orientador por excelencia, cuya misión social es ser un periódico digital encargado de mostrar la realidad holguinera en diversas esferas sociales, llevando información oportuna, certera y rigurosa con un ambiente de trabajo en equipo y motivador. Derivándose de ésta, como su principal objetivo, el de cubrir las necesidades informativas relacionadas con las actividades políticas, económicas, sociales, culturales, deportivas y científicas de la provincia.

En el rango de fechas de julio de 2009 a mayo de 2011 se publicaron alrededor de 6462 noticias, las que se encuentran en la base de datos. Hasta el momento no se tiene previsto

[10] http://www.vanguardia.co.cu/
[11] http://www.victoria.co.cu/
[12] http://www.guerrillero.co.cu/
[13] http://www.lademajagua.co.cu/
[14] http://www.periodico26.cu/
[15] http://www.escambray.cu/
[16] http://www.invasor.cu/
[17] http://www.5septiembre.cu/
[18] http://www.tribuna.co.cu/
[19] http://www.elhabanero.cubaweb.cu/

eliminar las noticias que caducan, sino que se pretende conservarlas siempre y cuando exista la capacidad en el servidor. Anterior a la fecha de julio del 2009 no existe información publicada en el sitio, por problemas tecnológicos que provocaron su desaparición.

Para el logro de una mejor organización de la información en la Web y cumplir el objetivo trazado por la entidad, se establecen las secciones: **Deporte**, **Cultura** y **Salud**, que abarcan los principales acontecimientos en dichas ramas; **Holguín**, refleja temas en campos de la producción, ciencia, técnica, masas estudiantiles o sociales; **Opinión**, abarca trabajos de opinión por género (por ejemplo, producción); **Especiales**, aglomera lo noticioso que no comprende el resto de las secciones (por ejemplo, historia de la localidad). En el caso de la información referente a los cinco prisioneros del imperio, se ubica en el sitio Web **Los Cinco**, inmerso en la página del **¡*ahora*!**.

Durante el proceso editorial del periódico **¡*ahora*!** digital, la **Editora Web principal** planifica y asigna las tareas a desarrollar (días antes a sus publicaciones) a los periodistas y (o) fotógrafos disponibles. Luego la **Editora Web principal** (o Editora Web, indistintamente) revisa, corrige (en caso de requerirlo) y publica los trabajos periodísticos en la Web. En el Anexo 1 se puede visualizar el **mapa de proceso** del flujo editorial del periódico **¡*ahora*!** digital.

En la fase de planificación y asignación, la **Editora Web principal** es la encargada de definir los temas noticiosos a tratar y decidir los periodistas y (o) fotógrafos que la elaborarán. Para la elección de los temas tiene en cuenta el plan editorial emitido anualmente por el Comité Central, las orientaciones mensuales suministradas por el PCC provincial, las prioridades fijadas por el Director semanalmente a partir de las líneas y políticas editoriales trazadas, y las iniciativas de los periodistas para satisfacer las necesidades informativas en la provincia. Por otra parte, se contemplan además las noticias transmitidas en el periódico impreso, las almacenadas que no lograron ser publicadas y las de otras fuentes periodísticas externas de la prensa nacional.

En ocasiones, dicha editora suele buscar en el sitio los periodistas que redactaron noticias análogas a las planificadas para realizar el proceso de asignación. El fin de realizar una búsqueda previa es garantizar una buena calidad en el desempeño de la tarea a ejecutar por el periodista a cargo y minimizar el tiempo de entrega debido a su experiencia en el tema.

Cada periodista, antes de iniciar el desarrollo de la tarea asignada, tiende a buscar en el sitio trabajos similares (ya sean escritos por él o no) al propuesto por la editora para realizar la tarea en menor tiempo, en consonancia con la línea editorial y con la calidad deseada por la **Editora Web Principal**. Esto conlleva la necesidad de agilizar la fase de revisión y corrección por las editoras.

En el momento de la publicación, los directivos o la **Editora Web principal** pueden decidir no publicar un artículo que había sido planificado con anterioridad, al surgir la necesidad u orientación de publicar otro en su lugar. De acuerdo con la inmediatez que requiere la elaboración de una nueva noticia (ya sea por la necesidad de reemplazo o cubrir espacios noticiosos libres) y el número reducido de periodistas disponibles, la editora generalmente busca manualmente la información adecuada entre las almacenadas (pendientes), así como las publicadas recientemente en el periódico impreso y fuentes periodísticas externas.

La información resultante del proceso editorial es destinada a cualquier lector con acceso a la Web desde cualquier parte del mundo. Con este propósito, el sitio Web del **¡*ahora*!** se actualiza en diversos momentos del día: a las 8:00 am, entre las 10:30 am y 1:30 pm por ser el horario de más tráfico en la página, 4:00 pm y por último, entre las 10:30 pm y 11:30 pm, por la diferencia de horario.

En las entrevistas realizadas (Anexo 2) a trabajadores inmersos en el proceso editorial del periódico **¡*ahora*!** digital, se pudo apreciar que las personas están contentas con la labor que realizan. La comunicación entre los compañeros es amena y transita en un ambiente agradable. La asignación de las tareas planificadas y fechas de entrega es comunicada a los periodistas y fotógrafos por la **Editora Web principal** durante la semana. En las fechas previstas, la editora se retroalimenta de los periodistas para saber el estado de los trabajos.

En particular, la **Editora Web principal** afirma que el proceso editorial del periódico digital se concibió desde un inicio bastante corto para poder conseguir de manera satisfactoria todas las actualizaciones necesarias por día. Sin embargo, destaca que una de las labores más complicada durante el proceso en general es encontrar la información oportuna en un tiempo breve. Principalmente en las fases de asignación de tareas y publicación de noticias (anteriormente explicados). Comenta además que este hecho sucede de manera análoga en las restantes entidades que llevan a su cargo un diario digital, y que los recursos humanos disponibles son comunes para este tipo de periódico pero también para otros que se editan en la entidad.

Aseguran que es inútil utilizar el buscador que contiene el sitio por su pobre capacidad de respuesta, en la mayoría de las veces los periodistas no encuentran lo que buscan, aun estando seguros de su publicación. De manera general, tanto los periodistas como la **Editora Web Principal** coinciden en la necesidad de sintetizar la búsqueda de información oportuna, con el fin de agilizar el proceso editorial del periódico **¡*ahora*!** digital. No obstante, reconocen que las condiciones de trabajo son las mejores, al contar con un local climatizado, computadora, acceso a internet (relativamente rápido), correo nacional e internacional, teléfono y acceso remoto desde sus casas. Además, se sienten estimulados con la labor que desempeñan.

Con el fin de comprobar el cumplimento de la misión y el objetivo trazado en el periódico **¡*ahora*!** digital, la **Editora Web principal** analiza la opinión de los lectores con los resultados de la encuesta sobre las noticias publicadas en el sitio. Además, supervisa la calidad y entrega en tiempo de los trabajos periodísticos. De no cumplirse con la realización de algún trabajo o la entrega en tiempo de acuerdo al plan, la **Editora Web principal** debe reemplazar de manera inmediata la noticia y como consecuencia, alterar el plan editorial. Por otra parte, los directivos verifican en cada publicación la calidad en general y si se cumplió en tiempo con la planificación prevista de acuerdo a las proyecciones de la entidad y la misión del periódico **¡*ahora*!** digital. Además están al tanto de las críticas recibidas por vía telefónica, correo o personal para analizarlas (de ser necesario) con los entes inmersos en el proceso editorial.

A raíz de los criterios emitidos en las entrevistas efectuadas a periodistas, fotógrafos y editores se pudo obtener información de las principales limitantes del proceso editorial del periódico **¡*ahora*!** digital. Los aportes arrojados por las entrevistas dieron lugar al diagnóstico del proceso de búsqueda de noticias en el sitio Web, con la intención de proponer una solución a dichas dificultades y garantizar un correcto funcionamiento del proceso editorial en general.

El buscador está destinado al usuario (periodistas, editores, directivos y lectores) que tenga necesidad de consultar información oportuna en un tiempo breve. Dicho proceso inicia una vez que el usuario ingresa en la interfaz la necesidad de información. Luego, el algoritmo de búsqueda realiza consultas a la base de datos a partir de filtros, que incluye a todas las palabras introducidas por el usuario, cualquier palabra o una frase exacta. El resultado de la búsqueda contiene los títulos de las noticias y fragmentos de textos correspondientes, que

coinciden con la consulta efectuada. El orden en que se visualizan las noticias en la interfaz del sistema puede ser de manera alfabética, por la popularidad, actualidad o antigüedad de las mismas. La cantidad de noticias a mostrar al usuario no es constante, sino que puede ser acotada por el usuario si lo deseara. En el Anexo 3 se muestra la interfaz de búsqueda de información del sitio Web del periódico **¡*ahora*!**.

Con respecto a los restantes periódicos digitales cubanos (mencionados al inicio del epígrafe), algunos permiten realizar la búsqueda de información similar al proceso descrito anteriormente. Sin embargo, otros solamente la ajustan al buscador de **google**[20] en internet.

El proceso de búsqueda concebido para el sitio Web del periódico **¡*ahora*!** actualmente no beneficia a los lectores o entes inmersos en el proceso editorial del mismo. Una de las causas es el no poder realizarla por autor o rango de fecha. No obstante, la deficiencia principal radica en que la mayoría de las veces el resultado de la búsqueda no concuerda con lo que anhelan los usuarios. Además, en ocasiones el sistema tiende a mostrar la mayor cantidad de noticias que están relacionadas con la consulta efectuada. De ahí que provoque en el usuario un alto agotamiento visual, al tener que revisar entre tantas noticias para encontrar la deseada o desistir de la búsqueda.

Otro de los problemas que posee es que admite consultas en el rango de 3 a 20 caracteres. Además, al realizar una búsqueda exacta asume que una misma palabra con tilde o sin tilde es completamente diferente, aunque semánticamente conserven el mismo significado. De manera análoga ocurre con las palabras que empiezan con mayúscula o minúscula. Por otra parte, la búsqueda de documentos que contengan todos los términos expresados en la consulta no suele ser exitosa, al mostrar en ocasiones documentos que no tienen ninguno de estos.

De manera general, no existe una feliz comprensión del sistema respecto a la necesidad del usuario, cuando los entes inmersos en el proceso editorial requieren de alta calidad en los resultados de la búsqueda. Como consecuencia, provoca que el usuario itere un número amplio de veces con el sistema hasta que encuentre la noticia que busca o hasta que se agote visualmente y termine la búsqueda. Irónicamente, a veces el que busca es un propio periodista que está seguro de que su trabajo está publicado en la Web, pero no lo encuentra. En todos estos casos provoca que el usuario pierda credibilidad en el proceso de búsqueda y lo deje de utilizar.

[20] http://www.google.com/

El proceso de búsqueda de información oportuna durante el proceso editorial del periódico, (como se explicó al inicio del epígrafe, la investigación utiliza el **¡*ahora*!** como caso de estudio, por ser representativo de los periódicos digitales cubanos) constituye un ejemplo donde se evidencia la necesidad de la aplicación de un SRI. No obstante, para propiciar un aumento en la calidad de respuesta de estos sistemas (como elemento fundamental tanto para el contexto que se explica como para cualquier otro análogo), se hace vital el mejorar la comprensión automática de la necesidad de información del usuario.

1.2 Introducción a la Recuperación de Información

En el año 1950 comienzan a surgir definiciones formales de Recuperación de Información para instaurar soluciones a la necesidad de acceder de manera rápida y precisa a una información a partir de la existencia de una gran cantidad de ésta. El autor Salton considera que la RI "es un campo relacionado con la estructura, análisis, organización, almacenamiento, búsqueda y recuperación de información" (Salton, 1983), mientras que Croft opina que es "el conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental, etc." (Croft, 1987).

Por otro lado, Korfhage la define como "la localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta". Para Ricardo Baeza-Yates y otros en (Baeza-Yates, 1999), "la Recuperación de Información trata con la representación, el almacenamiento, la organización y el acceso a ítems de información" (Korfhage, 1997).

El término **información** puede corresponderse a cualquier tipo de objeto, por ejemplo una imagen, un video e incluso un sonido. Sin embargo, en este contexto se interpretará como la representación textual de cualquier objeto y se denominará genéricamente **documento**. De esta manera, podrá ser un artículo, noticia, tesis, libro u otro ejemplo.

El objetivo esencial que se persigue en el área de la RI es: dada una necesidad de información del usuario y un conjunto de documentos, ordenar los documentos a partir del grado de relevancia que estos tengan y presentarlos al usuario. El término **relevancia** se refiere a la medida de cómo un documento se ajusta a una consulta, y esta última es la necesidad del usuario en cuestión. No obstante, al delimitar la exactitud del significado del

concepto de **relevante** y **no relevante**, se torna a veces complicado explicitarlos de forma clara y concisa dentro del ámbito de la RI. En ocasiones, un documento se considera relevante cuando el contenido del mismo posea alguna significación o importancia en relación con la consulta realizada por el usuario. Existen autores que manejan la idea de la utilidad de un documento recuperado, es decir, si el mismo le va a ser útil o no al usuario (Cooper, 1973).

1.2.1 Modelos de Recuperación de Información

Cercano al nacimiento de la disciplina de RI se instauró un conjunto de modelos de recuperación con el fin de especificar "cómo se obtienen las representaciones de los documentos y de la consulta, la estrategia para evaluar la relevancia de un documento respecto a una consulta, los métodos para establecer la importancia (orden) de los documentos de salida y los mecanismos que permiten una realimentación por parte del usuario para mejorar la consulta" (Villena, 1997). En la literatura se identifican tres modelos clásicos: el booleano, el del espacio vectorial y el probabilístico.

El **modelo booleano** es uno de los más simple que se conocen. La fundamentación teórica se sustenta en el Álgebra Booleana y en la Teoría de Conjuntos. El contenido de los documentos se representa como un conjunto de términos, donde el peso de cada uno toma valores binarios: 0 indica ausencia del término en el documento y 1 presencia. Las consultas se conforman con términos vinculados con alguno de los operadores lógicos **Y**, **O** y **NO**, y los resultados son referencias a documentos que satisfacen las restricciones lógicas de la expresión de búsqueda.

A partir de una consulta dada, se evalúa la expresión booleana con operaciones sobre los conjuntos (unión, intersección o complemento) formados por los documentos donde aparece cada término de la consulta. Los documentos resultantes son aquellos que hacen verdadera la consulta booleana: la medida de similitud $Sim(D_i, Q)$ devolverá 1 si el documento D_i hace verdad la expresión booleana Q, y cero, en caso contrario. Este modelo presenta como desventaja que todos los documentos resultantes poseen el mismo grado de relevancia, es decir, no hay un orden de relevancia sobre el conjunto de respuestas a una consulta.

Como solución a las deficiencias del primer modelo, Gerard Salton propone el **modelo del espacio vectorial** (Salton, 1971). En él, los documentos y las consultas son representados como vectores de términos. Cada documento puede ser visto como un vector que pertenece

a un espacio N-dimensional, donde N es la cantidad de términos que componen el vocabulario del conjunto de documentos. El **peso** es un valor numérico que refleja la importancia de un término en el documento, tal valor representa su poder de discriminación.

La forma de recuperar la información consiste en comparar el vector de consulta con los vectores de los documentos, a partir de una función de similitud. Entre las diferentes medidas existentes, la más común a utilizar en este ámbito es la del **coseno**, que devuelve el coseno del ángulo que forman ambos vectores en el espacio vectorial. La fórmula de dicha medida se muestra a continuación:

$$Sim(D_i, Q) = cos(D_i, Q) = \frac{D_i . Q}{\|D_i\|\|Q\|} = \frac{\sum_{k=1}^{N}\left(w_k^i . w_k^Q\right)}{\sqrt{\sum_{k=1}^{N}\left(w_k^i\right)^2 * \sum_{k=1}^{N}\left(w_k^Q\right)^2}} \quad (1.1)$$

Generalmente, en el modelo vectorial no se usan pesos negativos, de ahí que el denominador de la fórmula anterior permita normalizarla. El grado de similitud varía según la consulta que se realice. Cuanto más próximo a 1 esté el valor obtenido, más cercano a 0° será el ángulo formado por los vectores y, en consecuencia, más similares serán éstos; por el contrario, valores próximos a 0 implicarán que los vectores son ortogonales (la máxima separación posible en un espacio vectorial en el que todos los términos toman valores positivos). Con este modelo, a partir de una consulta se pueden obtener los documentos de forma ordenada según los valores resultantes de relevancia.

A pesar de ser un modelo antiguo, actualmente es el más popular por entregar buenos resultados, es simple, fácil y eficiente de implementar. Existen modelos más sofisticados, pero lo que se gana no justica el esfuerzo. Sin embargo, todos los modelos clásicos tienen ciertas falencias comunes, la más notoria es la incapacidad para capturar las relaciones entre términos. (Baeza-Yates, 2005)

Más adelante surge el **modelo probabilístico**, con el fin de representar el proceso de recuperación de información desde el punto de vista de las probabilidades. Específicamente, el modelo de Recuperación con Independencia Binaria (BIR, *Binary Independence Retrieval*, por sus siglas en inglés) propuesto por (Robertson, 1977), es conocido como el más simple. Cada documento es representado como un vector binario, el peso 0 indica ausencia del término en el documento y 1 presencia del término.

Este principio se fundamenta en el cálculo de la probabilidad de que el documento D_i sea relevante a la consulta Q realizada. En este caso, la medida de similitud entre la consulta y el documento será la probabilidad a favor de la relevancia. Esto puede ser expresado como la proporción entre la probabilidad de relevancia y la de no relevancia,

$$Sim(D_i, Q)\frac{Pr(w_1|D_i)}{Pr(w_2|D_i)} \tag{1.2}$$

donde $Pr(w_1|D_i)$ y $Pr(w_2|D_i)$ representan las probabilidades de que un documento D_i sea relevante o no relevante, respectivamente, dada la consulta Q. En este modelo, al igual que para el modelo del espacio vectorial se pueden ordenar los resultados obtenidos en función del grado de relevancia, debido a que se usan pesos.

1.2.2 Arquitectura de un Sistema de Recuperación de Información

La implementación de un modelo de recuperación de información (modelo booleano, del espacio vectorial, probabilístico u otro) para la obtención de un software origina el nacimiento de los SRI. En su arquitectura básica (Figura 1), se destacan tres módulos principales, tales como la base de datos documental, el subsistema de consultas y el mecanismo de recuperación.

Un SRI parte de una **base de datos documental**[21] sobre la cual se deben realizar operaciones de recuperación de información. Para poder realizar dichas operaciones, es necesario obtener primero una representación de todos sus documentos, cuyo texto tiene una estructura libre y sin formato. En un inicio están compuestos por sucesiones de palabras que forman estructuras gramaticales (por ejemplo, oraciones y párrafos). Tales documentos están escritos en lenguaje natural y expresan ideas de su autor sobre un determinado tema. La representación de estos documentos puede consistir finalmente en un conjunto de términos o descriptores que permitan caracterizarlos. La idea principal de estos descriptores es propiciarle una mayor eficiencia a la base de datos en cuestiones de tamaño y tiempo de indexación y recuperación.

Desde el punto de vista matemático, se considera la base de datos como una matriz cuyas columnas representan los términos o descriptores y las filas los documentos. El valor que se almacena en la intersección de una fila con una columna depende del modelo de recuperación de información de que se trate.

[21] Denominada en diferentes literaturas como corpus, colección o base de datos textual.

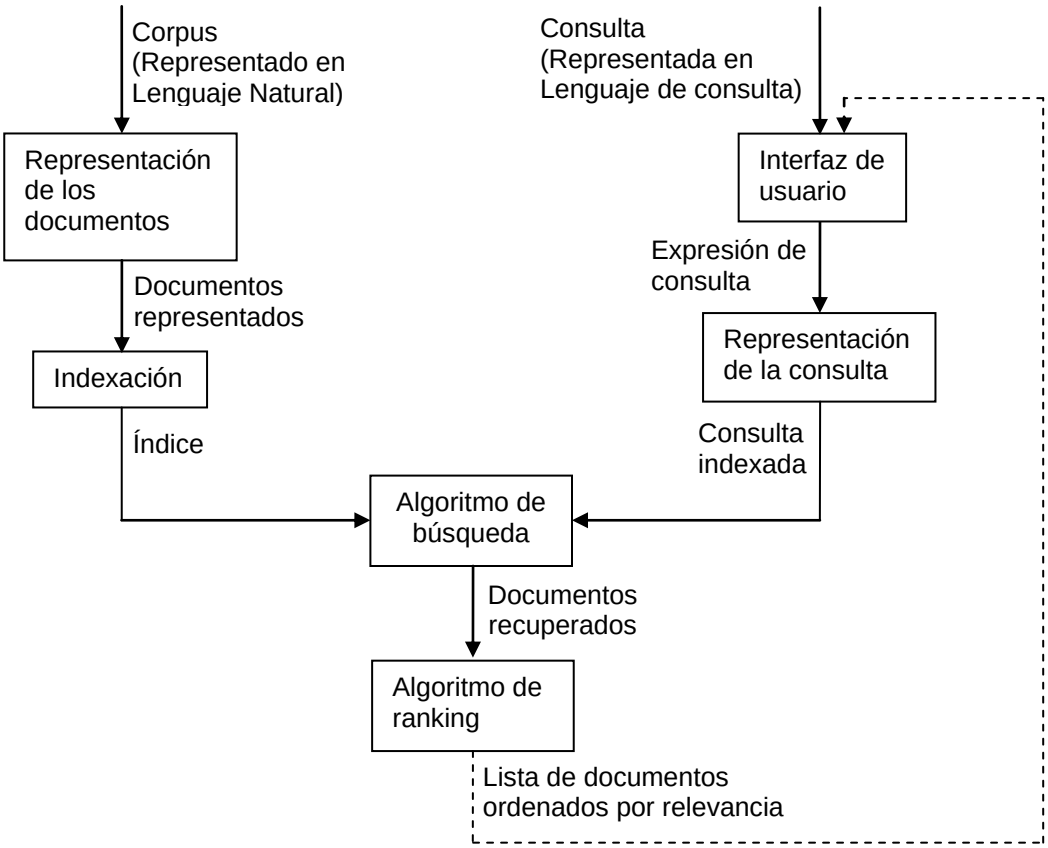


Figura 1: Arquitectura de un sistema de recuperación de información.

Para obtener la representación de los documentos existe un proceso llamado **indexación**, que permite la construcción de la **base de datos documental** (normalmente denominada **índice**) que la almacene. Entre los pasos de la indexación, se encuentra el del **análisis léxico** (llamado también *tokenization*), que tiene como objetivo transformar una cadena de caracteres en un conjunto de palabras o *tokens* (posibles descriptores), que a su vez son grupos de caracteres que presentan un significado colectivo. El analizador debe ser capaz de aplicar técnicas para tratar con caracteres especiales, dígitos, signos de puntuación, guiones y siglas. Además, de convertir el texto completo a mayúsculas o minúsculas.

La **eliminación de palabras vacías** (*stop words*, en inglés) es una de las técnicas que se aplican para reducir el tamaño del corpus. Permite eliminar palabras comunes a todos los dominios y fácilmente identificables, tales como artículos, preposiciones y conjunciones. En determinadas ocasiones se contemplan adverbios, pronombres, adjetivos, etc. En su totalidad, dependen del idioma y se almacenan en una lista construida previamente.

La **lematización** o **segmentación** (*stemming*, en inglés) son otras técnicas dependientes del idioma que reducen el número de términos del vocabulario y ahorran espacio en disco. Permiten identificar variantes morfológicas de una palabra, por ejemplo, democracia,

democrático y democratización, que se consideran palabras con un significado común y por tanto, pueden sustituirse por su raíz o lexema. Sin embargo, tienen como desventaja la pérdida de información sobre la palabra completa.

Específicamente, el ***stemming*** es un proceso heurístico ordinario que corta los finales de las palabras para reducirlos a su forma léxica o lexema. Así, "bibliotec" sería el lexema que agruparía biblioteca, bibliotecas, bibliotecario, bibliotecarios, bibliotecaria, bibliotecarias, Biblioteconomía y Bibliotecología. El empleo de esta técnica posibilita recuperar en el proceso de búsqueda documentos con términos de diversas variantes morfológicas.

En cambio, la **lematización** normalmente transforma la palabra a la base o al lema que pertenece. (Manning, 2008) Por ejemplo, las formas verbales las llevas a su infinitivo, los plurales a singulares y el género femenino al masculino. Las variantes morfológicas de la palabra y sus morfemas se descartan. Esta técnica es más precisa que la segmentación y necesita de las disciplinas de procesamiento del lenguaje natural y lingüística computacional.

Por último, se lleva a cabo la fase de **selección de términos** con el propósito de establecer criterios que permitan determinar si una palabra es un término de indexación válido, aumentar el poder discriminante entre los documentos y obtener los términos que representan mejor el contenido de los documentos. En el trabajo de (Zipf, 1949) se descubre que si se confeccionaba una lista con las palabras, junto con su frecuencia de aparición f en el documento, y se ordenaban de mayor a menor, se cumplía que la frecuencia de la palabra multiplicada por la posición p que ocupa en dicha ordenación, era igual a una constante C, es decir:

$$f * p = C \tag{1.3}$$

Más adelante, en (Luhn, 1958) se sugiere que las palabras que mejor describen el contenido se encuentran en un área comprendida entre las altamente frecuentes y las muy bajas frecuentes (Figura 2). La zona intermedia (entre ambos límites) es la que posee las palabras más significativas o de mayor contenido semántico. Generalmente, el límite superior se relaciona con el inicio de las palabras vacías y no se indexan por no tener poder discriminante entre los documentos. En el caso del límite inferior, se relaciona con las palabras de muy bajas frecuencias (o raras) y no se incluyen en el vocabulario de términos, al existir una baja probabilidad de que el usuario las use en una consulta. La mayoría de las veces, las personas utilizan un vocabulario sencillo para escribir, solo pocos autores

enriquecen el texto con palabras rebuscadas, las cuales tienden a enmarcarse en el límite inferior. Como consecuencia, la mayoría de las veces se eliminan las palabras que se encuentren en tres o menos documentos (Peña, 2003). Los límites anteriores deben establecerse con mucho cuidado, al propiciar una disminución de la exhaustividad por eliminar las palabras muy frecuentes y, en caso de no eliminarlas, puede provocar una precisión no deseada.

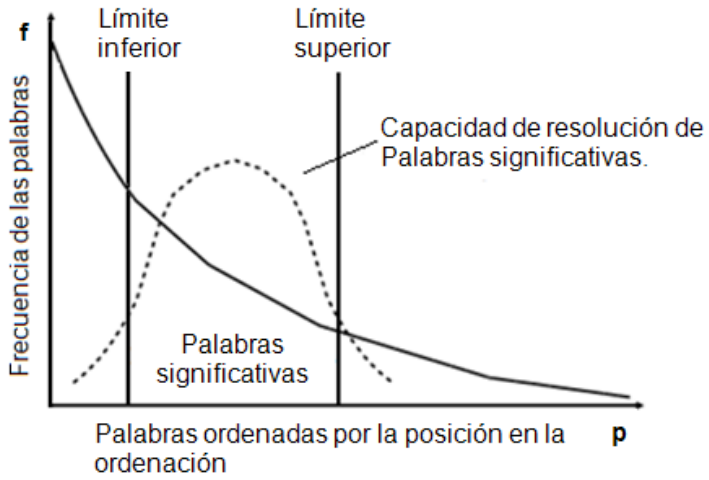


Figura 2: Representación gráfica de la frecuencia de los términos ordenados según su posición en la ordenación: ley de Zipf.

En la fase de selección es necesario determinar la importancia de un término en un documento para saber si incluirlo o no al vocabulario final. Para medir la importancia, se realiza la **ponderación de términos** o **asignación de pesos** que componen los índices de cada documento. Inicialmente, el peso de un término T_k en un documento D_i se puede determinar por la aparición o no de un término, o bien se calcula a partir de obtener su frecuencia de término, conocida como tf (*term frequency*, por sus siglas en inglés), que es la cantidad de veces que ocurre el término T_k en el documento D_i y se denota como $tf_{k,i}$.

En la medida anterior, todos los términos son considerados igualmente importantes cuando se realiza la evaluación de la relevancia en una consulta (Manning, 2008). Los términos tienen poco o ningún poder discriminante para la determinación de la relevancia. Como consecuencia, surge un enfoque inverso, que es el cálculo de la frecuencia documental inversa, conocida como idf (*inverse document frequency*, por sus siglas en inglés). En esta medida, un término T_k con poca frecuencia en la colección N tiene mayor poder discriminante respecto a un término que aparece en todos los documentos. De ahí, que surja la expresión,

$$idf_k = \log \frac{M}{n_k} \quad (1.4)$$

donde M es el número total de documentos de la colección y n_k el número de documentos donde se encuentra el término k-ésimo. La incorporación del logaritmo permite suavizar los valores arrojados de la proporción. Mientras menos términos aparezcan en la colección completa, el valor del idf será mayor. Así, el valor del idf será alto para términos raros y probablemente bajo para los muy frecuentes (Manning, 2008).

Una evolución de las medidas presentadas hasta el momento, es la denominada **frecuencia del término** por la **frecuencia documental inversa** $(tf * idf)$, que establece una relación entre la frecuencia de un término dentro de un documento $tf_{k,i}$ y su frecuencia en los documentos de la colección idf_k. Con el fin de normalizar las frecuencias para los documentos largos y cortos, la expresión quedaría de la siguiente manera:

$$tf_{k,i} * idf_k = \frac{tf_{k,i}}{largo(D_i)} \text{x} \log \frac{M}{n_k} \quad (1.5)$$

A medida que aumente el valor de esta medida, mejor será el término desde el punto de vista de la indexación. De todas las medidas explicadas, esta última es la más utilizada.

En el **subsistema de consultas**, el usuario puede formular en la interfaz del sistema su necesidad de información en un lenguaje de consulta. Sin embargo, en su forma original no puede procesarse de manera inmediata para seleccionar los documentos relevantes. La consulta debe pasar por un tratamiento previo, similar al de los documentos. De esta manera, se logra una correspondencia semántica entre el contenido de los documentos y la consulta, a la hora de verificar en el **índice** cuáles documentos pueden satisfacerla. La interfaz del usuario, además de permitir introducir la consulta, sirve para mostrar las respuestas retornadas por el sistema.

El **mecanismo de recuperación** es el responsable de evaluar el grado de similitud entre la consulta y los documentos. Como resultado de la evaluación, se recupera una lista ordenada con los documentos que considera **relevante**. Se establece que el primer documento de dicha lista corresponde al más relevante respecto a la consulta y así sucesivamente en orden decreciente. La forma de calcular el grado de similitud depende del modelo de recuperación utilizado (Epígrafe 1.2.1), que confecciona la representación de los documentos y de la

consulta, para luego aplicar sobre estos una estrategia que propicie evaluar la relevancia de un documento respecto a una consulta.

En el contexto experimental, generalmente se incluye en los SRI el proceso de evaluación de desempeño del sistema en función de diversos criterios. Tal es el caso del **espacio**, que es el tamaño que consumen los índices; la **usabilidad**, como la facilidad de uso del sistema; la **eficiencia**, el tiempo de indexación y de recuperación, así como la **eficacia** para evaluar la calidad de la recuperación.

En términos de **eficacia**, el concepto de relevancia es imprescindible para medir la efectividad de los SRI, donde la **efectividad** es puramente una medida de la capacidad del sistema para satisfacer al usuario en términos de la relevancia de los documentos recuperados (van Rijsbergen, 1999). De ahí, surgen las siguientes medidas de evaluación:

$$Exhaustividad\ (E) = \frac{Cantidad\ de\ documentos\ relevantes\ recuperados}{Cantidad\ de\ documentos\ relevantes} \qquad (1.6)$$

$$Precisión\ (P) = \frac{Cantidad\ de\ documentos\ relevantes\ recuperados}{Cantidad\ de\ documentos\ recuperados} \qquad (1.7)$$

La **exhaustividad** pretende plantear la proporción de documentos relevantes que se recuperan respecto a la cantidad de ellos que hay en la **colección documental**. La **precisión** es una de las más populares y se responsabiliza de medir la proporción de documentos que son realmente relevantes sobre los recuperados tras una consulta. Mientras más documentos relevantes se recuperen, más preciso será el SRI.

Empíricamente se ha comprobado que una alta exhaustividad se acompaña de una muy baja precisión y viceversa, es decir, existe una relación inversa entre ambas (Cleverdon, 1972). En (Martinez, 2004) se plantea que muchos usuarios consideran más importante la precisión, al no preocuparse tanto por los documentos que no se recuperan, mientras encuentren información relevante en un breve tiempo. Unido a este criterio, Cleverdon considera que para el usuario la precisión resulta mucho más interesante con respecto a la exhaustividad, al valorarse más las salidas sin **ruido**. No obstante, hay situaciones donde un usuario puede estar interesado en valores altos de exhaustividad. Existen otras medidas, algunas son definiciones nuevas y otras combinaciones de la exhaustividad y la precisión que propician un valor único de la efectividad de la recuperación de un sistema.

En la recuperación de información, uno de los problemas más significativos consiste en formular la consulta de tal manera que exprese de manera adecuada la necesidad de información del usuario. Algunas situaciones de sinonimia o polisemia (u otras menos importantes, como la homonimia, la antonimia, la hiperonimia, la hiponimia o la anáfora) provocan que el mismo concepto pueda expresarse con palabras diferentes y una misma palabra pueda aparecer en documentos que tratan sobre temas distintos. En esta situación, es muy común que los usuarios tengan que reformular la consulta para obtener mejores resultados.

A raíz de estas razones, se han propuesto mecanismos que permiten ayudar al usuario en la formulación de la consulta. Tal es el caso de la **expansión de consultas**, que realiza una ampliación de nuevos términos a la consulta inicial y determina nuevamente la importancia de cada término en la nueva consulta. El propósito general es ampliar el número de términos que mejor definan la necesidad informativa del usuario de acuerdo a la colección documental y al modelo de recuperación utilizado. Este proceso se puede auxiliar de los juicios de relevancia emitidos por el usuario.

Actualmente, en muchos SRI se emplea la Realimentación por Relevancia (RF, *Relevance Feedback*, por sus siglas en inglés), introducida por (Rui, 1998). Es un proceso interactivo e iterativo, que se ha convertido en un modo eficaz de modificar y expandir las consultas de usuario para mejorar la calidad de los sistemas de recuperación (Xu, 2003). El efecto de tal proceso es mover la consulta en dirección de los artículos relevantes y lejos de los no relevantes, en la expectativa de recuperar los artículos más queridos en una búsqueda posterior (Salton, 1990).

En el proceso de **realimentación por relevancia** (Figura 3) un usuario envía una consulta, el SRI retorna un conjunto inicial de documentos resultantes y le pide al usuario juzgar los documentos que son o no relevantes. Luego, el sistema reformula la consulta basada en los juicios de usuario, y retorna un conjunto con nuevos resultados. (Xiang, 2009) Este proceso continúa hasta que el usuario lo determine o se encuentre satisfecho, de ahí que se considere un proceso iterativo e interactivo, por el diálogo frecuente del usuario con el sistema.

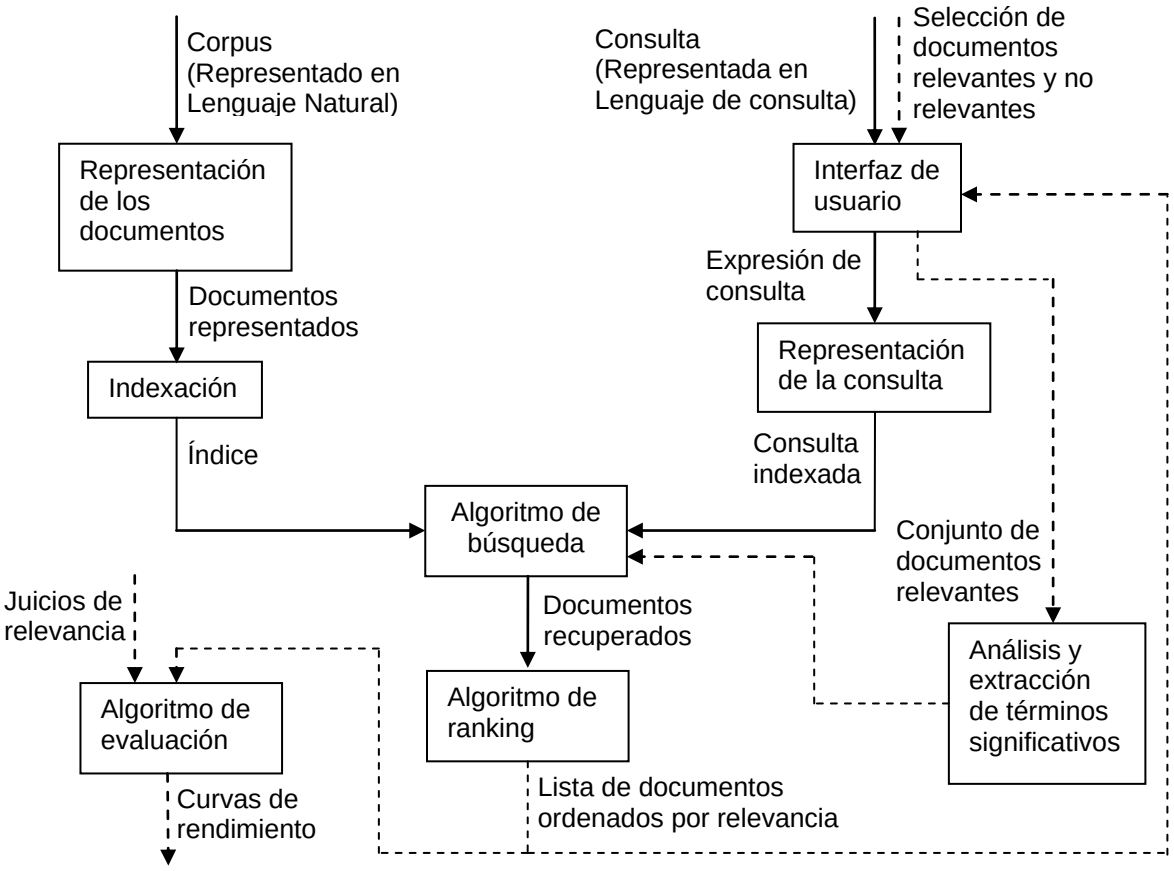


Figura 3: Proceso de realimentación por relevancia (se resalta con líneas discontinuas) en un SRI.

La realimentación por relevancia se clasifica de acuerdo a la manera en que se escogen los documentos relevantes en explícita, implícita y seudo. La **realimentación explícita** obtiene los documentos relevantes a partir de asesores[22] de relevancia, que indican la relevancia de un documento recuperado dada una consulta. Este tipo de realimentación es definida como explícita solamente cuando los asesores conocen que la realimentación proporcionada es interpretada como juicios de relevancia. En el caso de la **realimentación implícita** se infiere desde el comportamiento del usuario. Por ejemplo, notar qué documentos selecciona para la visualización y cuáles no, la duración del tiempo que empleó en visualizar un documento, y otras. Por último, la **realimentación seudo** es un proceso automático que asume como relevante a los s primeros documentos mostrados al usuario. En otras palabras, el sistema de recuperación asume que los primeros documentos recuperados por la consulta inicial del usuario son relevantes, y aplica la realimentación por relevancia para generar una nueva consulta, que es usada para crear el ranking de documentos a presentar al usuario.

Según (Salton, 1990), el procedimiento de realimentación por relevancia exhibe las siguientes ventajas principales:

[22] Se refiere a las personas que omiten los juicios de relevancia durante el proceso de realimentación.

1. Protege al usuario de los detalles del proceso de formulación de consulta y permite la construcción de expresiones de búsqueda útiles, sin el conocimiento íntimo de la colección y entorno de búsqueda.
2. Divide la operación de búsqueda en una secuencia de pequeños pasos de búsqueda, diseñados para acercarse gradualmente al área deseada.
3. Posee un proceso de modificación de consulta controlado, diseñado para enfatizar algunos términos y no destacar otros, como es requerido en entornos de búsqueda particulares.

En el **modelo del espacio vectorial**, la **realimentación por relevancia** es usualmente implementada a partir del algoritmo Rocchio (Rocchio, 1971). Es considerado uno de los algoritmos más temprano y el más clásico en este ámbito. En las iteraciones de realimentación se modifican los vectores de consulta, al incrementarse el peso de los términos contenido en los documentos relevantes y penalizarse el de los documentos no relevantes. A continuación se expone la fórmula a utilizar en el espacio vectorial,

$$\vec{Q}' = \alpha\vec{Q} + \frac{\beta}{|D_r|}\sum_{\vec{D}_i \in D_r} \vec{D}_i - \frac{\gamma}{|D_{nr}|}\sum_{\vec{D}_i \in D_{nr}} \vec{D}_i \qquad (1.8)$$

donde $\vec{Q}$ es el vector de la consulta original, $\vec{Q}'$ es el vector de la consulta realimentada, $\vec{D}_i$ es el vector que representa al $i-ésimo$ documento, D_r es el conjunto de documentos considerados relevantes por el usuario, D_{nr} es el conjunto de documentos considerados no relevantes, y α, β y γ son parámetros que permiten ajustar el impacto de los documentos relevantes y no relevantes. Los términos que resulten con peso negativo se ignoran.

Muchas extensiones fueron propuestas basadas en dicho algoritmo, tal es el caso de Ide regular y Ide dec-hi en (Ide, 1971). El primero no tiene en cuenta la normalización que realiza Rocchio con los pesos de los documentos, en el segundo se tiene en cuenta solo el documento no relevante que se sitúe en la posición más alta en la ordenación de documentos recuperados. Ambos fueron comparados experimentalmente con el algoritmo Rocchio en (Salton, 1990), los dos primeros arrojaron mejores resultados que Rocchio.

Para la **evaluación de la efectividad** en la realimentación por relevancia (Figura 3), se pueden desarrollar por dos estrategias. La primera consiste en comenzar con una consulta inicial $\vec{Q}$ y realizar una gráfica con los valores de **precisión** y **exhaustividad**. De igual

manera, se grafican estos valores pero para la próxima ronda de realimentación del usuario con la consulta modificada $\vec{Q}'$. La segunda gráfica generará curvas mucho mejor que en la primera, pero en realidad la segunda aporta poca información nueva. En parte, se debe a que los documentos relevantes conocidos (juzgado por el usuario) son clasificados ahora más alto. La imparcialidad exige que sólo se debiera evaluar con respecto a los documentos no vistos por el usuario.

La segunda estrategia es similar a la anterior, pero utiliza los documentos de la **colección residual** (es el conjunto de documentos menos aquellos que fueron juzgados relevantes) para la segunda ronda de evaluación

1.2.3 Colecciones de pruebas

Las colecciones de pruebas se utilizan en la fase experimental para la evaluación de los SRI. Se componen de: un conjunto de documentos que constituyen el corpus, consultas predefinidas y juicios de relevancia, que relacionan las consultas con los documentos del corpus que son relevantes a éstas.

La construcción de colecciones de pruebas es una tarea compleja. Esto se debe al alto costo que genera la identificación de los documentos relevantes o no relevantes, dada una consulta por un decisor humano; pues la mayoría de las veces los entes presentes en el proceso no están de acuerdo acerca de la relevancia.

En 1960, Gerard Salton y su grupo de desarrollo elaboraron en la Universidad de Cornell un SRI, denominado SMART[23] (*System for the Mechanical Analysis and Retrieval of Text*, por sus siglas en inglés). El nacimiento de SMART propició el surgimiento de colecciones de pruebas de diversos ámbitos. Entre ellas, se encuentran las colecciones de prueba TIME, ADI, CACAM, Cranfield, CISI y Medlars. De manera general, tales colecciones son pequeñas. En la Tabla 2 se muestran las principales características de las mismas.

[23] ftp://ftp.cs.cornell.edu/pub/smart/

Colección	Número de documentos	Número de términos	Número de consultas	Ámbito
ADI	82	828	35	Ciencias de la información
CACM	3204	7562	64	Ciencia de la computación
CISI	1460	4985	112	Biblioteconomía
CRANFIELD	1398	3857	225	Aeronáutica
MEDLARS	1033	7170	30	Medicina
TIME	425	5923	83	Noticias

Tabla 1: Principales características de las colecciones estándar de prueba.

En 1992 surge la Conferencia de Recuperación de Texto (TREC[24], *Text REtrieval Conferences*, por sus siglas en inglés), que es el foro anual de evaluación de la comunidad de RI, patrocinado por el Departamento de Defensa de los Estados Unidos y el Instituto Nacional de Estándares y Tecnología (NIST, *National Institute of Standards and Technology*, por sus siglas en inglés). TREC ha producido una serie de colecciones de pruebas con fuentes documentales diversas, desde periódicos como el *Wall Street Journal* o el *Financial Times*, documentos de Patentes de los Estados Unidos, hasta colecciones de páginas Web. Cada una de estas colecciones posee: un número grande de documentos, un número de temas (consultas) y juicios de relevancia (respuestas correctas).

1.2.4 Motores de búsqueda de código abierto

Actualmente existe un número amplio de aplicaciones de código abierto destinadas a la RI. No obstante, de acuerdo a la comparación realizada en (Middleton, 2007) con respecto a la cantidad de documentación disponible y actualizada, al amplio respaldo en la comunidad de desarrolladores e investigadores, el tiempo que emplean para la indexación y respuesta a las consultas, el tamaño de los índices y los valores que arrojan de precisión y exhaustividad, se destacan: Apache Lucene, Minion, Terrier, Indri, DataParkSearch, Swish-e, MG4J, mnGoSearch y Solr.

A raíz de los elementos antes mencionados, se decidió utilizar en la fase experimental el motor de búsqueda *Apache Lucene* para implementar los diferentes SRI en el Entorno de Desarrollo Integrado (IDE, *Integrated Development Environment*, por sus siglas en inglés) *NetBeans*. *Lucene* fue escrito originalmente por Doug Cutting y se integra a Apache en el 2001. Hasta el momento, se destaca por ser el más popular de RI basado en Java. Se soporta mediante una combinación del modelo espacio vectorial booleano puro; se utiliza en un gran número de sitios y aplicaciones, tales como el IDE Eclipse, la enciclopedia Británica,

[24] http://trec.nist.gov/

Wikipedia, Amazon, el buscador de código fuente como Krugle y Sourcerer, es el núcleo del motor de búsqueda Web Nutch y Solr.

1.3 Introducción al aprendizaje automático

El aprendizaje es la actividad que desarrolla el ser humano para aprender (Concepción, 2005). Según la Real Academia Española, el término **aprender** del latín *apprehendĕre*, significa adquirir conocimiento de algo por medio del estudio o de la experiencia. Con el conocimiento aprendido, el ser humano puede resolver nuevos problemas que antes eran insolubles, dar mejores soluciones o resolverlos eficientemente a menos costo.

Lamentablemente, las computadoras no aprenden o adquieren conocimiento de la información que almacenan y, por ende, carecen de habilidades para analizar e interpretar. Aún se considera una herramienta electrónica para almacenar (con una capacidad finita) información y prestar diversos servicios que favorecen el desarrollo de un conjunto amplio de actividades para las que fueron programadas. El análisis e interpretación manual de la información que se almacena en las computadoras se torna impráctico (lento, caro y subjetivo) por sus usuarios, en la medida que los volúmenes de información crecen exponencialmente.

A finales de la década de los 80, surge la creciente necesidad de automatizar el **proceso inductivo**[25]. El intento de solucionar problemas análogos al de aprender a reconocer palabras habladas, reconocer escrituras, clasificar documentos y otros procesamientos, abre una línea de investigación para el análisis inteligente de la información e impulsa las investigaciones al aprendizaje automatizado.

El Aprendizaje Automático (ML, *Machine Learning*, por sus siglas en inglés) es una rama de la Inteligencia Artificial (IA), cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Concretamente, trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Por tanto, es un proceso inductivo del conocimiento. (Suárez, 2005)

El **aprendizaje automático** tiene muchas áreas de aplicación. Específicamente en el Lenguaje Natural, es aplicable al análisis sintáctico y morfológico, recuperación de información, extracción de información, búsqueda de respuestas, traducción automática, creación de resúmenes, minería de textos, reconocimiento y generación de voz.

[25] El proceso o aprendizaje inductivo, parte de casos particulares (ejemplos) y obtiene casos generales (reglas o modelos) que generalizan o abstraen la evidencia.

1.3.1 Métodos de aprendizaje automático

En el aprendizaje automático inductivo existen varios tipos de técnicas. Generalmente, se clasifican en dos tipos: aprendizaje no supervisado y supervisado. En el primero se encuentra, entre otras tareas, la de **agrupamiento** (*clustering*, en inglés), y en el segundo tipo la de **clasificación**.

A continuación se exponen las definiciones de ambos tipos de aprendizaje automático basado en corpus. Se profundiza solamente en la explicación de las tareas respectivas del aprendizaje no supervisado y supervisado, que se mencionaron previamente.

1.3.1.1 Aprendizaje supervisado

En el aprendizaje supervisado, los documentos están previamente clasificados y se conoce la categoría a la que pertenece cada uno (Suárez, 2005). Una tarea del aprendizaje supervisado es la clasificación (Figura 4), que se alimenta de un conjunto de documentos que han sido clasificados manualmente de antemano y sirven como ejemplos para construir un clasificador, que después será utilizado para detectar a qué categoría[26] pertenecen los nuevos documentos entrantes (González, 2005).

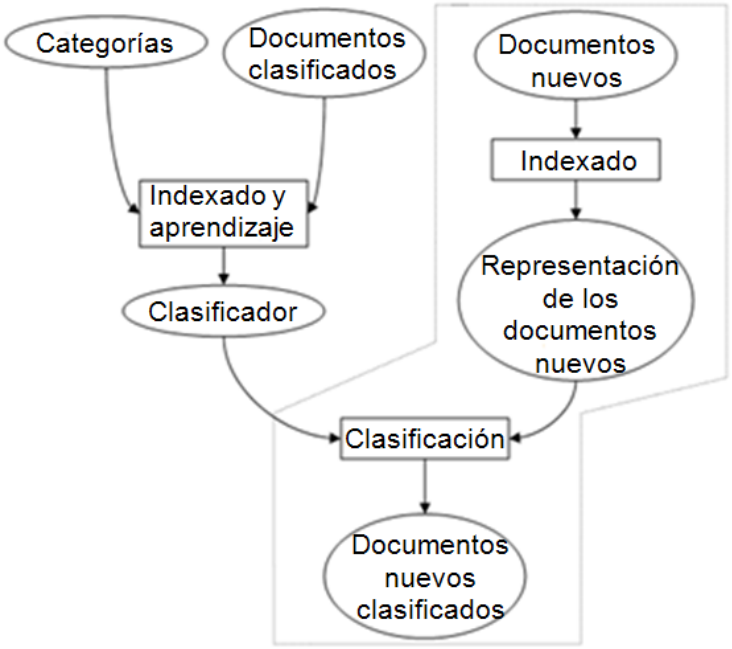


Figura 4: Proceso de clasificación de documentos.

En la clasificación textual, tanto los documentos categorizados como los nuevos a clasificar deben pasar por algunos pasos del **proceso de indexación** (análisis léxico, eliminación de

[26] Conocido en la literatura también como clase o etiqueta.

las palabras vacías, lematización o *stemming* y otras para reducir la dimensionalidad del corpus), explicado en el Epígrafe 1.2.2. La intención es llevar los documentos originales a un formato concreto, que pueda ser adecuado para los algoritmos de aprendizaje supervisado, tales como vectores de pesos de términos, como en el **modelo del espacio vectorial**.

En la literatura de este ámbito, es usual encontrar el término **clase** o **etiqueta** en lugar de **categoría**. En la clasificación automática se hallan tres tipos de clasificación: el de **una-etiqueta**, que sitúa cada documento estrictamente en una clase; **multi-etiqueta**, fija documentos a cualquier número de clases, y **binaria**, que es un tipo especial del primero, que asigna el documento a una clase o a su complemento (Sebastiani, 2002).

Recientemente, los algoritmos de clasificación, tales como **Máquinas de Vectores de Soporte** (SVM, *Support Vector Machine*, por sus siglas en inglés), han recibido gran atención en la comunidad del aprendizaje automático. Son capaces de resolver problemas de alta dimensionalidad con muy pocos ejemplos y además trabajar eficazmente cuando los ejemplos son abundantes.

El algoritmo de clasificación binaria SVM, desarrollado por (Vapnik, 1995), construye un modelo que predice si un documento nuevo pertenece a una categoría o a la otra. Los documentos de entrada son vistos como un vector p-dimensional (una lista de p documentos). El objetivo final de un algoritmo SVM es separar en dos grupos el contenido del vector. De esta forma, quedaría en un grupo los documentos positivos e_p y en el otro los negativos e_n (Figura 5). En el contexto de la Figura 5, los documentos (positivos y negativos) más cercanos a las dos líneas continuas son llamados vectores de soporte y la línea discontinua es llamado **hiperplano de decisión**. En este caso, el propósito del aprendizaje en SVM es encontrar el hiperplano con el margen máximo (óptimo).

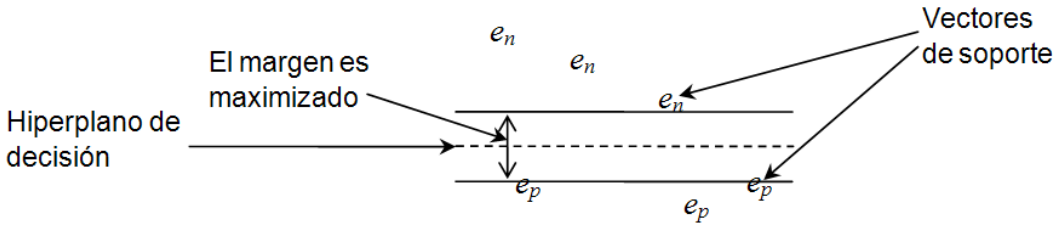


Figura 5: Los vectores de soporte y la separación de hiperplanos.

1.3.1.2 Aprendizaje no supervisado

En el aprendizaje no supervisado no hay una clasificación previa de los documentos, los cuales se agrupan automáticamente (Suárez, 2005). Una tarea del aprendizaje no supervisado es el agrupamiento. Tiene como objetivo formar una colección de grupos o subconjuntos, que cumplan con las propiedades de la homogeneidad dentro de los grupos y la heterogeneidad entre los grupos. En la primera, los documentos que pertenecen al mismo grupo deben ser tan similares como se pueda; en la segunda, los documentos que corresponden a grupos diferentes deben ser tan disímiles como se pueda. En otras palabras, trata de seguir el principio de maximizar la similitud de los documentos en cada grupo y minimizar la similitud entre los grupos.

Generalmente, los problemas de agrupamiento pueden ser divididos en dos categorías: **agrupamiento duro** y **agrupamiento borroso**. En el primero, un documento pertenece a uno y solamente un grupo; mientras que en el segundo, un documento puede pertenecer a dos o más grupos con probabilidades. (Gan, 2007)

De igual manera que para la clasificación, los documentos originales a agrupar deben ser previamente representados en un formato concreto, que pueda ser adecuado para los algoritmos de agrupamiento. Además, la aplicación de algunas técnicas para la reducción de la dimensionalidad del corpus repercute gradualmente en la eficiencia del agrupamiento y la clasificación, al disminuir el número de términos a procesar y aumentar el poder discriminante de los documentos.

1.3.1.3 Aprendizaje semi-supervisado

El propósito actual de los investigadores en el ámbito del aprendizaje supervisado es lograr entrenar un clasificador capaz de sustituir a un experto humano. Para ello, es vital tener en cuenta un número suficientemente grande de ejemplos para la etapa de entrenamiento, lamentablemente en la práctica no suele ser así. La mayoría de la veces hay un número pequeño de documentos previamente clasificados y otra parte mucho mayor sin clasificar, y como consecuencia, generan clasificadores de muy baja calidad.

Una solución a esta problemática fue la fusión de los algoritmos del aprendizaje supervisado con los del no supervisado, el cual se denominó en la literatura **aprendizaje semi-supervisado** (SSL, *Semi-Supervised Learning*, por sus siglas en inglés). Es un paradigma

del aprendizaje automático, en el cual el modelo es construido con un número pequeño de documentos etiquetados y un número grande de documentos no etiquetados.

La idea clave en el aprendizaje semi-supervisado es etiquetar documentos no etiquetados con el uso de ciertas técnicas y de esta manera incrementar la cantidad de documentos etiquetados para el entrenamiento del clasificador. Este tipo de aprendizaje ha sido empleado, entre otras aplicaciones, a la clasificación de documentos (Nigam, 2000a, Li, 2003, Liu, 2003).

1.3.2 Aprendizaje automático en la recuperación de la información

Los avances actuales del aprendizaje automático introdujeron nuevas alternativas para los algoritmos de realimentación por relevancia en la Recuperación de Información. Una de ellas fue la incorporación del aprendizaje supervisado a partir de la aplicación satisfactoria del algoritmo de clasificación SVM a problemas de realimentación por relevancia en (Hong, 2000, Zhang, 2001, Drucker, 2001, Chen, 2001, Hoi, 2004).

El surgimiento de este enfoque se originó por el interés de utilizar la información que aportaban también los documentos no relevantes. Hasta ese entonces en la realimentación original habían sido ignorados, al tener en cuenta solamente los documentos juzgados relevantes por los usuarios. En la perspectiva de la realimentación por relevancia con SVM, los documentos relevantes y no relevantes se consideran ejemplos positivos y negativos, respectivamente, y la realimentación es trasladada a un problema de clasificación binaria.

Con el nuevo enfoque de realimentación, el usuario tiene la opción de etiquetar (en **relevante** y **no relevante**) algunos de los documentos recuperados a partir de la consulta inicial. Luego son suministrados al procedimiento de aprendizaje supervisado (SVM) con el fin de generar un clasificador, que es usado para producir un nuevo orden en los resultados. De esta manera se recuperan documentos de mayor utilidad para el usuario que en la búsqueda original. Las iteraciones de realimentación continúan hasta que el usuario lo decida.

A continuación se muestra, en la Figura 6, el proceso de realimentación por relevancia con SVM inmerso en un SRI después de la primera iteración. El algoritmo Rocchio se diferencia del nuevo enfoque en que el primero refina la consulta y no tiene en cuenta la información de los documentos no relevantes, mientras que el segundo refina el clasificador y soluciona la exclusión de los ejemplos no relevantes.

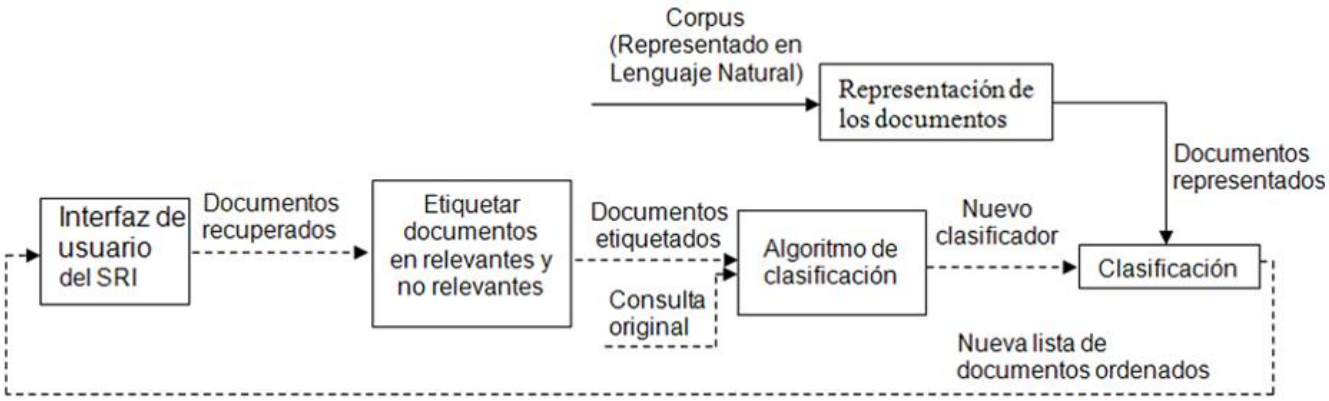


Figura 6: Proceso de realimentación por relevancia con el algoritmo de clasificación SVM.

Por otra parte, el aprendizaje semi-supervisado ha sido utilizado para clasificar documentos relevantes en la **realimentación por relevancia seudo** (Huang, 2006, Li, 2008).

1.4 Introducción al aprendizaje activo

En el contexto del aprendizaje automático se plantea la necesidad de encontrar modos de entrenar un clasificador cuando el conjunto de entrenamiento es combinado con un número pequeño de ejemplos etiquetados y un gran número de no etiquetados. Los métodos tradicionales de aprendizaje supervisado o no supervisado no satisfacen resolver tales problemas. En particular, cuando el problema tiene que ver con datos en un espacio de dimensión alta. En años recientes, muchos métodos han sido propuestos, tal es el caso de los algoritmos con aprendizaje semi-supervisado (Epígrafe 1.3.1.3) y los de aprendizaje activo.

El **aprendizaje activo** en la tarea de clasificación puede mejorar el clasificador con menos ejemplos etiquetados de entrenamiento, si se permite elegir activamente los ejemplos de los cuales él aprende (Settles, 2010). En general, el aprendizaje activo toma como entrada un conjunto de entrenamiento de ejemplos etiquetados, así como un conjunto más grande de ejemplos no etiquetados. La salida del algoritmo es un clasificador y un conjunto relativamente pequeño de ejemplos etiquetados, que son añadidos al conjunto de entrenamiento para actualizar el clasificador.

Este tipo de aprendizaje ha sido motivado por el aprendizaje automático. Por ende, su aplicación ha sido enfocada ampliamente en las tareas de clasificación de texto (Lewis, 1994, Lewis, 1995, Liere, 1997, McCallum, 1998a, Nigam, 2000b, Schohn, 2000, Tong, 2001a, Tong, 2001b, Roy, 2001, Zhu, 2003, Hoi, 2006, Zhu, 2008a, Zhu, 2008b, Purpura, 2008). No obstante, con el transcurso del tiempo el estudio del aprendizaje activo ha sido generalizado

a tareas del Procesamiento del Lenguaje Natural (NLP, *Natural Language Processing*, por sus siglas en inglés), entre las que se encuentran:

- Filtrado de información adaptable (Zhang, 2003)
- Extracción de información (Thompson, 1999, Scheffer, 2001, Jones, 2003, Finn, 2003, Culotta, 2006, Settles, 2008b)
- Reconocimiento de entidades (Shen, 2004, Hachey, 2005, Becker, 2005a, Vlachos, 2006, Kim, 2006, Laws, 2008)
- Etiquetamiento de partes de la oración (Dagan, 1995, Argamon-Engelson, 1999, Ringger, 2007)
- Análisis sintáctico (Thompson, 1999, Tang, 2002, Steedman, 2003, Hwa, 2003, Osborne, 2004, Becker, 2005b, Reichart, 2007)
- Desambiguación del significado de la palabra (Fujii, 1998, Chen, 2006, Zhu, 2007, Chan, 2007, Zhu, 2008a, Zhu, 2008b)
- Comprensión del lenguaje hablado (Tur, 2003, Tur, 2005, Wu, 2006)
- Transliteración automática (Kuo, 2006)
- Segmentación de secuencia (Sassano, 2002)
- Recuperación de Información (Xu, 2003, He, 2004, Xie, 2004, Cord, 2005, Shen, 2005, Yu, 2005, Chang, 2005)

De acuerdo al estudio realizado, con el transcurrir de los años se puede apreciar la vinculación del aprendizaje activo con las diferentes tareas del ML y NLP. Es válido destacar, la participación inicial e inmediata del aprendizaje activo en la clasificación de texto, como un foco de estudio permanente.

1.4.1 Protocolos

El aprendizaje activo se refiere a una nueva forma de aprender de los algoritmos de aprendizaje (nombrado en ocasiones **aprendiz**). En este enfoque, el algoritmo de aprendizaje (supervisado) tiene un grado de control sobre los ejemplos en el cual es entrenado, es decir, tiene la autonomía para seleccionar cuáles ejemplos son adicionados a su conjunto de entrenamiento.

El algoritmo de aprendizaje activo puede comenzar con un número pequeño de ejemplos etiquetados, seleccionar cuidadosamente pocos ejemplos adicionales para los cuales solicita etiquetas y aprender del resultado de dicha respuesta. Luego utiliza nuevamente el conocimiento recién ganado para escoger cuidadosamente los próximos ejemplos a seleccionar. De esta manera, el aprendizaje activo pretende alcanzar un alto desempeño en las tareas del aprendizaje automático, usando tan pocos ejemplos etiquetados como sea posible.

En la literatura se describen varios enfoques o protocolos (Figura 7), que reflejan este tipo de aprendizaje, tal es el caso de:

1. Síntesis de consulta de membresía (Angluin, 1988)
2. Muestreo selectivo basado en flujo (Cohn, 1990, Cohn, 1994)
3. Muestreo basado en fondo (Lewis, 1994)

En el primero se consulta aleatoriamente cada ejemplo no etiquetado en el espacio de entrada, para luego ser etiquetado. Generalmente es usado en tareas de aprendizaje de regresión. A menudo es manejable y eficiente para problemas de dominio finito, aunque el etiquetamiento aleatorio de ejemplos puede ser poco práctico para problemas de dominio real si el **oracle**[27] es un anotador humano. En ocasiones, para fines experimentales, el **oracle humano** es simulado usando ejemplos pre-etiquetados, los cuales se ocultan hasta efectuarse la consulta.

Las limitantes que se reflejan en **síntesis de consulta de membresía**, dieron lugar a los restantes protocolos. En el **aprendizaje activo secuencial** (como es llamado a veces el segundo protocolo), los ejemplos no etiquetados salen de la fuente de datos (posiblemente infinita) secuencialmente, uno por uno, y el aprendiz escoge cuál de estos utilizará para el aprendizaje. Típicamente el aprendiz debe decidir en cada uno de ellos si solicitar o no su etiqueta. El empleo limitado de este protocolo en determinados problemas radica en permitir al aprendiz consultar cada ejemplo individualmente, es decir, una vez durante la vida del mismo. No obstante es apropiado, entre otras aplicaciones, en el reconocimiento de voz.

[27] En la mayoría de las aplicaciones, el **oracle** es un humano que se encarga de etiquetar los ejemplos seleccionados.

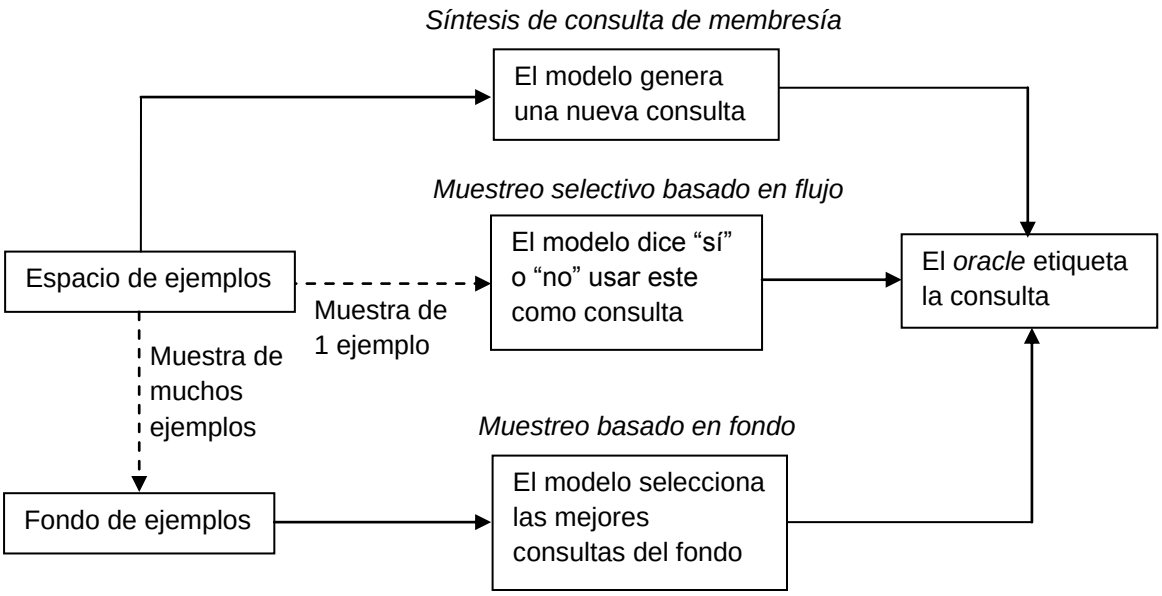


Figura 7: Protocolos del aprendizaje activo.

Una alternativa que ayuda con este problema es el tercer protocolo (Figura 8), que supone que hay un pequeño conjunto de ejemplos etiquetados y un fondo grande de ejemplos no etiquetados disponibles. En este aprendizaje se seleccionan los mejores ejemplos del fondo de datos no etiquetados.

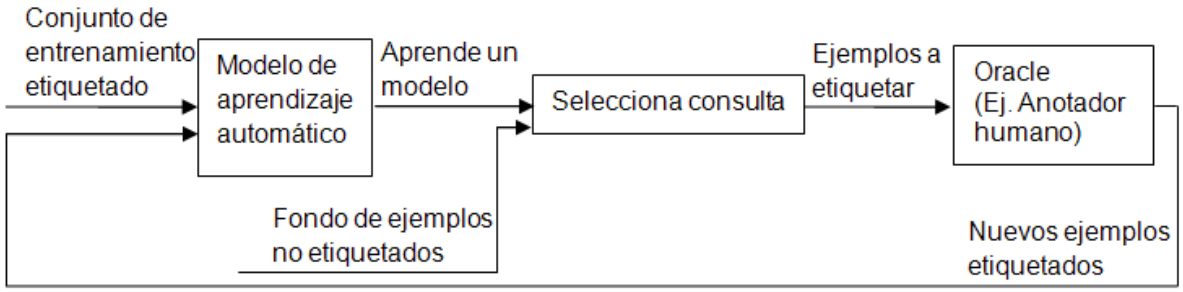


Figura 8: El ciclo de aprendizaje activo basado en fondo.

La diferencia principal entre el aprendizaje activo **basado en flujo** y el **basado en fondo** es que el primero realiza la exploración por los ejemplos secuencialmente y toma decisiones de consulta de manera individual, mientras que el otro evalúa y ordena el fondo completo de ejemplos no etiquetados antes de seleccionar las mejores consultas. El protocolo **basado en fondo** ha sido mucho más común entre aplicaciones de documentos, específicamente en tareas de clasificación de texto (McCallum, 1998a, McCallum, 1998b, Roy, 2001, Tong, 2001b, Hoi, 2006, Purpura, 2008) y extracción de información (Thompson, 1999, Scheffer,

2001, Finn, 2003, Settles, 2008b). No obstante, pueden imaginarse ajustes donde el acercamiento **basado en flujo** es más apropiado. Por ejemplo, cuando la memoria o el poder de procesamiento pueden ser limitados, como con los dispositivos móviles e incrustados.

1.4.2 Estrategias de consultas

El punto clave del aprendizaje activo se encuentra en el criterio de selección de muestra a etiquetar. De ahí, que en todos los protocolos anteriormente explicados se emplee el término **consulta** como la evaluación de información de los ejemplos no etiquetados. En tal sentido, ha sido expuestas en la literatura diferentes alternativas para formular las estrategias de consultas, tal es el caso de:

1. Consulta por comité (Seung, 1992, Freund, 1997b)
2. Muestra de incertidumbre (Lewis, 1994)
3. Reducción de la varianza (MacKay, 1992, Cohn, 1996)
4. Cambio esperado del modelo (Cohn, 1996)
5. Reducción esperada del error (Roy, 2001)

1.4.2.1 Consulta por comité

La **consulta por comité** (QBC, *Query By Committee* por sus siglas en inglés) contiene un comité de modelos de aprendizaje $C^* = \{\theta^{(1)}, \dots, \theta^{(C)}\}$, los cuales son entrenados en el conjunto actual de ejemplos etiquetados T, pero representando hipótesis a competir. Cada miembro del comité tiene permitido votar en las etiquetas de consultas candidatas. Se consideran consultas más informativas aquellos ejemplos en los cuales los miembros del comité discrepan más.

El algoritmo de la estrategia QBC (Anexo 4) para dos comités es propuesto por primera vez en (Seung, 1992) y extendido posteriormente en (Freund, 1997b) para un escenario de **muestreo selectivo basado en flujo**. En cada iteración el algoritmo obtiene aleatoriamente un ejemplo x no etiquetado. Luego llama al **algoritmo Gibbs** dos veces y compara las dos predicciones para la etiqueta de x. Si las dos predicciones son iguales, rechaza el ejemplo y procede a la próxima iteración. En caso contrario, solicita la etiqueta correcta de x y adiciona el ejemplo etiquetado al conjunto de entrenamiento.

Los modelos que se construyen en el comité representan diferentes regiones del espacio de versión (Figura 9), el cual es una representación de toda la información contenida en los

ejemplos observados por el algoritmo de aprendizaje. El porcentaje que se corresponde con las disminuciones en el tamaño del espacio de versión se considera una medida buena del progreso del proceso de aprendizaje. De ahí, que la premisa fundamental en la estrategia QBC sea minimizar dicho espacio, que es el conjunto de hipótesis que son consistentes con el etiquetado actual de los ejemplos de entrenamiento.

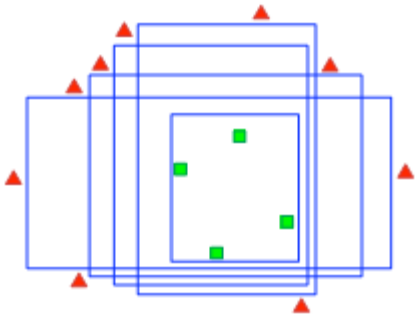


Figura 9: Ejemplo de regiones del espacio de versión. Todas las hipótesis son consistentes con los datos de entrenamiento etiquetado en T (como se indican en las formas de polígonos), pero cada una representa un modelo diferente en el espacio de versión.

La estrategia QBC reduce de manera exponencial el número de ejemplos etiquetados necesario para adquirir un aprendizaje deseado, comparado con la selección aleatoria de ejemplos de entrenamiento. Sin embargo, suele fallar con frecuencia al perder tiempo eliminando áreas que no tienen efecto en la tasa de error. Además, la propuesta inicial de QBC (Seung, 1992) está sobre el marco de trabajo de **muestreo selectivo basado en flujo**, donde cada instancia es considerada solamente una vez para decidir si consultar por su etiqueta o no. Esto es significativo cuando hay un número grande de ejemplos no etiquetados.

Para determinar el nivel de desacuerdo, en (Dagan, 1995, Argamon-Engelson, 1999) se propone la medida de **entropía de voto** en la tarea de **etiquetamiento de partes de la oración**. El método en general es para entrenar clasificadores probabilísticos. Selecciona para el entrenamiento solamente los ejemplos más informativos en un flujo de ejemplos no etiquetados.

En (Liere, 1997) se describe un método de aprendizaje activo para la categorización de texto. Emplea un comité de clasificadores basado en *Winnow.* Para determinar la predicción del comité se utiliza la **votación por mayoría**. Más adelante, en (Liere, 1998) se incorpora a la experimentación el algoritmo de aprendizaje activo pero con un comité de clasificadores

basado en *Perceptron*, el cual resultó superior al de *Winnow* y necesita un número inferior de ejemplos etiquetados para el entrenamiento.

Por otra parte, en (McCallum, 1998a) se modifica el método QBC al usar un fondo grande de documentos no etiquetados para mejorar la clasificación de texto cuando los datos de entrenamiento etiquetados son escasos. Para medir el grado de desacuerdo del comité consideran la **entropía de voto**, empleada por (Dagan, 1995), pero al no contemplar la confianza de los miembros del comité y no arrojar un buen desempeño para la clasificación de documentos se propone como medida de desacuerdo la divergencia de *Kullback-Leibler* (KLD (Kullback, 1951), *Kullback-Leibler Divergence*, por sus siglas en inglés). Además, para la selección de documentos no sólo se tiene en cuenta el desacuerdo de clasificación, sino la **densidad** de la región alrededor de un documento. En la propuesta combinan el aprendizaje activo con el algoritmo de Maximización de la Expectativa (EM, *Expectation-Maximization*, por sus siglas en inglés), con el fin de completar las etiquetas de aquellos documentos que permanecen no etiquetados.

Desde otra perspectiva, en (Abe, 1998) se combina la idea de QBC con los métodos de aprendizaje de ensamble *boosting* (Freund, 1997a) y *bagging* (Breiman, 1996) para construir los comités, los cuales se denominaron *Query-by-boosting* y *Query-by-bagging.* Posteriormente, en (Melville, 2004) se considera que para un efectivo aprendizaje activo es crucial que el comité esté hecho de hipótesis consistentes que sean muy diferentes una de la otra. De ahí que en el artículo se proponga el algoritmo *ActiveDecorate*, el cual usa un comité de *Decorate*[28] para seleccionar buenos ejemplos de entrenamiento. Para evaluar la utilidad esperada de los ejemplos no etiquetados se utilizan los márgenes como en (Abe, 1998), pero en este caso lo definen como la diferencia entre la probabilidad más alta y la segunda más alta predicha por el comité. Los resultados experimentales con un fondo de datos no etiquetados demuestran que en general *ActiveDecorate* supera a *Query-by-boosting* y *Query-by-bagging.*

Más adelante, en (Melville, 2005) se plantea el uso de la divergencia de Jensen-Shannon (JSD, *Jensen-Shannon Divergence*, por sus siglas en inglés) como una medida de desacuerdo o de la utilidad de adquirir ejemplos etiquetados para aprender estimaciones de probabilidad de clase. De manera general, mide la distancia entre dos distribuciones de

[28] Es un método de ensamble que construye un comité diverso, en el que cada hipótesis es tan diferente como sea posible usando datos de entrenamiento artificiales.

probabilidad, la cual puede ser generalizada para medir la distancia entre un número finito de distribuciones. Es una extensión natural de KLD para un conjunto de distribuciones. KLD está definida entre dos distribuciones, y JSD, de un conjunto de distribuciones es la media de KLD de cada distribución a la media del conjunto. En los experimentos realizados demostraron que JSD efectivamente captura la incertidumbre de la estimación de probabilidad de clase y permite identificar ejemplos particularmente informativos, lo cual mejora significativamente los modelos de estimación de distribución de clase. Posteriormente, en (Settles, 2008b) se modificó la medida JSD para el **etiquetamiento de secuencia**.

En los artículos (Bachrach, 2003, Bachrach, 2005) se extienden los resultados de (Bachrach, 2002) y se muestra que es posible implementar el algoritmo QBC para separadores lineares de aprendizaje con una complejidad que depende solamente del número de consultas realizadas, por tanto, es aplicable con *kernels* también. La técnica que proponen la denominan *Kernel Query By Committee* (KQBC), que permite dar un paso amplio en la implementación aplicable de QBC, ya que una implementación *naive* (ejemplo, K-Vecino más cercano (K-NN, *K-Nearest Neighbors*, por sus siglas en inglés)) de QBC tiene un tiempo de complejidad irracional. Sin embargo, KQBC es capaz de aprender activamente en problemas de gran escala usando el protocolo de **muestreo selectivo**. En el trabajo se demuestra como KQBC trabaja en tareas de clasificación binaria, tales como clasificación lineal sintética y de imágenes de cara para la etiqueta femenina o masculina. En ambos casos, el algoritmo KQBC aprende más rápido y es superior al de SVM.

En (Purpura, 2008), se desarrollaron dos enfoques para el aprendizaje activo. Cada uno extiende el algoritmo QBC para la categorización de texto y utiliza un escenario **basado en fondo**. La primera extensión explota la naturaleza jerárquica de un algoritmo de clasificación que se llamó *Hierarchical Query By Committee* (HQBC), y la segunda *Hierarchical Query by Committee by Clustering* (HQBCBC) construye un HQBC. A partir de los experimentos realizados, ambos métodos exceden al desempeño del QBC estándar.

No existe un consenso general en la literatura sobre el tamaño apropiado del comité a usar, el cual puede de hecho variar por tipo de modelo o aplicación. Sin embargo, hasta pequeños tamaños de comité (2 o 3) han demostrado en la práctica trabajar bien (Seung, 1992, McCallum, 1998a, Settles, 2008b).

Los investigadores de este ámbito proponen hacer los métodos más exactos. Una posibilidad es aumentar el número de hipótesis que dan comportamientos predictivos diferentes. Por

otro lado, enfatizan en la necesidad de encontrar un modo de manejar el **ruido** y el hecho de que los datos no son linealmente separables. Además, en el contexto de la recuperación de información sugieren su adaptación.

1.4.2.2 Muestra de incertidumbre

La estrategia de consulta **muestra de incertidumbre** se considera la más simple y comúnmente usada. Es una heurística alternativa de QBC que contempla un solo clasificador o modelo de aprendizaje. El modelo a utilizar no solo toma decisiones de clasificación sino que estima además una certeza, la cual puede ser utilizada para seleccionar ejemplos. En otras palabras, la estrategia puede ser usada con cualquier tipo de clasificador que prediga una clase y suministre una medida de cuán cierta esta predicción es. Tal es el caso de clasificadores probabilísticos, *fuzzy*, vecino más cercano, y neuronales, junto con muchos otros, que satisfagan este criterio o puede ser fácilmente modificado para hacerlo así (Lewis, 1994).

Los clasificadores que utilizan **muestra de incertidumbre** deben asociar un valor de certeza en cada una de sus clasificaciones, que indica la confianza (certeza) del sistema que el ejemplo pertenece a la clase asignada. Los valores de certeza típicamente se encuentran en el rango de 0 a 1: 0 indica que el sistema está seguro de que el ejemplo no pertenece a la clase en cuestión; 1, que está completamente seguro de que el ejemplo pertenece a la clase. En cada iteración del proceso de aprendizaje activo, los valores de certeza de cada ejemplo son calculados y aquellos que son más inciertos o menos confiables son seleccionados para el etiquetamiento. La filosofía detrás de este acercamiento es que un mejor clasificador puede ser construido reduciendo la incertidumbre en el conjunto de datos. (Hu, 2008)

En cada iteración del algoritmo, la versión actual del clasificador es aplicada a cada ejemplo no etiquetado del fondo. Luego se escogen los $b/2$ ejemplos más cercanos a 0.5 por encima de su valor, y los $b/2$ ejemplos más cercanos a 0.5 por debajo de su valor, donde b es el número de ejemplos seleccionados en cada iteración. Se escogen los más cercanos a 0.5 porque este valor corresponde al clasificador de más incertidumbre para asignar una etiqueta. (Anexo 5) Concretamente, en esta estrategia se consultan los ejemplos que para el modelo actual es **menos confiable** en su etiquetamiento más probable (Settles, 2008b).

Muestra de incertidumbre es específicamente significativa para usarlo en la clasificación de texto (Lewis, 1994, Segal, 1994, Schohn, 2000). No obstante, su aplicación ha sido extensiva a algunas tareas del Procesamiento del Lenguaje Natural, como **análisis sintáctico**

(Thompson, 1999, Tang, 2002, Osborne, 2004, Becker, 2005b), **extracción de información** (Settles, 2008b), **segmentación de documentos** (Carvalho, 2004, Settles, 2008b) y **etiquetamiento de partes de la oración** (Ringger, 2007).

La estrategia de (Lewis, 1994), a pesar de mostrarse completamente eficaz, puede incurrir en altos costos computacionales. Especialmente, cuando corre en grandes volúmenes de ejemplos no etiquetados que el algoritmo de aprendizaje tiene que reclasificar durante cada iteración de aprendizaje. Además, al igual que QBC, falla con frecuencia seleccionando ejemplos que se consideran informativos por el aprendiz pero contienen poca información acerca del resto de la distribución de ejemplos. En la literatura de este ámbito, dichos ejemplos seleccionados suelen llamarse ***outliers*** (en inglés).

1.4.2.3 Reducción de la varianza

La **reducción de la varianza** se encuentra motivada por el resultado de (Geman, 1992), el cual muestra que el error de generalización puede ser descompuesto en tres componentes:

$$E_{T[(\hat{y}-y)^2|x]} = E[(y - E[y|x])^2] + (E_L[\hat{y}] - E[y|x])^2 + E_L[(\hat{y} - E_L[\hat{y}])^2] \qquad 2.2$$

El primer término en el miembro derecho de la ecuación 2.2 es el **ruido** (*noise*, en inglés), que es la varianza de la etiqueta correcta y dado x, el cual no depende del modelo o los datos de entrenamiento. El segundo término es la **predisposición** (*bias*, en inglés) elevada al cuadrado, el tercero es la **varianza** del modelo. Estos dos últimos términos comprenden el significado del error cuadrado del aprendiz con respecto a la función de regresión $E[y|x]$. Cuando el segundo término es cero, se dice que el aprendiz no tiene predisposición (*unbiased*, en inglés). Se asume que los clasificadores son aproximadamente *unbiased*, es decir, su *bias* al cuadrado es insignificante cuando se compara con el error cuadrado. La estrategia de consulta **reducción de la varianza** se enfoca en algoritmos que minimicen el error del aprendizaje minimizando su varianza:

$$\sigma_{\hat{y}}^2 \equiv \sigma_{\hat{y}}^2(x) = E_L[(\hat{y} - E_L[\hat{y}])^2] \qquad (2.3)$$

Cuando una nueva entrada $\tilde{x}$ se selecciona, se consulta y el resultado $(\tilde{x}, \tilde{y})$ se adiciona al conjunto de entrenamiento, $\sigma_{\hat{y}}^2$ debería cambiar, la cual se denota como:

$$\tilde{\sigma}_{\hat{y}}^2 = E_{L\cup(\tilde{x},\tilde{y})}\left[\sigma_{\hat{y}}^2|\tilde{x}\right] \qquad (2.4)$$

En el algoritmo de la estrategia se elige como los más informativos a los ejemplos que más reducen la varianza del modelo, y por ende, de manera indirecta reducen el error de

generalización. Este enfoque se aplica solamente a tareas de regresión y al protocolo **síntesis de consulta de membresía**. Sin embargo, para la clasificación de texto se emplearon métodos similares en (Zhang, 2000, Hoi, 2006, Schein, 2007).

A diferencia de la estrategia de la **reducción esperada del error***,* se percibe como ventaja que el modelo no necesita ser re-entrenado. No obstante, su empleo práctico presenta algunas desventajas en términos de complejidad computacional. Cuestión que intentan eliminar algunos investigadores al reducir el espacio de dimensionalidad (Paass, 1995, Hoi, 2006, Settles, 2008b). Aun así, los métodos que sustentan esta estrategia todavía son empíricamente mucho más lentos que otras más simples, como **muestra de incertidumbre**.

1.4.2.4 Cambio esperado del modelo

En el enfoque **cambio esperado del modelo** se selecciona el ejemplo que daría el mejor cambio al modelo actual si se conociera su etiqueta. Un ejemplo de estrategia de consulta en este enfoque es la Longitud del Gradiente Esperada (EGL, *Expected Gradient Length*, por sus siglas en inglés) para tipos de modelos probabilísticos discriminatorios. Fue introducido por (Settles, 2008a) para el aprendizaje activo en el escenario de instancia–múltiple (MI, *multiple-instance*, por sus siglas en inglés), para tareas de clasificación de texto y recuperación de imágenes basado en contenido.

En un problema de aprendizaje MI, las instancias son organizadas en bolsas en vez de instancias individuales, que son las utilizadas para el etiquetamiento y entrenamiento del modelo. Una bolsa es etiquetada negativa sí y solamente sí todas sus instancias son negativas. Sin embargo, una bolsa es positiva si al menos una de sus instancias es positiva. Fue formalizado por (Dietterich, 1997) y desde entonces ha sido aplicado a una variedad de tareas, lo que incluye la recuperación de imágenes basada en contenido (CBIR, *Content-Based Image Retrieval*, por sus siglas en inglés) (Maron, 1998, Andrews, 2003, Rahmani, 2006) y clasificación de texto (Andrews, 2003, Ray, 2005).

Para estas tareas de aprendizaje MI es posible obtener etiquetas al nivel de bolsa y directamente al de instancia. Este enfoque es útil cuando las etiquetas de bolsa son fácilmente adquiridas y las etiquetas de instancias pueden ser obtenidas, pero son costosas.

La estrategia de consulta EGL en un escenario MI considera identificar la instancia que debería impartir mejor cambio al modelo actual, si se conociera su etiqueta. Por lo general,

realiza el entrenamiento usando la optimización basada en el gradiente, el cambio impartido al modelo puede ser medido por la longitud del gradiente de entrenamiento.

Durante la ejecución de la estrategia, el aprendiz debería consultar la instancia x que, de etiquetada y ser adicionada al conjunto de entrenamiento T, resultaría el nuevo gradiente de entrenamiento de la magnitud más grande. Por $\nabla\ell(L;\theta)$ se denota el gradiente de la función objetivo ℓ respecto a los parámetros del modelo θ. Ahora, $\nabla\ell(L \cup \langle x,y\rangle;\theta)$ es el nuevo gradiente que debería ser obtenido al adicionar $\langle x,y\rangle$ a L. Como el algoritmo de consulta no conoce de antemano la etiqueta verdadera y, se debe calcular la longitud como una expectativa sobre todos los etiquetamientos posibles,

$$\phi^{EGL}(x) = \sum_i P(y_i|x)\|\nabla\ell(L \cup \langle x,y_i\rangle;\theta)\| \qquad (2.7)$$

donde $\|\nabla\ell(L \cup \langle x,y_i\rangle;\theta)\|$ es la norma Euclidiana de cada vector de gradiente que resulta. La intuición detrás de esta estrategia es que preferirá los ejemplos que probablemente influirán más en el modelo (o tendrán el mayor impacto en sus parámetros), sin tener en cuenta la etiqueta de consulta que resulta. Este enfoque puede ser computacionalmente costoso, si el espacio de rasgos y el conjunto de etiquetamientos es muy grande.

Más adelante, en (Druck, 2009) se propone un enfoque de aprendizaje activo en el que se solicita etiqueta para rasgos, en vez de instancias. Experimentan para dos tareas de etiquetamiento de secuencia y demuestran que el etiquetamiento de rasgos activo es más efectivo que el de rasgos pasivos, así como del aprendizaje activo tradicional con instancias.

1.4.2.5 Reducción esperada del error

Una estrategia más directa es la **reducción esperada del error** propuesta por (Roy, 2001) para la clasificación de texto con el protocolo **basado en fondo**, usando Naïve Bayes. El algoritmo mide cuánto el error de generalización será probablemente reducido. Para ello, estima la pérdida (0/1 o log) que debería resultar al adicionar el ejemplo candidato $x \in U$, con su etiqueta y a T. El ejemplo candidato que causa el error esperado más bajo es el seleccionado para el etiquetamiento. Al usar la pérdida de log, el algoritmo elegiría,

$$\tilde{E}_{\hat{P}_{T^*}} = \frac{1}{|U|}\sum_{x\in U}\sum_{y\in Y}\hat{P}_{T^*}(y|x)\log\left(\hat{P}_{T^*}(y|x)\right) \qquad (2.5)$$

y para la pérdida 0/1 sería,

$$\tilde{E}_{\hat{P}_{T^*}} = \frac{1}{|U|}\sum_{x\in U} 1 - \max_{y\in Y} \hat{P}_{T^*}(y|x) \qquad (2.6)$$

donde Y es el conjunto de etiquetas y $\hat{P}_{T^*}(y|x)$ es la estimación sobre todas las posibles etiquetas dentro del modelo candidato entrenado en $T + (x, y)$, que se considera T^*. El algoritmo general de la estrategia se encuentra en el Anexo 6.

En muchos casos, desafortunadamente este enfoque de consulta lleva un aumento drástico del costo computacional. No solamente requiere estimar la reducción de error esperado sobre U para cada consulta, sino un nuevo modelo debe ser incrementalmente re-entrenado para cada posible etiquetamiento de consulta, que por su parte itera sobre el fondo entero. A causa de esto, las aplicaciones de esta estrategia son poco prácticas y se han considerado solamente en tareas simple de clasificación binaria. La estrategia ha sido aplicada en tareas de clasificación de texto a partir de su combinación con el aprendizaje semi-supervisado (Zhu, 2003), supervisado (Moskovitch., 2007), y otros como (Guo, 2007).

1.4.3 Análisis general de las estrategias de consulta

Un importante obstáculo en el aprendizaje activo es la posibilidad de adquirir las etiquetas de ejemplos ***outlier*** o con **ruido** (por ejemplo, ejemplos mal etiquetados por el **oracle humano** al realizarlo de manera distraída). Este problema está presente especialmente en **muestra de incertidumbre**, **consulta por comité** y **cambio esperado del modelo**. Según (Hu, 2010), **muestra de incertidumbre** por ser la estrategia de consulta más simple y popular, resuelve este problema combinando la estrategia con información de densidad (Fujii, 1998, Nguyen, 2004, Settles, 2008b); con información de diversidad (Brinker, 2003, Dagli, 2005, Osugi, 2005, Shen, 2005) o con ambas (Shen, 2004, Xu, 2007).

De acuerdo a la literatura consultada, **muestra de incertidumbre**, QBC y sus variantes son atractivas debido a su aplicación en el aprendizaje automático, pero en ocasiones poco robustos al consultar ***outliers***. En cambio, la **reducción de la varianza**, **cambio esperado del modelo** y **reducción esperada del error** son bastante robustos en muchas situaciones, pero de muy alto costo computacional.

1.4.4 Aprendizaje activo en la recuperación de información

Los sistemas de recuperación de información tradicionales tratan de estimar la información que beneficia al usuario a partir de la consulta inicial. En ocasiones, el resultado recuperado puede no corresponderse con la necesidad del usuario y obligarlo a continuar consultando en

una larga sesión de búsqueda. Con el tiempo se ha intentando mejorar la efectividad del sistema de recuperación con la incorporación del proceso de **realimentación por relevancia**. Según (Xu, 2007), existen dos problemas principales cuando se emplea la realimentación por relevancia. El primero es cómo seleccionar el primer conjunto de documentos para ser presentado al usuario para la realimentación, el segundo es cómo efectivamente utilizar la información de realimentación por relevancia para reformular la consulta.

Muchas de las investigaciones realizadas en este ámbito se enfocan en el segundo problema, actualizando la consulta. También ha sido beneficiada a partir de la incorporación de técnicas de clasificación como tarea del aprendizaje supervisado, pero centrándose en refinar el clasificador más que la consulta. Recientemente, el aprendizaje activo ha ocupado un lugar importante en la Recuperación de Información, relacionándose con el primer problema de la realimentación por relevancia. De manera general, la incorporación de este tipo de aprendizaje tiene como objetivo resolver el problema de investigación: qué documentos presentar al usuario de modo que la realimentación del usuario en los documentos pueda impactar considerablemente el desempeño de la realimentación por relevancia (Xu, 2007). El aprendizaje activo es usado para maximizar la exactitud de los SRI mientras se minimiza la cantidad de realimentación requerida (Tian, 2011).

Un trabajo interesante en aplicar el aprendizaje activo a la realimentación por relevancia es el de (Siegelmann, 2001), donde un usuario es asumido para iterativamente escoger grupos, y la responsabilidad del aprendizaje activo es diseñar buenos grupos a partir del sistema. El proceso de recuperación es inicializado por el usuario que envía una consulta, para que luego el sistema encuentre un pequeño subconjunto de grupos a presentar al usuario, junto con sus resúmenes. El sistema espera hasta que el usuario selecciona uno o varios de los grupos presentados, para luego usar como evidencia las selecciones y actualizar la distribución sobre los documentos o evaluación de relevancia. La salida está dada por los documentos ordenados por sus pesos, y la iteración continúa hasta que termine el usuario o el sistema. En cada iteración, el sistema presenta al usuario una consulta y éste responde el grupo que es más relevante para él.

Más tarde, en (Siegelmann, 2005), se propone la metodología Recuperación de Información Activa (AIR, *Active Information Retrieval*, por sus siglas en inglés). Un sistema AIR toma el rol

activo de hacer preguntas[29] al usuario para clarificar las necesidades de información, disminuir las sesiones de búsquedas y aumentar la efectividad del sistema. El sistema considera el tipo de consulta inversa explicado en (Siegelmann, 2001), donde el sistema escoge un conjunto de grupos a ser presentado al usuario, y éste elige el grupo, que será incluido en el conjunto de consulta que es más relevante para el usuario. En el algoritmo activo utilizan la estrategia de consulta **muestra de incertidumbre**, que elige la consulta inversa que se espera que disminuirá más la incertidumbre sobre el vector de relevancia. Miden el nivel de incertidumbre del sistema sobre la necesidad del usuario, usando la función de entropía. Además optimizan la selección del conjunto de grupos con un método de aproximación que satisfactoriamente encuentra el mejor próximo grupo a incluir en el conjunto de grupos. Como trabajo futuro desean implementar el sistema en Bioinformática, donde los documentos en la colección no tienen una métrica natural de similitud entre ellos porque los documentos pueden ser imágenes, textos o ficheros de genes. Además, quisieran considerar consultas inversas más sofisticadas, tales como preguntar al usuario para ordenar por la relevancia a los grupos.

En (Onoda, 2002, Onoda, 2005) se utiliza la realimentación por relevancia con el algoritmo de clasificación SVM, pero desde el punto de vista del aprendizaje activo. Los documentos etiquetados como relevantes y no relevantes por el usuario, desde una lista de documentos recuperados, son considerados ejemplos positivos y negativos, respectivamente, y a su vez, utilizados para generar un clasificador SVM. Luego produce un nuevo orden, que recupera documentos relevantes con una calidad superior a la recuperación original. En este enfoque, la realimentación por relevancia es transportada a un problema de clasificación binaria. En este artículo no se discute teóricamente cómo realizan la selección de documentos que influyen en el aprendizaje efectivo y en el desempeño de la RI. Sin embargo, dejan claro en la experimentación realizada con el conjunto de artículos "*Los Angels Times*" (es usado en TREC), que el SRI con la realimentación de SVM activa es superior al desempeño de los SRI clásicos sin realimentación y con realimentación por relevancia basado en el algoritmo *Rocchio*. Para la obtención de estos resultados utilizaron las medidas de evaluación **precisión** y **exhaustividad**.

Posteriormente, en (Onoda, 2004, Onoda, 2006) se presenta un método efectivo de realimentación nombrado **realimentación de no-relevancia**, que aprovecha solamente la

[29] Llaman a las consultas realizadas por el sistema al usuario **consultas inversas**.

información de los documentos no relevantes. Consideran la situación de recuperación como un problema de clasificación SVM de una clase. En la experimentación se emplea el mismo conjunto de artículos de (Onoda, 2002) y comparan la propuesta con un SRI sin realimentación por relevancia y con uno con realimentación basada en Rocchio. El algoritmo propuesto es el de mayor desempeño en términos del número de iteraciones para recuperar documentos relevantes. De ahí que concluyan que el enfoque basado en SVM de una clase es muy útil para la recuperación de documentos, usando solamente la información de los documentos no relevantes. Por otra parte, el SRI sin realimentación se comporta mejor que el de realimentación basada en Rocchio, ya que este último no trabaja bien cuando el sistema recibe solamente la información de los documentos no relevantes. Dejan como trabajo futuro, al igual que en (Onoda, 2002), la explicación teórica de cómo realizan la selección de ejemplos.

Sin embargo, en (Onoda, 2005) no se tienen en cuenta solo los documentos no relevantes, sino también los relevantes. Además, se adoptan varias representaciones del modelo SVM y reglas para mostrar los documentos en cada iteración. Acorde a los experimentos realizados, se opta por una representación binaria de SVM, y la regla según la cual los documentos que son discriminados relevantes y se encuentran en el área de margen de SVM son mostrados al usuario. Recientemente, en (Donmez, 2009) se expone un método activo para el aprendizaje de SVM, que se basa en la relación propuesta por (Steck, 2007).

En el trabajo de (Xu, 2003) se introduce un nuevo enfoque de Realimentación por Relevancia Híbrido usando un SVM (HRFSVM, *Hybrid Relevance Feedback approach using a Support Vector Machine*, por sus siglas en inglés), que selecciona activamente de un fondo de documentos no etiquetados un número pequeño de documentos para solicitar las etiquetas correspondientes al usuario. Utilizan la estrategia de consulta **muestra de incertidumbre** para escoger los documentos a etiquetar del fondo, pero el esquema **muestra de margen** (Scheffer, 2001), por ser el más simple en términos de costo computacional. El sistema presenta al usuario los documentos más lejanos y más cercanos del lado positivo del hiperplano de SVM. En la experimentación se utiliza el conjunto de datos *Reuters-21578* y para determinar la efectividad del proceso de realimentación se emplea la medida de **precisión**. Los resultados muestran que HRFSVM se desempeña significativamente mejor que el algoritmo de Realimentación por Relevancia SVM (SVMRF, *SVM Relevance Feedback*, por sus siglas en inglés) y de Realimentación por Relevancia SVM Activo

(ActiveSVMRF, *Active SVM Relevance Feedback*, por sus siglas en inglés). El método propuesto es una heurística que para el futuro proponen mejorar con tecnología de optimización.

En (Shen, 2003, Shen, 2005) se estudia el problema de realimentación por relevancia activa, donde el sistema de recuperación se responsabiliza de escoger activamente los mejores documentos para la realimentación por relevancia. Para ello, derivaron la estrategia de consulta **muestra de incertidumbre** y los algoritmos prácticos *Top K, Gapped Top K* y *K Cluster Centroid*. La evaluación realizada con los conjuntos de datos TREC2003 HARD, AP88-89 y AP90 arrojó que tanto *Gapped Top K* como *K Cluster Centroid* superan al algoritmo *Top K* con pocos documentos relevantes juzgados, sugiriendo que la diversidad en los documentos presentados es una propiedad deseable. No obstante, tienen interés de explorar en varias direcciones, tales como, aprender de **documentos no relevantes** juzgados por el usuario para hacer uso completo del esfuerzo del usuario y la realimentación; explorar otras estrategias para seleccionar documentos y, por último, tratar de combinar la realimentación seudo y activa. El TREC HARD Track (Allan, 2003) ha simulado algunos trabajos a lo largo de la línea de realimentación activa que incluye (Robertson, 2003, Shen, 2003).

En (Xu, 2007) se presenta el enfoque de aprendizaje activo nombrado *Active-RDD* (*Active Learning to achieve Relevance, Diversity and Density*, por sus términos en inglés). Escogen los documentos de mayor relevancia, los que maximizan la distancia entre el nuevo documento y los seleccionados, pero con altas regiones de densidad. La forma en que determinan la diversidad no es igual a la de los algoritmos *Gapped Top k* y *Cluster Centroid* en (Shen, 2003, Shen, 2005, Shen, 2007). En el algoritmo *Active-RDD* el factor de relevancia se determina a partir de la medida *KL-divergence,* que es dada por la primera ronda de búsqueda como el *score* de relevancia. El factor diversidad es determinado a partir del método que mide la distancia mínima entre el documento y algún otro documento del conjunto (se corresponde al método *single linkage* en el agrupamiento jerárquico). La densidad es la media de J-divergence entre un documento y todos los otros documentos, que mide el grado de solapamiento entre un documento y todos los otros. Con el fin de combinar los tres factores, realizan una combinación lineal de sus medidas respectivas.

El algoritmo *Active-RDD* extiende el algoritmo de Relevancia Marginal Máxima (MMR, *Maximal Marginal Relevance*, por sus siglas en inglés), que permite ordenar los documentos

por su grado de relevancia y evitar al mismo tiempo la redundancia. Esta extensión se sugiere en (Shen, 2005) pero no es implementada. Para la experimentación emplearon el conjunto de datos *TREC HARD 2005 Track* y *TREC HARD 2003 Track* y Lemur como SRI. Compararon el desempeño del algoritmo *Active-RDD* con los de realimentación activa *Top K,Gapped Top K* y *Cluster Centroid* (Shen, 2005) a través de las medidas de evaluación *Mean Average Precision* (MAP) y precisión, empleadas por (Shen, 2005). El resultado experimental muestra que el algoritmo propuesto se desempeña significativamente mejor que el resto. En el artículo se sugiere mejorar la realimentación por relevancia haciendo uso completo de la realimentación de usuario a través del aprendizaje de **documentos no relevantes**, y combinar la **realimentación implícita** con el aprendizaje activo. Más tarde, en (Xu, 2008) se propone un algoritmo de realimentación activa incorporando la primera sugerencia propuesta en (Xu, 2007).

En la generación de palabras claves para anuncios en la Web (Wu, 2009) se elabora un modelo interactivo para explorar la realimentación por relevancia. Utilizan un modelo de regresión para entrenarlo con los términos etiquetados y predecir los resultados de relevancia de los candidatos no etiquetados. Retornan al usuario para etiquetar los términos seleccionados con el enfoque de aprendizaje activo TED (*Transductive Experimental Design*, por sus siglas en inglés) presentado en (Yu, 2006). Tiende a seleccionar términos no etiquetados con una predicción dura y representativa. La propuesta de generación de palabras claves puede ser extendida a otras aplicaciones, tales como el agrupamiento de términos. Hasta el momento, los términos de una sola palabra son considerados. De ahí que declaren querer explorar el uso de frases como trabajo futuro.

Para TREC 2009, en (Cormack, 2010) se utiliza la búsqueda interactiva y los juicios para encontrar un conjunto grande y diverso de ejemplos de entrenamiento. Luego emplearon el proceso de aprendizaje activo para encontrar los documentos más relevantes.

Por otra parte, en (Xu, 2010) se explora cómo integrar el aprendizaje activo en los métodos de aprendizaje de preferencia, que pueda modelar dependencias de representaciones de vectores de rasgos, así como de relaciones. El objetivo es aprender de manera activa una función de ordenamiento general para reducir el costo de entrenamiento y ordenar correctamente. Existen algunos enfoques para el aprendizaje activo con preferencias, pero la mayoría supone que las entidades (tales como, documentos, páginas Web, productos, canciones, etc.) son independientes una de las otras y no toman en cuenta las relaciones

entre ellas. La estrategia de selección activa se enfoca en la mayor pérdida esperada. Los resultados experimentales muestran para la recuperación Web en TREC 2004 que el método propuesto aprende considerablemente más rápido que algoritmos que no tienen en cuenta la dependencia.

En el artículo de (Tian, 2011) se propone una estrategia de realimentación por relevancia con aprendizaje activo para maximizar la exactitud relativa al esfuerzo del usuario. Se investiga el método de aprendizaje activo de **margen** con el algoritmo de clasificación SVM, que toma los documentos más cercanos a la **frontera de decisión** cuya relevancia es máximamente incierta. Además, presentan una mejora denominada **estructura local**, que captura la idea de que los ejemplos útiles a etiquetar también deberían ser lejanos a los ejemplos ya etiquetados y cercanos a los no etiquetados. La realimentación por relevancia la simularon con las colecciones de prueba Robust04 TREC para mostrar que el acercamiento de aprendizaje activo dominó a varios *baseline* de RF estándares de Rocchio con relación a la cantidad de realimentación proporcionada por el usuario. Las medidas utilizadas para evaluar la efectividad fueron MAP, *Top-10 precision* (P@10) y *Mean Reciprocal Rank* (MRR).

En el ámbito de la recuperación de información, el aprendizaje activo ha sido estudiado con mayor intensidad en la recuperación de imágenes, que para otros tipos de datos, tales como, documentos, canciones y videos. En (Tong, 2001c), se expone la combinación de un algoritmo de aprendizaje activo con el de clasificación SVM denominado SVM_{Active}. El algoritmo en general selecciona las imágenes más informativas para pedirle al usuario que las etiquete en **relevante** y **no relevante** respecto al concepto de consulta del usuario, para rápidamente refinar el concepto de frontera. Luego de concluir las rondas de realimentación por relevancia se recuperan las imágenes más relevantes (o las relevantes) que se encuentran más lejos de la frontera SVM. Durante el aprendizaje activo se consideran imágenes más informativas aquellas que estén más cerca de la frontera. Para ello, aplican en un escenario **basado en fondo** la estrategia de consulta **muestra de incertidumbre**.

En la experimentación utilizan los conjuntos de datos de imágenes: *four-category*, *ten-category* y *fifteen-category* y aplican la medida de evaluación **precisión**. Los resultados experimentales muestran que el algoritmo propuesto adquiere un nivel de exactitud superior al de los esquemas de refinamiento de consulta tradicionales de realimentación por relevancia. Además, obtienen rápidamente el clasificador SVM con pocas imágenes etiquetadas. No obstante, el algoritmo es poco práctico en bases de datos de imágenes

demasiado grandes. Otra limitante es que asumen tener inicialmente al menos una imagen relevante y una no relevante. Estas limitantes las dejan como trabajos futuros, e incluso, comentan algunas de sus posibles soluciones pero no la implementan.

En (Cord, 2004) se presenta el método de aprendizaje activo RETIN AL para la recuperación de imágenes basada en el contenido (CBIR, *Content-Based Image Retrieval*, por sus siglas en inglés). El método propuesto se basa en el principio del algoritmo SVM_{Active} en (Tong, 2001c) pero sin usar la frontera de SVM para encontrar el umbral de las imágenes más cercanas a éste. De hecho, proponen un ajuste adaptativo del umbral, en el que suponen que el mejor umbral se corresponde a la frontera buscada. Con este umbral pueden mostrar tanto imágenes relevantes como no relevantes. De esta manera se considera un umbral bueno, si y solo sí, el conjunto de imágenes seleccionadas se encuentra bien balanceado (entre imágenes relevantes y no relevantes). Esta propiedad es la utilizada para ajustarlo. Además, optimizan el conjunto de imágenes a presentar al usuario para su etiquetamiento. Para ello aplican al conjunto el algoritmo de agrupamiento LBG (Patan-e, 2001) y el sistema selecciona solamente las imágenes más relevantes a cada grupo para mostrar al usuario. La base de datos de imágenes utilizada para el experimento es extraída de la base de datos de fotos COREL. El desempeño del sistema CBIR fue determinado a partir de curvas de **precisión**/**exhaustividad**. Los resultados experimentales muestran que RETIN AL presenta mejores valores de precisión con respecto al algoritmo SVM_{Active} de (Tong, 2001c).

Posteriormente, en (Cord, 2005) se plantea un esquema de aprendizaje activo en el algoritmo de clasificación SVM para la realimentación por relevancia. Combina cada una de las estrategias de consulta **muestra de incertidumbre** y **reducción del error** con la diversidad de selección, haciendo uso del esquema de diversidad de ángulo y una corrección de frontera propuesta para tratar con escasos datos de entrenamiento. Los resultados experimentales muestran la eficiencia de las combinaciones propuestas, en especial la de la estrategia que usa la corrección de frontera, muestra de incertidumbre y diversidad de ángulo. Además, muestran que el tiempo computacional de ésta es significativamente menor con respecto a la otra. En (Cord, 2006) se incorpora al sistema de recuperación de imágenes un aprendizaje semántico a largo plazo, ya que hasta el momento incrementaban el desempeño del sistema, pero solamente durante la sesión de recuperación actual.

En (Cord, 2007, Cord, 2008) se refina la propuesta obtenida hasta el momento, con el fin de presentar un esquema general de aprendizaje activo de RETIN. Las principales

contribuciones conciernen a la corrección de frontera para hacer el proceso de recuperación más robusto y la introducción de un nuevo criterio de selección de imagen que represente mejor el objetivo de la CBIR.

Más adelante, en (Cord, 2009, Cord, 2010) se muestra una estrategia para acelerar el sistema de aprendizaje activo en bases de datos grandes, y así superar la limitante de escalabilidad en este contexto. Para ello, en el paso de la selección de imágenes a mostrar al usuario para el etiquetamiento obtienen imágenes del fondo que son cercanas a las relevantes en el conjunto de entrenamiento. Esto lo realizan usando la búsqueda K NN para todas las imágenes en el conjunto de entrenamiento. Muchas de las imágenes seleccionadas no son buenas, pero pueden ser fácilmente filtradas. Con este fin, determinan la relevancia de cada una de las imágenes seleccionadas usando el clasificador SVM y mantienen solamente las N imágenes más relevantes. Para hacer rápido el proceso de K NN, en vez de hacer una exploración lineal, usan un esquema de indexación eficiente basado en Hashing Localmente Sensible (LSH, *Locally Sensitive Hashing*, por sus siglas en inglés) para la medida de distancia X^2, que permite combinar la búsqueda K NN rápida con el *kernel* basado en la distancia X^2.

La segunda cuestión de escalabilidad que tratan es que seleccionan entre las N imágenes más relevantes, la más incierta y de más diversidad, utilizando la diversidad de ángulo de (Brinker, 2003). Si al usuario cuando le muestran el resultado no está satisfecho, entonces escogen otra incierta de las restantes N imágenes relevantes para ayudar a mejorar la función de relevancia. Los resultados experimentales, en una base de datos de 178 500 imágenes, muestran que el método propuesto es 45 veces más rápido que los que tienen una exactitud similar a él en el estado de arte que emplean el ángulo de diversidad.

En (Ferecatu, 2004, Ferecatu, 2005, Ferecatu, 2007) se extiende el método de aprendizaje activo para la realimentación por relevancia usando SVM, propuesto por (Tong, 2001b, Tong, 2001c). Específicamente, modifican el criterio de selección, que originalmente consiste en seleccionar el documento que está más cercano al hiperplano definido por SVM. Dicho criterio de selección lo denominaron el candidato **más ambiguo** (MA, *Most Ambiguous*, por sus siglas en inglés). La modificación realizada al MA consistió en minimizar la redundancia entre las imágenes candidatas a mostrar al usuario, nombrada finalmente como los candidatos **más ambiguos y ortogonal** (MAO, *Most Ambiguous and Orthogonal*). Los experimentos confirman la efectividad de la propuesta de selección haciendo uso de una

base de datos de imágenes satelitales. En (He, 2004) se elabora un nuevo método de aprendizaje activo, nombrado **espacio de versión medio** (*mean version space,* por sus términos en inglés), que pretende seleccionar la imagen óptima en cada ronda de la realimentación por relevancia.

Se presenta en (Dagli, 2005) un sistema de aprendizaje activo para la recuperación de imágenes basada en la realimentación por relevancia, incorporando la noción de diversidad al conjunto de consulta a mostrar al usuario. Proponen emplear la diversidad angular expuesta en (Brinker, 2003) en un escenario de aprendizaje activo que utiliza el Análisis Discriminante Predispuesto (BDA, *Biased Discriminant Analysis*, por sus siglas en inglés), al cual denominaron BDA basado en diversidad (*diversity-based BDA, en inglés*) o DBDA. Para la experimentación se utiliza una base de datos de imágenes extraída de COREL. A partir de las medidas de desempeño **exhaustividad** y **precisión** muestran en los resultados que el sistema híbrido DBDA es superior al de DBA y SVM_{Active} por un margen amplio en las primeras ronda de realimentación y uno cercano en las últimas rondas, donde DBDA coincide con BDA.

Por otra parte, en (Hoi, 2005) se fusiona el aprendizaje semi-supervisado y el algoritmo SVM con un aprendizaje activo, denominado SVM-SSAL (*SVM-Semi-Supervised Active Learning*, por sus siglas en inglés). En la primera ronda de realimentación, la selección de los ejemplos no etiquetados es realizada a partir de las distancias más grandes de SVM. Lo combinan con el enfoque semi-supervisado con el fin de incrementar el número de ejemplos positivos. En las restantes rondas la selección se realiza aplicando la estrategia **muestra de incertidumbre**, es decir, seleccionan los ejemplos no etiquetados con las distancias más pequeñas de SVM. Posteriormente, aplican el algoritmo de aprendizaje semi-supervisado y muestran las imágenes escogidas por el sistema al usuario. Para la evaluación empírica de la propuesta tomaron imágenes de los CDs de COREL y determinaron el desempeño con la medida MAP. Los resultados experimentales muestran que el algoritmo SVM-SSAL es superior al de TSVM-SAL (*Transductive SVM based active learning*, por sus siglas en inglés) y SAL (*SVMActive Learning*, por sus siglas en inglés).

En (Hoi, 2008) se plantea un nuevo esquema para el aprendizaje activo con el fin de direccionar los problemas principales que posee SVM_{Active}. El primero, relacionado al desempeño, que usualmente está limitado por el número de entrenamiento. Además está diseñado para seleccionar un solo ejemplo en cada iteración de aprendizaje. No obstante,

podría seleccionar imágenes similares al escoger ejemplos múltiples. Más adelante, en (Hoi, 2009) se emplea la matriz de información *Fisher* como medida del modelo de incertidumbre, y escoge el conjunto de ejemplos que efectivamente reduce la información de *Fisher.* Este enfoque lo aplican a la realimentación por relevancia en la CBIR y a la categorización de texto de manera efectiva. Dejan pendiente la extensión de la metodología propuesta a otros problemas de aprendizaje automático y aplicaciones de multimedia.

Se extiende en (Zhou, 2006) la investigación preliminar (Zhou, 2004), al integrar en el método Ssaira (*Semi-Supervised Active Image Retrieval with Asymmetry*, por sus términos en inglés) los méritos del aprendizaje supervisado y aprendizaje activo en el proceso de realimentación por relevancia. Se emplean dos aprendices bastante diversos entre ellos, y con estos determinan las imágenes a mostrar al usuario. Para ello seleccionan las imágenes en que los aprendices discrepan más, o en las que ambos aprendices tienen una confianza baja. A fin de mejorar la confianza, las imágenes etiquetadas no son trasladadas del conjunto de imágenes no etiquetadas al de entrenamiento. Solo se utilizarán como ejemplos de entrenamiento etiquetados temporalmente, porque en la próxima ronda de realimentación serán considerados nuevamente como datos no etiquetados. Por otra parte, adoptan un mecanismo especial para generalizar los ejemplos negativos. Concretamente, los K NN con una distancia euclidiana de ejemplos no etiquetados son identificados para cada ejemplo negativo, y entonces, el vector de rasgos de este k+1 ejemplos son promediados para derivar un ejemplo virtual, que es usado por el aprendiz en lugar del ejemplo negativo original. Los resultados experimentales sobre imágenes de COREL confirman que SSAIRA es el mejor entre los métodos comparados, incluyendo SVM_{Active}. Desean direccionar el problema de muestra de entrenamiento asimétrico a partir de la generalización de los ejemplos negativos usando K NN, ya que se contemplan ejemplos con propiedades similares.

Por otra parte, para direccionar la escalabilidad de la realimentación por relevancia basada en el aprendizaje activo, en (Crucianu, 2008) construyen un M-tree en el espacio de rasgo asociado al *kernel* y los K NN del hiperplano, que es la frontera dada por el SVM son los recuperados. Este método logra recuperar de manera rápida las imágenes más ambiguas. Las evaluaciones desempeñadas en dos bases de datos de imágenes reales de la tierra muestran que la búsqueda es significativamente más rápido con este enfoque, permitiendo la selección en tiempo real de las imágenes retornadas al usuario desde la base de datos. En el

artículo dejan pendiente si la condición de baja redundancia puede ser usada durante el proceso de búsqueda.

En (Liu, 2008) se formula un enfoque diferente para seleccionar las imágenes que se mostraron al usuario. Se valora que puede existir redundancia entre las imágenes más ambiguas y puede afectar negativamente el proceso de recuperación. Para darle solución a este inconveniente proponen un método de aprendizaje no supervisado para agrupar las imágenes cercanas a la frontera de clasificación SVM y seleccionar una imagen por cada grupo para ser etiquetadas por el usuario.

En (He, 2009) se introduce un nuevo algoritmo de aprendizaje de subespacio activo (ASL, *Active Subspace Learning*, por sus siglas en inglés), que activamente selecciona los ejemplos más informativos en el subespacio óptimo, motivado por la conexión entre el aprendizaje de subespacio y la regresión lineal. Utilizan técnicas de Diseño Experimental Óptimo (DEO, *Optimal Experimental Design*, por sus siglas en inglés) para seleccionar las imágenes que minimizan los errores predictivos esperados. Luego solicitan el etiquetamiento de los usuarios para las imágenes seleccionadas y lo usan para el aprendizaje de un subespacio óptimo. Los resultados experimentales en una base de datos de imágenes extraídas de COREL muestran que el algoritmo propuesto supera a tres algoritmos con aprendizaje pasivo y a dos con aprendizaje activo. Con estos resultados demuestran que el aprendizaje activo basado en el Diseño Experimental Óptimo es más efectivo que SVM_{Active} para seleccionar los ejemplos más informativos.

En (Piras, 2011) se presenta un enfoque de aprendizaje activo para el vecino más cercano (NN, *Nearest Neighbor*, por sus siglas en inglés). El método propuesto escoge entre las mejores imágenes aquellas que están en un área no demasiado lejos y no demasiado cerca a la consulta. Aplican un enfoque max-min, que selecciona una imagen de conjunto evaluando todas las distancias entre las imágenes más relevantes y las imágenes del conjunto y escogiendo para cada una de ellas la más corta. Las imágenes son ordenadas acorde a estas distancias y con esto la de máxima distancia es seleccionada y mostrada al usuario.

Por otra parte, en la recuperación de música (Mandel, 2006) se describe un sistema de recuperación para el desempeño flexible de consultas similares de música y emplean la realimentación por relevancia con SVM_{Active} (Tong, 2001b) para aprender un clasificador por cada consulta. El objetivo de la propuesta es minimizar el número de canciones que un

usuario debe etiquetar para una consulta. En los resultados experimentales, un subconjunto de la colección *uspop2002* muestra que una selección inteligente de ejemplos requiere durante el aprendizaje activo la mitad como ejemplos etiquetados para adquirir la misma exactitud que un esquema estándar.

En (Chen, 2008) se desarrolla un nuevo esquema para la recuperación de música basada en contenido (CBMR, *content-based music retrieval*, por sus siglas en inglés). Se utiliza la realimentación por relevancia con una clase SVM con aprendizaje activo, que concierne solamente a los ejemplos relevantes. No obstante, le integran un espacio adaptativo que toma en cuenta los ejemplos relevantes y no relevantes, es decir, transforma el espacio de rasgo original en otro, que se corresponde mejor a las necesidades y especificidades de usuario. Los resultados experimentales en una base de datos de géneros musicales muestran la efectividad del enfoque propuesto. Más adelante, en (Chen, 2009) se propone una estrategia de selección que obliga que los ejemplos seleccionados no sean solamente los cercanos a la frontera de decisión de SVM, sino que también sean diversos sobre el espacio de rasgo asociado a la función *kernel* de SVM, maximizando la distancia entre ellos. En los resultados experimentales demuestran que el desempeño del algoritmo es superior al SVM_{Activo} con una selección aleatoria, con una estrategia de muestra de incertidumbre y con la diversidad de ángulo propuesta en (Brinker, 2003).

De manera general, en la realimentación por relevancia activa, se destaca la aplicación del protocolo **basado en fondo** y la estrategia de consulta **muestra de incertidumbre** o extensiones de ella. Dicho tipo de recuperación, ha sido enfocado en mayor grado para imágenes que para otros tipos de objetos como documentos y música (Figura 10). En la recuperación de documentos, el aprendizaje activo se refleja en pocos trabajos. Sin embargo, en la recuperación de imágenes un conjunto diverso de autores experimentan en el aprendizaje activo, entre los que se destaca Cord de manera constante en el intervalo de fecha del 2004 al 2010.

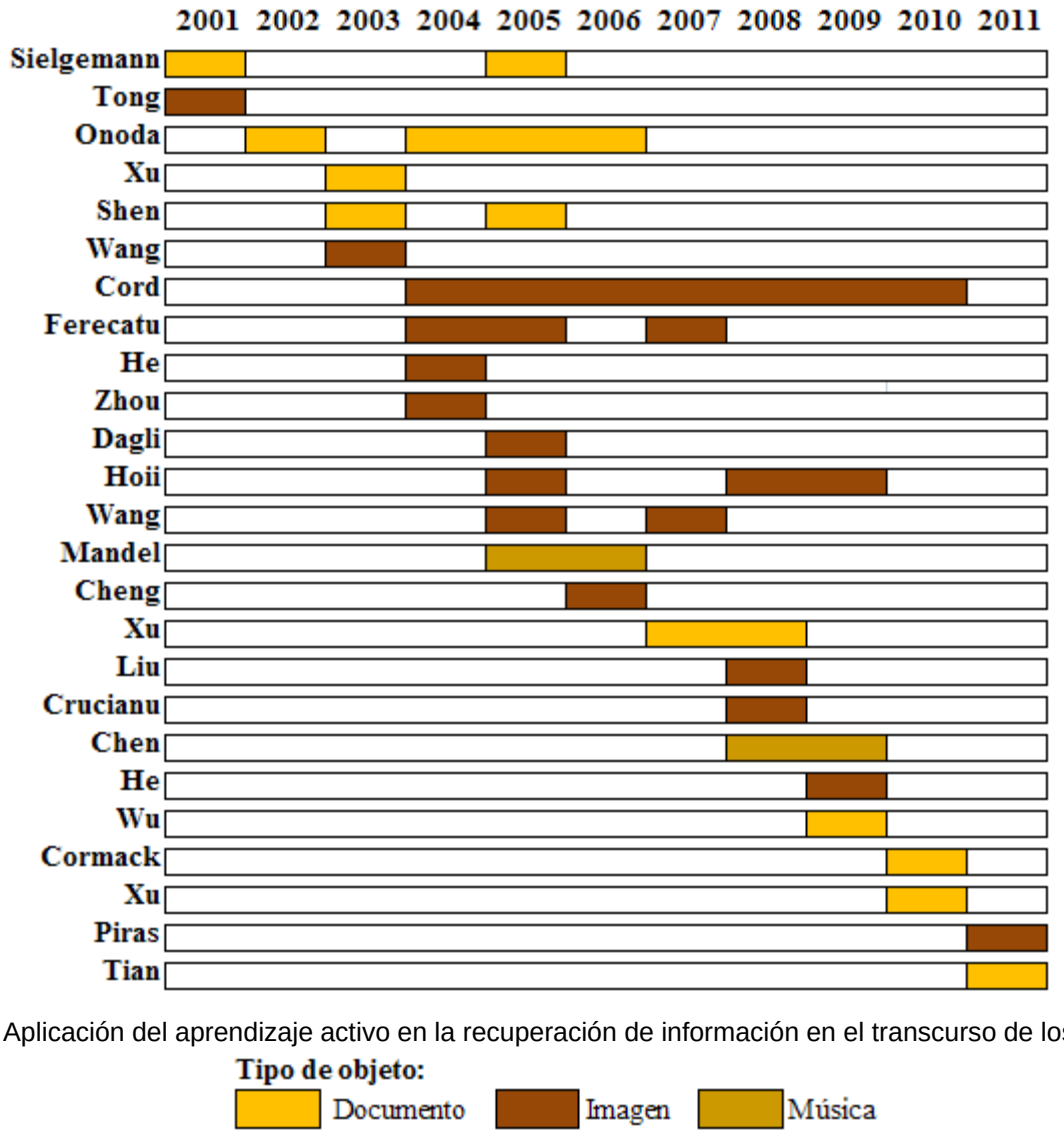


Figura 10: Aplicación del aprendizaje activo en la recuperación de información en el transcurso de los años.

1.5 Conclusiones parciales

En este capítulo se analizaron los conceptos básicos e importantes asociados a la recuperación de información, aprendizaje automático y activo, con la óptica de elevar la calidad de la búsqueda de información oportuna en un entorno periodístico. La sustitución exitosa de los algoritmos tradicionales de la realimentación por relevancia por algoritmos del aprendizaje supervisado en los SRI permitió la incorporación del aprendizaje activo como una técnica primordial para aprender el concepto de relevancia del usuario y disminuir el número de ciclos de consultas. La atención del aprendizaje activo en la recuperación de información se expresa en pocos trabajos, destacándose en la mayoría la aplicación del protocolo **basado en fondo** y la estrategia de consulta **muestra de incertidumbre** o extensiones de ella.

CAPÍTULO 2: DISEÑO DE UN MODELO DE REALIMENTACIÓN POR RELEVANCIA BASADO EN APRENDIZAJE ACTIVO

En el presente capítulo se describen los resultados obtenidos en el estudio experimental para la evaluación de diversos Sistemas de Recuperación de Información sobre la colección noticiosa TIME con medidas de **precisión** y **exhaustividad**. Además, se presenta el diseño del modelo propuesto como solución al problema científico plasmado en la investigación. Se explica cómo insertarlo en cualquier SRI clásico para un contexto periodístico, como lo es para el sitio Web del periódico **¡*ahora*!**, de gran utilidad durante el proceso editorial. Por último, se exponen las principales conclusiones obtenidas a partir de los resultados experimentales arrojados y lo tratado en el capítulo.

2.1 Descripción de la experimentación

De acuerdo a (Hull, 1993), los tres ingredientes de un experimento de recuperación de información controlado son:

1. Una colección de prueba de recuperación de información.
2. Una o más medidas de evaluación, que asignan valores a la efectividad de la búsqueda.
3. Una metodología estadística, que determina si las diferencias observadas en el desempeño entre los métodos investigados son estadísticamente significativas.

A continuación se describe el empleo de estos ingredientes para evaluar la calidad de los documentos recuperados en un SRI clásico sin realimentación por relevancia (método base), con respecto a otros que incorporan la realimentación por relevancia, pero con el empleo de diferentes técnicas.

Para llevar a cabo el estudio experimental se implementaron seis Sistemas de Recuperación de Información, empleando la librería Lucene y como entorno de desarrollo el NetBeans. Durante el proceso de indexación de cada uno de estos sistemas se aplicaron algunas de las técnicas de preprocesamiento inmersas en Lucene. Para el análisis léxico se empleó un método estándar, que tiene a su vez filtros para llevar todas las palabras a minúscula, eliminar caracteres especiales (signos de puntuación y guiones) y palabras vacías (usando un arreglo de preposiciones, artículos, adverbios, pronombres y conjunciones en el idioma inglés). Los números no se eliminan, al considerarse una información útil a almacenar en el contexto noticioso. El analizador utilizado permitió además mantener las siglas y correos

electrónicos como términos que poseen un significado colectivo, al no ser fraccionados por los signos y caracteres especiales inmersos.

Con el fin de reducir la cantidad de términos en la colección textual se aplicó el filtro de segmentación. En Lucene es implementado a través del algoritmo de Segmentación de Porter (Porter, 1980), que transforma las palabras a su forma léxica o lexema.

Los sistemas implementados se sustentan en el modelo del espacio vectorial, donde cada documento o consulta se encuentra expresada en forma de vector. La medida de similitud a utilizar para la búsqueda de documentos que se ajustan a la consulta es la del coseno del ángulo que forman los vectores de documentos y la consulta actual en el sistema. Los pesos de los términos en los documentos fueron determinados por la medida $tf * idf$ (explicada en el Epígrafe 1.2.2). Luego de culminado el proceso de indexación, se obtiene el **índice**, que es la estructura de datos que contiene los documentos previamente preprocesados.

El primer sistema implementado es un SRI clásico sin realimentación por relevancia; el segundo incorpora la realimentación por relevancia clásica implementada por (Rocchio, 1971); el tercero contiene el algoritmo de clasificación probabilístico Naive Bayes, que considera como ejemplos de entrenamiento los juicios de relevancia emitidos por el usuario (relevante o no relevante). La idea de este último es entrenar de manera iterativa e interactiva por parte del usuario un clasificador binario, hasta que el usuario considere que el sistema comprendió su concepto de necesidad de información. El cuarto incorpora al SRI una realimentación por relevancia con el algoritmo de clasificación previamente comentado, pero con un aprendizaje activo, aplicando la estrategia de consulta **muestra de incertidumbre** en un escenario **basado en fondo**. El otro sistema es igual al anterior, pero emplea la estrategia **consulta por comité** en un escenario de **muestreo selectivo basado en flujo**. Por último, el sexto SRI contempla la idea anterior, pero en un escenario **basado en fondo**. En las pruebas realizadas a los SRI que incluyen la realimentación por relevancia se ejecutaron dos iteraciones, para mitigar las posibles variaciones de los resultados.

2.1.1 Colección de prueba TIME

Para evaluar la calidad de los documentos recuperados en el SRI clásico, así como de los otros que incorporan diversas técnicas de realimentación por relevancia, se seleccionó la

colección de prueba TIME[30] de SMART, que contiene artículos de noticias mundiales de la revista americana TIME a partir del año 1963.

La colección contiene 425 documentos con un promedio de 27.3 oraciones/documento, y 83 consultas con un promedio de 3.9 documentos relevantes/consulta. Los vectores de consulta contienen pocos términos con respecto a los vectores de documentos. Incluye además los documentos relevantes a cada consulta, denominados **juicios de relevancia**.

Por tener en cuenta el objeto de estudio de la presente investigación, se decidió elegir la colección de prueba TIME. A pesar de esta última encontrarse en idioma inglés, abarca un contexto noticioso. Es válido aclarar que no se utilizó una colección en el idioma español porque no se encontró ninguna que cumpliera específicamente esta característica.

2.1.2 Medidas de evaluación utilizadas

Con el fin de medir la efectividad de cada uno de los sistemas antes mencionados, se utilizaron las medidas de evaluación **precisión** y **exhaustividad**. Para la primera y segunda iteración se evaluó la precisión (P, *precision*, por su significado en inglés) y la exhaustividad (R, *recall*, por su significado en inglés) del:

1. SRI clásico (SRI_C)
2. SRI con realimentación por relevancia Rocchio (SRI_ROCCHIO)
3. SRI con realimentación por relevancia, que utiliza el clasificador probabilístico Naive Bayes (SRI_NAIVE)
4. SRI con realimentación por relevancia, que emplea el clasificador probabilístico Naive Bayes con una estrategia de aprendizaje activo **muestra de incertidumbre** en un escenario **basado en fondo** (SRI_UNC)
5. SRI con realimentación por relevancia, que usa el clasificador probabilístico Naive Bayes con una estrategia de aprendizaje activo **consulta por comité** en un escenario de **muestreo selectivo basado en flujo** (SRI_QBC_F)
6. SRI con realimentación por relevancia, que aplica el clasificador probabilístico Naive Bayes con una estrategia de aprendizaje activo **consulta por comité** en un escenario **basado en fondo** (SRI_QBC _P)

Por cada una de las iteraciones se aplicó una realimentación seudo y explícita. En la seudo, se contemplaron los 10 primeros documentos como relevantes para el usuario; en la

[30] ftp://ftp.cs.cornell.edu/pub/smart/time/

explícita, la selección de los documentos relevantes fue realizada por el usuario para ayudar en la comprensión del sistema de la necesidad de información en cuestión. A continuación se visualizan en las tablas (2-5) los promedios de la precisión y exhaustividad por cada iteración y tipo de realimentación por relevancia de los sistemas evaluados.

	SRI_C	SRI_ROCCHIO	SRI_NAIVE	SRI_UNC	SRI_QBC_F	SRI_QBC_ P
P	0,122525557	0,09836955	0,053534395	0,027545231	0,032114762	0,032023646
R	0,903571191	0,72464127	0,86501621	0,8648577	0,84156526	0,84063848

Tabla 2: Promedio de la precisión (P) y exhaustividad (R) en la primera iteración para una realimentación seudo.

	SRI_C	SRI_ROCCHIO	SRI_NAIVE	SRI_UNC	SRI_QBC_F	SRI_QBC_ P
P	0,122525557	0,05650762	0,30113788	0,2378997	0,030884676	0,030896674
R	0,903571191	0,85949834	0,71545104	0,79802737	0,83996913	0,84156526

Tabla 3: Promedio de la precisión (P) y exhaustividad (R) en la segunda iteración para una realimentación seudo.

	SRI_C	SRI_ROCCHIO	SRI_NAIVE	SRI_UNC	SRI_QBC_F	SRI_QBC_ P
P	0,122525557	0,09836955	0,858355699	0,753006149	0,763257843	0,761778033
R	0,903571191	0,72464127	0,945239783	0,88134443	0,93432554	0,93219939

Tabla 4: Promedio de la precisión (P) y exhaustividad (R) en la primera iteración para una realimentación explícita.

	SRI_C	SRI_ROCCHIO	SRI_NAIVE	SRI_UNC	SRI_QBC_F	SRI_QBC_ P
P	0,122525557	0,05650762	0,975903614	0,908975899	0,762706156	0,763895018
R	0,903571191	0,85949834	0,90357119	0,841572366	0,93030947	0,92428538

Tabla 5: Promedio de la precisión (P) y exhaustividad (R) en la segunda iteración para una realimentación explícita.

En la Tabla 2 y 3 se puede apreciar cómo la **realimentación por relevancia seudo** no favorece la **precisión** y la **exhaustividad** de los documentos recuperados en la primera iteración de búsqueda, respecto al SRI clásico sin realimentación por relevancia (SRI_C). Sin embargo, en la segunda iteración (Tabla 3) el SRI que emplea en la realimentación por relevancia el clasificador Naive Bayes (SRI_NAIVE) se desempeña con una precisión superior al SRI clásico (SRI_C). De manera análoga pero con resultados inferiores a esta última, la realimentación por relevancia que emplea el clasificador Naive Bayes con una estrategia de consulta **muestra de incertidumbre** en un escenario **basado en fondo** es superior al SRI clásico (SRI_C).

De manera general, el SRI con realimentación Rocchio tiende a ser inferior en cuanto a la precisión y exhaustividad, a raíz de la cantidad de términos iniciales que contiene la consulta, con un promedio de 14 términos. En la primera iteración se expande la consulta con un término y en la segunda, con dos.

En cuanto a los resultados reflejados en las tablas 4 y 5, el SRI que emplea en la realimentación por relevancia el algoritmo de clasificación Naive Bayes (SRI_NAIVE) y los que utilizan el clasificador técnicas de aprendizaje activo (SRI_UNC, SRI_QBC_F y SRI_QBC_P) se comportan notablemente superiores respecto a la precisión de la recuperación del SRI clásico (SRI_C) para su primera y segunda iteración. Además, en términos de exhaustividad los dos sistemas de recuperación (SRI_QBC_F y SRI_QBC_P) que emplean la realimentación con el clasificador Naive Bayes, pero con una estrategia **consulta por comité** en un escenario **basado en fondo** y en uno **basado en flujo**, se comportan superior que el SRI clásico (SRI_C).

Desde otra perspectiva, si se analizan los resultados de las tablas 2 y 4, se puede evidenciar el incremento de la calidad (precisión y exhaustividad) del SRI, al emplear en la realimentación un clasificador, y en este último técnicas del aprendizaje activo, como **muestra de incertidumbre** y **consulta por comité**. La mejora paulatina de estos tres últimos casos la marca el simple hecho de utilizar una **realimentación por relevancia explícita** en vez de **seudo**. De manera análoga a esto ocurre con los resultados mostrados en las tablas 3 y 5 para una segunda iteración.

2.1.3 Validación estadística

En los experimentos realizados para la comparación de los diversos sistemas de RI se utilizaron pruebas no paramétricas, tanto para k muestras relacionadas (prueba de Friedman (Friedman, 1937, Friedman, 1940)), como para dos muestras relacionadas (Prueba de Wilcoxon (Wilcoxon, 1945)). Cuando se realizaron las pruebas de Friedman y de Wilcoxon se aplicó el método de Monte Carlo con intervalos de confianza del 95 y 99%, respectivamente, y un número de muestras igual a 10 000.

Se comprobó la hipótesis nula de que entre los algoritmos (una pareja de ellos o entre todos, según prueba) no hay diferencias significativas; en caso contrario se rechaza la misma. Se consideró altamente significativo un resultado menor que 0,01; significativo cuando es menor que 0,05 y mayor que 0.01; medianamente significativa, un resultado menor que 0,1 y mayor que 0,05; no significativo, un resultado mayor que 0,1.

Para el procesamiento de los resultados de los experimentos se utilizó la herramienta SPSS (*Statistical Package for Social Sciences,* por sus términos en inglés), versión 15.0 para Windows. Actualmente se considera uno de los paquetes de procesamiento estadístico más rigurosos y completos.

2.1.3.1 Resultados experimentales obtenidos

De manera general, los experimentos efectuados consisten en comparar los SRI: sin realimentación por relevancia (SRI_C), con los que aplican realimentación por relevancia, empleando el método Rocchio (SRI_ROCCHIO), el método de clasificación Naive Bayes (SRI_NAIVE), el método de clasificación Naive Bayes con una estrategia de aprendizaje activo **muestra de incertidumbre** en un escenario **basado en fondo** (SRI_UNC), el método de clasificación Naive Bayes con una estrategia de aprendizaje activo **consulta por comité** en un escenario de **muestreo selectivo basado en flujo** (SRI_QBC_F) y el método de clasificación Naive Bayes con una estrategia de aprendizaje activo **consulta por comité** en un escenario **basado en fondo** (SRI_QBC_P).

Los valores de los parámetros utilizados para el método Rocchio en la realimentación por relevancia del SRI, fueron de $\alpha = 1$, $\beta = 0{,}75$ y $\gamma = 0{,}25$. Además, se utilizó la colección residual en cada iteración ejecutada de realimentación.

Experimento 1: Comparar los SRI sin realimentación por relevancia con los que aplican la realimentación por relevancia seudo, respecto a los valores de precisión obtenidos en el resultado de búsqueda por cada consulta en la primera iteración.

Objetivos: Determinar si los SRI con realimentación por relevancia que aplican algoritmos de clasificación con un aprendizaje activo son significativamente superior al de un SRI clásico, un SRI con realimentación Rocchio y uno que aplica realimentación, pero empleando el clasificador probabilístico Naive Bayes.

Se aplicó la prueba de Friedman entre todos los sistemas respecto a los valores de precisión por cada consulta para la primera iteración y arrojó que existen diferencias significativas entre ellos.

Se efectuó la prueba de Wilcoxon dos a dos, respecto a los valores de precisión por cada consulta para la primera iteración y se corrobora la existencia de diferencias significativas entre:

- SRI_C y SRI_QBC_P

- SRI_C y SRI_QBC_F
- SRI_QBC_P y SRI_UNC
- SRI_ QBC_P y SRI_NAIVE

También, diferencias altamente significativas entre:

- SRI_ROCCHIO y SRI_QBC_P
- SRI_ROCCHIO y SRI_QBC_F
- SRI_ROCCHIO y SRI_UNC
- SRI_ROCCHIO y SRI_NAIVE
- SRI_C y SRI_UNC
- SRI_C y SRI_NAIVE
- SRI_ QBC_F y SRI_UNC
- SRI_ QBC_F y SRI_NAIVE

A partir de los resultados arrojados se considera que el SRI_ROCCHIO supera a los restantes sistemas en cuanto a la medida de precisión en la primera iteración con una realimentación seudo. En el Anexo 7 se muestran las tablas de las pruebas estadísticas relacionadas con el experimento 1.

Experimento 2: Comparar los SRI sin realimentación por relevancia con los que aplican la realimentación por relevancia seudo, respecto a los valores de exhaustividad obtenidos en el resultado de búsqueda por cada consulta en la primera iteración.

Objetivos: Determinar si los SRI con realimentación por relevancia que aplican algoritmos de clasificación con un aprendizaje activo son significativamente superior al de un SRI clásico, un SRI con realimentación Rocchio y uno que aplica realimentación pero empleando el clasificador probabilístico Naive Bayes.

Se aplicó la prueba de Friedman entre todos los sistemas, respecto a los valores de exhaustividad por cada consulta para la primera iteración y arrojó que existen diferencias significativas entre ellos.

Se empleó además la prueba de Wilcoxon dos a dos, respecto a los valores de exhaustividad por cada consulta para la primera iteración y corrobora la existencia de diferencias significativas entre: SRI_ROCCHIO y SRI_QBC_F, y SRI_ROCCHIO y SRI_QBC_P; medianamente significativas entre: SRI_ROCCHIO y SRI_C, SRI_ ROCCHIO y SRI_UNC, y

SRI_ROCCHIO y SRI_NAIVE. Mientras que entre SRI_ROCCHIO y SRI_QBC_F, y SRI_ROCCHIO y SRI_QBC_P, existen diferencias altamente significativas.

A partir de los resultados arrojados se considera que el SRI_C supera a los restantes sistemas en cuanto a la exhaustividad para la primera iteración con una realimentación seudo. En el Anexo 7 se muestran las tablas de las pruebas estadísticas relacionadas con el experimento 2.

Experimento 3: Comparar los SRI sin realimentación por relevancia con los que aplican la realimentación por relevancia seudo, respecto a los valores de precisión obtenidos en el resultado de búsqueda por cada consulta en la segunda iteración.

Objetivos: Determinar si los SRI con realimentación por relevancia que aplican algoritmos de clasificación con un aprendizaje activo son significativamente superior al de un SRI clásico, un SRI con realimentación Rocchio y uno que aplica realimentación pero empleando el clasificador probabilístico Naive Bayes.

Se aplicó la prueba de Friedman entre todos los sistemas, respecto a los valores de precisión por cada consulta para la segunda iteración y arrojó que existen diferencias significativas entre ellos.

Se realizó también la prueba de Wilcoxon dos a dos, respecto a los valores de precisión por cada consulta para la segunda iteración. Se corrobora la existencia de diferencias significativas entre: SRI_ROCCHIO y SRI_C, y SRI_QBC_F y SRI_C. Se observaron diferencias altamente significativas entre:

- SRI_QBC_P y SRI_NAIVE
- SRI_QBC_P y SRI_UNC
- SRI_QBC_P y SRI_C
- SRI_QBC_P y SRI_ROCCHIO
- SRI_QBC_P y SRI_QBC_F
- SRI_NAIVE y SRI_UNC
- SRI_NAIVE y SRI_C
- SRI_NAIVE y SRI_ ROCCHIO
- SRI_NAIVE y SRI_QBC_F
- SRI_UNC y SRI_C
- SRI_UNC y SRI_ROCCHIO

- SRI_UNC y SRI_QBC_F
- SRI_ROCCHIO y SRI_QBC_F

A raíz de los resultados arrojados se considera que el SRI_QBC_P supera a los restantes sistemas en cuanto a la precisión para la segunda iteración con una realimentación seudo. En el Anexo 7 se muestran las tablas de las pruebas estadísticas relacionadas con el experimento 3.

Experimento 4: Comparar los SRI sin realimentación por relevancia con los que aplican la realimentación por relevancia seudo, respecto a los valores de exhaustividad obtenidos en el resultado de búsqueda por cada consulta en la segunda iteración.

Objetivos: Determinar si los SRI con realimentación por relevancia que aplican algoritmos de clasificación con un aprendizaje activo son significativamente superior al de un SRI clásico, un SRI con realimentación Rocchio y uno que aplica realimentación pero empleando el clasificador probabilístico Naive Bayes.

Se aplicó la prueba de Friedman entre todos los sistemas, respecto a los valores de exhaustividad por cada consulta para la segunda iteración y arrojó que existen diferencias significativas entre ellos.

Se efectuó, además, la prueba de Wilcoxon dos a dos, respecto a los valores de exhaustividad por cada consulta para la segunda iteración. Se corrobora la existencia de diferencias significativas entre: SRI_QBC_P y SRI_UNC, y SRI_QBC_F y SRI_UNC; medianamente significativo entre SRI_ROCCHIO y SRI_UNC. Entre SRI_C y SRI_UNC, SRI_C y SRI_NAIVE, SRI_ROCCHIO y SRI_NAIVE, SRI_QBC_P y SRI_NAIVE, SRI_QBC_F y SRI_NAIVE, y SRI_UNC y SRI_NAIVE existen diferencias altamente significativas.

A partir de los resultados obtenidos se considera que el SRI_C supera a los restantes sistemas en cuanto a la exhaustividad para la segunda iteración con una realimentación seudo. En el Anexo 7 se muestran las tablas de las pruebas estadísticas relacionadas con el experimento 4.

Experimento 5: Comparar los SRI sin realimentación por relevancia con los que aplican la realimentación por relevancia explícita, respecto a los valores de precisión obtenidos en el resultado de búsqueda por cada consulta en la primera iteración.

Objetivos: Determinar si los SRI con realimentación por relevancia que aplican algoritmos de clasificación con un aprendizaje activo son significativamente superior al de un SRI clásico, un SRI con realimentación Rocchio y uno que aplica realimentación, pero empleando el clasificador probabilístico Naive Bayes.

Se aplicó la prueba de Friedman entre todos los sistemas, respecto a los valores de precisión por cada consulta para la primera iteración y arrojó que existen diferencias significativas entre ellos.

Se utilizó la prueba de Wilcoxon dos a dos, respecto a los valores de precisión por cada consulta para la primera iteración y se corrobora la existencia de diferencias altamente significativas entre:

- SRI_NAIVE y SRI_QBC_P
- SRI_NAIVE y SRI_QBC_F
- SRI_NAIVE y SRI_UNC
- SRI_NAIVE y SRI_ROCCHIO
- SRI_NAIVE y SRI_C
- SRI_QBC_P y SRI_ROCCHIO
- SRI_QBC_P y SRI_C
- SRI_QBC_F y SRI_ROCCHIO
- SRI_QBC_F y SRI_C
- SRI_UNC y SRI_ROCCHIO
- SRI_UNC y SRI_C

A raíz de los resultados alcanzados, se concluye que SRI_NAIVE supera a los restantes sistemas en cuanto a la medida de precisión en la primera iteración del sistema con una realimentación explícita. En el Anexo 7 se muestran las tablas de las pruebas estadísticas relacionadas con el experimento 5.

Experimento 6: Comparar los SRI sin realimentación por relevancia con los que aplican la realimentación por relevancia explícita, respecto a los valores de exhaustividad obtenidos en el resultado de búsqueda por cada consulta en la primera iteración.

Objetivos: Determinar si los SRI con realimentación por relevancia que aplican algoritmos de clasificación con un aprendizaje activo son significativamente superior al de un SRI clásico,

un SRI con realimentación Rocchio y uno que aplica realimentación pero empleando el clasificador probabilístico Naive Bayes.

Se empleó la prueba de Friedman entre todos los sistemas, respecto a los valores de exhaustividad por cada consulta para la primera iteración y arrojó que existen diferencias significativas entre ellos.

Se realizó la prueba de Wilcoxon dos a dos, respecto a los valores de exhaustividad por cada consulta para la primera iteración y se corrobora la existencia de diferencias significativas entre: SRI_UNC y SRI_NAIVE, SRI_QBC_P y SRI_C, y SRI_C y SRI_QBC_F. Existen diferencias altamente significativas entre:

- SRI_UNC y SRI_C
- SRI_UNC y SRI_ QBC_F
- SRI_UNC y SRI_ROCCHIO
- SRI_QBC_P y SRI_QBC_F
- SRI_QBC_P y SRI_ROCCHIO
- SRI_NAIVE y SRI_C
- SRI_NAIVE y SRI_QBC_F
- SRI_NAIVE y SRI_ROCCHIO
- SRI_C y SRI_ROCCHIO

Al analizar los resultados, SRI_UNC supera a los restantes sistemas en cuanto a la exhaustividad para la primera iteración con una realimentación explícita. En el Anexo 7 se muestran las tablas de las pruebas estadísticas relacionadas con el experimento 6.

Experimento 7: Comparar los SRI sin realimentación por relevancia con los que aplican la realimentación por relevancia explícita, respecto a los valores de precisión obtenidos en el resultado de búsqueda por cada consulta en la segunda iteración.

Objetivos: Determinar si los SRI con realimentación por relevancia que aplican algoritmos de clasificación con un aprendizaje activo son significativamente superior al de un SRI clásico, un SRI con realimentación Rocchio y uno que aplica realimentación pero empleando el clasificador probabilístico Naive Bayes.

Se aplicó la prueba de Friedman entre todos los sistemas, respecto a los valores de precisión por cada consulta para la segunda iteración y arrojó que existen diferencias significativas entre ellos.

Se efectuó también la prueba de Wilcoxon dos a dos, respecto a los valores de precisión por cada consulta para la segunda iteración. Se corrobora la existencia de diferencias significativas entre: SRI_C y SRI_ROCCHIO, mientras que existen diferencias altamente significativas entre:

- SRI_NAIVE y SRI_UNC
- SRI_NAIVE y SRI_QBC_P
- SRI_NAIVE y SRI_QBC_F
- SRI_NAIVE y SRI_C
- SRI_NAIVE y SRI_ROCCHIO
- SRI_UNC y SRI_QBC_P
- SRI_UNC y SRI_QBC_F
- SRI_UNC y SRI_C
- SRI_UNC y SRI_ROCCHIO
- SRI_QBC_P y SRI_C
- SRI_QBC_P y SRI_ROCCHIO
- SRI_QBC_F y SRI_C
- SRI_QBC_F y SRI_ROCCHIO

A partir de los resultados arrojados, SRI_NAIVE supera a los restantes sistemas en cuanto a la precisión para la segunda iteración con una realimentación explícita. En el Anexo 7 se muestran las tablas de las pruebas estadísticas relacionadas con el experimento 7.

Experimento 8: Comparar los SRI sin realimentación por relevancia con los que aplican la realimentación por relevancia explícita, respecto a los valores de exhaustividad obtenidos en el resultado de búsqueda por cada consulta en la segunda iteración.

Objetivos: Determinar si los SRI con realimentación por relevancia que aplican algoritmos de clasificación con un aprendizaje activo son significativamente superior al de un SRI clásico, un SRI con realimentación Rocchio y uno que aplica realimentación pero empleando el clasificador probabilístico Naive Bayes.

Se efectuó la prueba de Friedman entre todos los sistemas, respecto a los valores de exhaustividad por cada consulta para la segunda iteración y arrojó que existen diferencias significativas entre ellos.

Se llevó a cabo también la prueba de Wilcoxon dos a dos, respecto a los valores de exhaustividad por cada consulta para la segunda iteración. Se corrobora la existencia de diferencias significativas entre: SRI_QBC_F y SRI_C, y SRI_QBC_F y SRI_NAIVE, y SRI_QBC_F y SRI_ROCCHIO; medianamente significativas entre SRI_QBC_P y SRI_ROCCHIO. Entre: SRI_QBC_F y SRI_UNC, SRI_QBC_P y SRI_UNC, SRI_C y SRI_UNC, y SRI_NAIVE y SRI_UNC existen diferencias altamente significativas.

Al observar los resultados, SRI_QBC_F supera a los restantes sistemas en cuanto a la exhaustividad para la segunda iteración con una realimentación explícita. En el Anexo 7 se muestran las tablas de las pruebas estadísticas relacionadas con el experimento 8.

Al término de las experimentaciones descritas anteriormente, se puede evidenciar el diseño adecuado de un SRI que emplee técnicas de aprendizaje activo para incrementar la calidad de la búsqueda de información en un entorno periodístico. Tal es el caso, de un SRI que requiera alta precisión e involucre la realimentación por relevancia seudo con un clasificador Naive Bayes y la estrategia **consulta por comité** en un escenario **basado en fondo**, garantizando al menos dos iteraciones (experimento 3). También se puede diseñar el SRI con realimentación por relevancia explícita que contemple un clasificador Naive Bayes y una estrategia de consulta **muestra de incertidumbre** en un escenario **basado en fondo** o una estrategia de **consulta por comité** en un escenario **basado en flujo** (experimentos 6 y 8). Ambos son convenientes aplicarlos siempre y cuando el objetivo del sistema sea aumentar la exhaustividad más que la precisión, siendo suficiente una iteración para la realimentación que contemple la estrategia de consulta **muestra de incertidumbre** y dos iteraciones para la de **consulta por comité**.

Por otra parte, los resultados experimentales permiten afirmar que no solo la estrategia de consulta **muestra de incertidumbre** en un escenario **basado en fondo** no es la única técnica de aprendizaje activo que favorece la calidad de respuesta en un sistema de esta índole, sino que la estrategia **consulta por comité** también la beneficia, e incluso en determinados contextos supera su desempeño. El introducir estas técnicas de aprendizaje activo al clasificador inmerso en la realimentación por relevancia de un SRI permite que éste se entrene con pocos ejemplos, pero con un alto grado de calidad para su posterior clasificación, y de estar manera evitar el agotamiento por parte de los usuarios, al tener que iterar e interactuar con el sistema muchas veces para entrenar de manera adecuada el clasificador.

2.2 Incorporación de modelos de realimentación por relevancia basados en aprendizaje activo en un Sistema de Recuperación de Información

El diseño del SRI que se propone en la investigación se sustenta en el modelo del espacio vectorial, donde cada documento o consulta será expresada en forma de vector. La medida de similitud a utilizar para la búsqueda de documentos que se ajustan a la consulta es la del coseno del ángulo que forman los vectores de documentos y la consulta actual en el sistema.

A continuación se desglosan las técnicas y el orden que se proponen aplicar para el desarrollo del SRI en un contexto periodístico, basado en el estudio experimental anteriormente comentado. Durante el proceso de indexación se deben contemplar las siguientes técnicas:

1. Análisis léxico
2. Eliminación de las palabras vacías
3. Segmentación
4. Ponderación de los términos
5. Creación del índice

En el SRI por defecto debe realizarse una búsqueda tradicional como en otros sistemas de recuperación clásico, haciendo uso de la medida de similitud coseno. Una vez que devuelva los resultados de la primera consulta, se puede incrementar la precisión de la búsqueda de información al incluir en el sistema una **realimentación por relevancia seudo** (para los 10 primeros documentos), que incorpora el algoritmo de clasificación probabilístico Naive Bayes con la estrategia de **consulta por comité** en un escenario **basado en fondo**. A continuación se muestra un esbozo del algoritmo de **realimentación por relevancia seudo** propuesto a insertar en el sistema, para ser ejecutado de manera automática luego de la primera iteración donde se le muestran al usuario los resultados de la búsqueda:

Entrada: Documentos recuperados por el sistema en la primera iteración de la búsqueda.

1. Llama a un $\boldsymbol{Experto\ automático}$ para obtener de manera seudo los 10 primeros documentos etiquetados como $\boldsymbol{relevante}$ y los otros diez restantes como $\boldsymbol{no\ relevante}$.
2. Se entrenan los 7 clasificadores $\boldsymbol{Naive\ Bayes}$ que componen el comité con los documentos etiquetados en 1.

3. Se escogen aleatoriamente dos miembros del comité para ambos predecir las etiquetas de los documentos del índice que no han sido mostrados al usuario.
4. Todos los documentos que discrepen en la predicción de la etiqueta por los dos miembros, son analizados por el ***Experto automático*** para etiquetarlos correctamente y adicionarlos al conjunto de entrenamiento.
5. Se actualizan todos los miembros del comité con el nuevo conjunto de entrenamiento formado.
6. Si el usuario decide terminar el entrenamiento de los miembros del comité, cada uno de los miembros clasifican el índice del SRI en ***relevante*** o ***no relevante*** a partir de la mayoría de votos de éstos y se muestran los documentos resultantes relevantes al usuario. Si no, va al paso 3.

Algoritmo 1: Realimentación por relevancia seudo, que emplea el algoritmo de clasificación probabilístico Naive Bayes con la estrategia **consulta por comité** en un escenario **basado en fondo**.

En el caso de quererse incrementar la exhaustividad de la búsqueda de información en el SRI, se puede incluir en él una **realimentación por relevancia explícita** (para ser escogidos los documentos manualmente por el usuario), que utiliza el algoritmo de clasificación probabilístico Naive Bayes con la estrategia de consulta muestra de incertidumbre en un escenario basado en fondo. En lo que sigue se expone el algoritmo, que es iniciado por el usuario luego de la primera iteración, donde se le muestra los resultados de la búsqueda:

Entrada: Documentos recuperados por el sistema en la primera iteración de la búsqueda.

1. El usuario escoge los documentos que considera ***relevante***, así como los ***no relevantes*** a la consulta.
2. Se entrena el clasificador ***Naive Bayes*** con los documentos etiquetados en 1.
3. El clasificador predice las etiquetas de los documentos del índice que no han sido mostrados al usuario.
4. Se escogen los b documentos para los cuales el clasificador es menos confiable y se le muestran al usuario.

5. Se aplica el paso 1 y se adicionan los documentos etiquetados al conjunto de entrenamiento.
6. Se actualiza el clasificador ***Naive Bayes*** con el nuevo conjunto de entrenamiento formado.
7. Si el usuario decide terminar el entrenamiento del clasificador, se clasifica el índice del SRI en ***relevante*** o ***no relevante*** y se le muestran al usuario los documentos resultantes relevantes. Si no, va al paso 3.

Algoritmo 2: Realimentación por relevancia explícita, que utiliza el algoritmo de clasificación probabilístico Naive Bayes con la estrategia de consulta **muestra de incertidumbre** en un escenario **basado en fondo**.

También se puede emplear una **realimentación por relevancia explícita** (para ser escogidos los documentos manualmente por el usuario), utilizando el algoritmo de clasificación probabilístico Naive Bayes con la estrategia de consulta por comité en un escenario basado en flujo.

Entrada: Documentos recuperados por el sistema en la primera iteración de la búsqueda.

1. El usuario escoge los documentos que considera ***relevante***, así como los ***no relevantes*** a la consulta.
2. Se entrenan los 7 clasificadores ***Naive Bayes***, que componen el comité.
3. Se escoge aleatoriamente un documento no etiquetado del índice, que no ha sido mostrado al usuario.
4. Se escogen aleatoriamente dos miembros del comité para ambos predecir la etiqueta del documento escogido.
5. Si las dos predicciones son iguales, entonces rechaza el documento y escoge otro.
6. Si no, se le muestra al usuario el documento para que lo etiquete correctamente, y se adiciona el documento etiquetado al conjunto de entrenamiento del clasificador.
7. Se actualizan todos los miembros del comité con el nuevo conjunto de entrenamiento formado.

8. Si el usuario decide terminar el entrenamiento de los miembros del comité, cada uno de los miembros clasifican el índice del SRI en ***relevante*** o ***no relevante*** a partir de la mayoría de votos de éstos y le muestran los documentos relevantes resultantes al usuario. Si no, va al paso 3.

Algoritmo 3: Realimentación por relevancia explícita, que emplea el algoritmo de clasificación probabilístico Naive Bayes con la estrategia **consulta por comité** en un escenario **basado en flujo**.

Desde la perspectiva de la investigación (aumentar la calidad de la búsqueda de información), los resultados obtenidos y la experimentación realizada, se sugiere implantar en el sitio Web del periódico **¡*ahora*!** el modelo reflejado en el algoritmo 1. Este algoritmo se destaca por los valores de **precisión** obtenidos respecto a los restantes, elemento que resulta imprescindible en momentos del proceso editorial donde se requiera buscar información oportuna en breve tiempo. El modelo es un SRI que incluye la **realimentación por relevancia seudo**, basada en el algoritmo de clasificación probabilístico Naive Bayes con la estrategia **consulta por comité** en un escenario **basado en fondo**. No obstante, se deja a consideración de los clientes finales la utilización de cualquiera de los tres algoritmos para el momento de la implantación. Los restantes algoritmos 2 y 3 se diferencian del primero en que son adecuados cuando se requiere obtener la mayor cantidad de documentos que se ajustan a la consulta y no los más oportuno o precisos.

Sustituir el proceso de búsqueda actual (inmerso en el sitio Web del **¡*ahora*!**) por el SRI que se propone con realimentación por relevancia basado en aprendizaje activo, permitirá sintetizar momentos de búsqueda de información oportuna durante el proceso editorial del periódico **¡*ahora*!** digital. Específicamente en la fase de asignación de tareas por la **Editora Web principal**, lo que le permitirá saber qué periodistas redactaron noticias análogas a las planificadas a publicar. Por otra parte, los periodistas podrán obtener en la búsqueda, trabajos publicados en la Web (ya sea por él o no) para redactar otros similares a los encontrados y poder realizar la tarea en menor tiempo y con las características deseadas por la **Editora Web Principal**. La aplicación del modelo beneficiará también la búsqueda de noticias, ya sea por la necesidad de reemplazar alguna de las que habían sido previstas a publicar de manera inmediata o por la de cubrir algún espacio noticioso libre.

2.3 Conclusiones parciales

En el presente capítulo se propone el diseño de un modelo de realimentación por relevancia basado en aprendizaje activo, que permite aumentar la calidad de la búsqueda de información oportuna en un entorno periodístico. En los resultados obtenidos de la experimentación realizada con la colección de prueba TIME se constata la eficacia en tres variantes de realimentación por relevancia activa. La primera permite un aumento de la precisión, al incorporarse la **realimentación por relevancia seudo**, que utiliza la técnica de clasificación probabilística Naive Bayes con la estrategia **consulta por comité** en un escenario **basado en fondo**. La segunda y la tercera garantizan incrementar la exhaustividad, si se asocia con la **realimentación por relevancia explícita**, que utiliza la técnica de clasificación probabilística Naive Bayes con la estrategia de consulta **muestra de incertidumbre** en un escenario **basado en fondo**, o se vincula con la estrategia **consulta por comité** en un escenario **basado en flujo**.

CONCLUSIONES

Como conclusiones generales de la presente investigación, se obtienen las siguientes ideas:

Para solucionar el problema científico planteado se asumieron los fundamentos teóricos de la disciplina Recuperación de Información como punto de partida.

La realimentación por relevancia basada en aprendizaje activo es actualmente muy investigada en el área de la Recuperación de Información, permite un aumento considerable de la precisión y exhaustividad en los Sistemas de Recuperación de Información y con ello la calidad general del proceso de búsqueda de información, como lo evidencian los resultados obtenidos.

El diseño del modelo propuesto que incorpora técnicas de aprendizaje activo incrementa la calidad de la búsqueda de información dentro de los sistemas de recuperación en el entorno periodístico, resultado respaldado con las evaluaciones experimentales realizadas.

El empleo de las pruebas no paramétricas de Friedman y Wilcoxon permitió realizar el análisis estadístico de la comparación entre los diversos Sistemas de Recuperación de Información e identificar aquellos que significativamente resultaran superiores, teniendo en cuenta la precisión y la exhaustividad de los resultados arrojados en la búsqueda para un dominio noticioso.

RECOMENDACIONES

Como resultado de la investigación realizada y los resultados obtenidos, se recomienda:

Evaluar la efectividad de los resultados de búsqueda en sistemas de recuperación de información con realimentación por relevancia empleando técnicas de aprendizaje activo para el caso de otras fuentes documentales y otros métodos de clasificación que se adecuen en el ámbito tratado.

Implantar el modelo propuesto en el sitio Web del periódico **¡*ahora*!** y en el resto de los medios digitales del país.

BIBLIOGRAFÍA

ABE, N. A. M. H. (1998) Query Learning Strategies using Boosting and Bagging. **In Proceedingsof the International Conference on Machine Learning (ICML)**, pp. 1–9.

ALLAN, J. (2003) HARD track overview in TREC2003. **Proceedings of TREC 2003**, pp.

ANDREWS, S., I. TSOCHANTARIDIS, AND T. HOFMANN (2003) Support vector machines for multiple-instance learning. **In Advances in Neural Information Processing Systems (NIPS)**, 15(pp. 561–568.

ANGLUIN, D. (1988) Queries and concept learning. **Machine Learning**, 2(pp. 319–342.

ARGAMON-ENGELSON, S. A. I. D. (1999) Committee-based sample selection for probabilistic classifiers. **Journal of Artificial Intelligence Research**, 11(pp. 335–360.

BACHRACH, R., S. FINE AND E. SHAMIR (2002) Query by committee, linear separation and random walks. **TCS**, 284(1)(pp.

BACHRACH, R. G., AMIR NAVOT AND NAFTALI TISHBY (2005) Query By Committee Made Real. pp.

BACHRACH, R. G., NAFTALI TISHBY AND AMIR NAVOT (2003) Kernel Query By Committee (KQBC). pp.

BAEZA-YATES, R. A. N., GONZALO (2005) Recuperación de la Información: Modelos, Estructuras de Datos, Algoritmos y Búsqueda en la Web. pp.

BAEZA-YATES, R. A. R.-N., B. (1999) Modern Information Retrieval. **ACM Press. Addison Wesley**, pp.

BECKER, M., BEN HACHEY, BEATRICE ALEX AND CLAIRE GROVER (2005a) Optimising selective sampling for bootstrapping named entity recognition. **Proceedings of the ICML 2005 Workshop on Learning with Multiple Views**, pp. 5-11.

BECKER, M. A. M. O. (2005b) A two-stage method for active learning of statistical grammars. **Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence**, pp. 991–996.

BREIMAN, L. (1996) Bagging predictors. **Machine Learning**, 24(2)(pp. 123–140.

BRINKER, K. (2003) Incorporating Diversity in Active Learning with Support Vector Machines. **Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)**, pp. 59-66.

CARLSON, C. (2003) Information overload, retrieval strategies and Internet user empowerment. **In Haddon, Leslie, Eds. Proceedings The Good, the Bad and the Irrelevant (COST 269)**, 1(pp. 169-173.

CARVALHO, V. R. A. W. W. C. (2004) Learning to extract signature and reply lines from email. **Proceedings of the Conference on Email and Anti-Spam (CEAS)**, pp.

CLEVERDON, C. W. (1972) On the inverse relationship of recall and precision. **Journal of Documentation**, 28(pp. 195-201.

COHN, D., LES ATLAS, RICHARD LADNER, M. A. EL-SHARKAWI, R. J. MARKS II, M. E. AGGOUNE AND D. C. PARK (1990) Training connectionist networks with queries and selective sampling. **In Advances in Neural Information Processing Systems (NIPS)**, pp. 566-573.

COHN, D. A. (1994) Neural network exploration using optimal experiment design. **In Advances in Neural Information Processing Systems (NIPS). Morgan Kaufmann**, 6(pp. 679-686.

COHN, D. A., ZOUBIN GHAHRAMANI AND MICHAEL I. JORDAN (1996) Active learning with statistical models. **Journal of Artificial Intelligence Research**, 4(pp. 129–145.

CONCEPCIÓN, G. M. R. A. F. R. E. (2005) **Rol del profesor y sus estudiantes en el proceso de enseñanza aprendizaje**, Holguín, Cuba, Ediciones Holguín.

COOPER, W. S. (1973) On selecting a Measure of Retrieval Effectiveness. **Journal of the American Society for Information Science**, 24(pp. 87-92.

CORD, M., DAVID GORISSE, AND FREDERIC PRECIOSO (2009) Optimization on active learning strategy for object category retrieval. **in IEEE ICIP**, pp.

CORD, M., DAVID GORISSE, AND FREDERIC PRECIOSO (2010) SCALABLE ACTIVE LEARNING STRATEGY FOR OBJECT CATEGORY RETRIEVAL. pp.

CORD, M., PHILIPPE HENRI GOSSELIN AND SYLVIE PHILIPP-FOLIGUET (2007) Stochastic exploration and active learning for image retrieval. **Image and Vision Computing**, 25(pp. 14–23.

CORD, M. A. P. H. G. (2004) RETIN AL: An active learning strategy for image category retrieval. **In IEEE International Conference on Image Processing**, 4(pp. 2219–2222.

CORD, M. A. P. H. G. (2005) Active Learning Techniques for User Interactive Systems: Application to Image Retrieval. **Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content**, pp.

CORD, M. A. P. H. G. (2006) Image retrieval using long-term semantic learning. pp.

CORD, M. A. P. H. G. (2008) Active learning methods for Interactive Image Retrieval. pp.

CORMACK, G. V. A. M. M. (2010) Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks. pp.

CROFT, W. B. (1987) Approaches to intelligent information retrieval. **Information Proccesing & Management**, 23, 4(pp. 249-254.

CRUCIANU, M., DANIEL ESTEVEZ, VINCENT ORIA AND JEAN-PHILIPPE TAREL (2008) Speeding Up Active Relevance Feedback with Approximate kNN Retrieval for Hyperplane Queries. pp.

CULOTTA, A., TRAUSTI KRISTJANSSON, ANDREW MCCALLUMA AND PAUL VIOLA (2006) Corrective feedback and persistent learning for information extraction. **Journal of Artificial Intelligence 170** 14(pp. 1101–1122.

CHAN, Y. S. A. H. T. N. (2007) Domain adaptation with active learning for word sense disambiguation. **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)**, pp. 49–56.

CHANG, E. Y., SIMON TONG, KINGSHY GOH AND CHENG-WEI CHANG (2005) Support Vector Machine Concept-Dependent Active Learning for Image Retrieval. **IEEE Transactions on Multimedia**, pp.

CHEN, G., TIAN-JIANG WANG, AND PERFECTO HERRERA (2009) Music Retrieval Based on a Multi-samples Selection Strategy for Support Vector Machine Active Learning. pp.

CHEN, G., TIANJIANG WANG AND PERFECTO HERRERA (2008) Relevance Feedback in an Adaptive Space with One-Class SVM for Content-Based Music Retrieval **IEEE Xplore**, pp.

CHEN, J., ANDREW SCHEIN, LYLE UNGAR AND MARTHA PALMER (2006) An empirical study of the behavior of active learning for word sense disambiguation. **Proceedings of the Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL 2006)**, pp. 120–127.

CHEN, Y., XIANG ZHOU, AND THOMAS S. HUANG (2001) One-Class SVM for learning in image retrieval. **In Proc. IEEE Int'l Conf. on Image Processing** pp.

DAGAN, I. A. S. P. E. (1995) Committee-based sampling for training probabilistic classifiers. **Proceedings of the Twelfth International Conference on Machine Learning**, pp. 150–157.

DAGLI, C. K., SHYAMSUNDAR RAJARAM AND THOMAS S. HUANG (2005) Combining Diversity-Based Active Learning with Discriminant Analysis in Image Retrieval. **Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)** pp. 173-178.

DIETTERICH, T., R. LATHROP, AND T. LOZANO-PEREZ (1997) Solving the multiple-instance problem with axis-parallel rectangles. **Artificial Intelligence**, 89(pp. 31–71.

DONMEZ, P. A. J. G. C. (2009) Active Sampling for Rank Learning via Optimizing the Area Under the ROC Curve. pp.

DRUCK, G., BURR SETTLES, AND ANDREW MCCALLUM (2009) Active learning by labeling features. **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 81–90.

DRUCKER, H., BEHZAD SHAHRARY AND DAVID C. GIBBON (2001) Relevance Feedback using Support Vector Machines **Proc. 18th International Conf. On Machine Learning**, pp. 122-129.

FERECATU, M., MICHEL CRUCIANU AND NOZHA BOUJEMAA (2004) Reducing the Redundancy in the Selection of Samples for SVM-based Relevance Feedback. pp.

FERECATU, M., MICHEL CRUCIANU AND NOZHA BOUJEMAA (2005) Active SVM-based Relevance Feedback with Hybrid Visual and Conceptual Image Representation. pp.

FERECATU, M. A. N. B. (2007) Interactive Remote Sensing Image Retrieval Using Active Relevance Feedback. **IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (DRAFT)**, pp.

FINN, A. A. N. K. (2003) Active learning selection strategies for information extraction. **Proceedings of the International Workshop on Adaptive Text Extraction and Mining (ATEM-03)**, pp. 18–25.

FREUND, Y. A. R. E. S. (1997a) A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences**, 1(pp. 119–139.

FREUND, Y. S., H.; SHAMIR, E.; AND TISHBY, N. (1997b) Selective sampling using the query by committee algorithm. **Machine Learning**, 28(pp. 133-168.

FRIEDMAN, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the American Statistical Association**, 32(pp. 675–701.

FRIEDMAN, M. (1940) A comparison of alternative tests of significance for the problem of m rankings. **Annals of Mathematical Statistics**, 11(pp. 86–92.

FUJII, A., TAKENOBU TOKUNAGA, KENTARO INUI AND HOZUMI TANAKA (1998) Selective sampling for example-based word sense disambiguation. **Computational Linguistics**, 4(pp. 573–597.

GAN, G., CHAOQUN MA, AND JIANHONG WU (2007) **Data Clustering: Theory, Algorithms and Applications**, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

GEMAN, E. B., AND R. DOURSAT (1992) Neural networks and the bias/variance dilemma. **Neural Computation**, 4(pp. 1-58.

GONZÁLEZ, J. J. C., J. R. MÉNDEZ REBOREDO AND F. FDEZ-RIVEROLA (2005) Modelos Anti-Spam de Inteligencia Artificial. **Conferencia IADIS Ibero-Americana WWW/Internet 2005**, pp. 548-551.

GUO, Y. A. R. G. (2007) Optimistic active learning using mutual information. **In Proceedings of International Joint Conference on Artificial Intelligence (IJ-CAI). AAAI Press**, pp. 823–829.

HACHEY, B., BEATRICE ALEX AND MARKUS BECKER (2005) Investigating the effects of selective sampling on the annotation task. **Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)**, pp. 144–151.

HE, J., MINGJING LI, HONG-JIANG ZHANG, HANGHANG TONG, CHANGSHUI ZHANG (2004) Mean Version Space: a New Active Learning Method for Content-Based Image Retrieval. pp.

HE, X., AND DENG CAI (2009) Active Subspace Learning. **IEEE 12th International Conference on Computer Vision (ICCV)**, pp.

HOI, C.-H., CHI-HANG CHAN, KAIZHU HUANG, MICHEL R. LYU AND IRWIN KING (2004) Biased Support Vector Machine for Relevance Feedback in Image Retrieval. **Proc. Int'l Joint Conf. Neural Networks**, pp. 3189-3194.

HOI, S. C. H., RONG JIN AND MICHAEL R. LYU (2006) Large-scale text categorization by batch mode active learning. **Proceedings of the 15th International World Wide Web Conference (WWW 2006)**, pp. 633–642.

HOI, S. C. H., RONG JIN, AND MICHAEL R. LYU (2009) Batch Mode Active Learning with Applications to Text Categorization and Image Retrieval. **IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**, 21(9) pp. 1233-1248.

HOI, S. C. H., RONG JIN, JIANKE ZHU AND MICHAEL R. LYU (2008) Semi-Supervised SVM Batch Mode Active Learning with Applications to Image Retrieval. **ACM Transactions on Information Systems**, pp.

HOI, S. C. H. A. M. R. L. (2005) A Semi-Supervised Active Learning Framework for Image Retrieval. **In Proc. CVPR 2005**, pp.

HONG, P., QI TIAN, THOMAS S. HUANG (2000) Incorporate Support Vector Machines to Content-Based Image Retrieval with Relevance Feedback. **Proceedings of IEEE 2000 International Conference on Image Processing (ICIP'2000)**, 3(pp. 750-753.

HU, R., BRIAN MAC NAMEE, AND SARAH JANE DELANY (2008) Sweetening the Dataset: Using Active Learning to Label Unlabelled Datasets. pp.

HU, R., SARAH JANE DELANY, AND BRIAN MAC NAMEE (2010) EGAL: Exploration Guided Active Learning for TCBR. **Proceedings of ICCBR**, pp. 156-170.

HUANG, X., HUANG, Y. R., WEN, M., AN, A., LIU, Y., & POON, J. (2006) Applying data mining to pseudo-relevance feedback for high performance text retrieval. **In Proceedings of the 6th IEEE international conference on data mining**, pp. 295-306.

HULL, D. (1993) Using statistical testing in the evaluation of retrieval experiments. **In Proceedings of the 16th ACM Conference on Research and Development in Information Retrieval (SIGIR´93)**, pp. 329-338.

HWA, R., MILES OSBORNE, ANOOP SARKAR AND MARK STEEDMAN (2003) Corrected co-training for statistical parsers. **Proceedings of the ICML Workshop on the "Continuum from Labeled to Unlabeled Data"**, pp. 95-102.

IDE, E. (1971) New experiments in relevance feedback. **In The SMART Retrieval System: Experiments in Automatic Document Processing**, pp. 337–354.

JONES, R., RAYID GHANI, TOM MITCHELL AND ELLEN RILOFF (2003) Active learning for information extraction with multiple view feature sets. **Proceedings of the 20th International Conference on Machine Learning (ICML 2003)**, pp. 21-24.

KIM, S., YU SONG, KYUNGDUK KIM, JEONG-WON CHA AND GARY GEUNBAE LEE (2006) MMR-based active machine learning for bio named entity recognition. **Proceedings of the Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL 2006)**, pp. 69–72.

KORFHAGE, R. R. (1997) Information Storage and Retrieval. **New York. Wiley Computer Publishing**, pp.

KULLBACK, S. A. R. A. L. (1951) On information and sufficiency. **Annals of Mathematical Statistics**, 22(pp. 79–86.

KUO, J.-S., HAIZHOU LI AND YING-KUEI YANG (2006) Learning transliteration lexicons from the web. **Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association of Computational Linguistics**, pp. 1129–1136.

LAWS, F. A. H. S. (2008) Stopping criteria for active learning of named entity recognition. **Proceedings of the 22nd International Conference on Computational Linguistics (CoLing)**, pp. 465-472.

LEWIS, D. D. (1995) A sequential algorithm for training text classifiers: Corrigendum and additional data. ACM SIGIR Forum 29 2(pp. 13-19.

LEWIS, D. D. A. W. A. G. (1994) A Sequential Algorithm for Training Text Classifiers. **Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval**, pp. 3–12.

LI, M., HANG LI, ZHI-HUA ZHOU (2008) Semi-Supervised Document Retrieval. **Preprint submitted to Information Processing & Management**, pp.

LI, X., & LIU, B. (2003) Learning to classify text using positive and unlabeled data. **In: Proceedings of the 19th international joint conference on artificial intelligence**, pp. 587–594.

LIERE, R. A. P. T. (1997) Active learning with committees for text categorization. **Proceedings of the fourteenth national conference on artificial intelligence**, pp. 591–597.

LIERE, R. A. P. T. (1998) Active Learning with Committees in Text Categorization: Preliminary Results in Comparing Winnow and Perceptron. **AAAI Technical Report WS-98-05. American Association for Artificial Intelligence**, pp.

LIU, B., DAI, Y., LI, X., LEE, W. S., & YU, P. S. (2003) Building text classifiers using positive and unlabeled examples. **In Proceedings of the third IEEE international conference on data mining**, pp. 179–188.

LIU, R., Y. WANG, T. BABA, D. MASUMOTO, S. NAGATA (2008) SVM-based active feedback in image retrieval using clustering and unlabeled data. **Pattern Recognition 41**, 8(pp. 2645–2655.

LUHN, H. P. (1958) The automatic creation of literature abstracts. **IBM Journal of Research and Development**, 2(pp. 59-165.

MACKAY, D. J. C. (1992) Information-based objective functions for active data selection. **Neural Compution 4**, 4(pp. 590-604.

MAES, P. (1994) Agents that Reduce Work and Information Overload. **Communications of the ACM**, 37(pp. 30-40.

MANDEL, M. I., GRAHAM E. POLINER AND DANIEL P.W. ELLIS (2006) Support Vector Machine Active Learning for Music Retrieval. **Multimedia Systems manuscript**, pp. 3–13.

MANNING, D., RAGHAVAN, P. AND SCHÜTZE, H. (2008) An Introduction to Information Retrieval. **Cambridge University Press, England**, pp.

MARON, O. A. T. L.-P. (1998) A framework for multiple-instance learning. **In Advances in Neural Information Processing Systems (NIPS)**, 10(pp. 570–576.

MARTINEZ, F. J. Y. R., J.V. (2004) Reflexiones sobre la Evaluación de los Sistemas de Recuperación de Información: Necesidad, Utilidad y Viabilidad. **Anales de Documentación**, 7(pp. 153-170.

MCCALLUM, A. A. K. N. (1998a) Employing EM and pool-based active learning for text classification. **Proceedings of the 15th International Conference on Machine Learning (ICML-98)**, pp. 350–358.

MCCALLUM, A. A. K. N. (1998b) Pool-Based Active Learning for Text Classification. **Conf. Automated Learning and Discovery (CONALD-98)**, pp.

MELVILLE, P., AND RAYMOND J. MOONEY (2004) Diverse Ensembles for Active Learning. **In Proceedings of the 21st International Conference on Machine Learning, Banff, Canada**, pp.

MELVILLE, P., S. M. YANG, M. SAAR-TSECHANSKY, AND R. MOONEY (2005) Active learning for probability estimation using Jensen-Shannon divergence. **In Proceedings of the European Conference on Machine Learning (ECML)**, pp. 268–279.

MIDDLETON, C., AND R. BAEZA-YATES, (2007) A Comparison of Open Source Search Engines. pp.

MOSKOVITCH., R., N. NISSIM, D. STOPEL, C. FEHER, R. ENGLERT, AND Y. ELOVICI (2007) Improving the detection of unknown computer worms activity using active learning. **In Proceedings of the German Conference on AI**, pp. 489–493.

NGUYEN, H. T., ARNOLD SMEULDERS (2004) Active learning using pre-clustering. **Proceedings of the 21st International Conference on Machine Learning (ICML)**, pp. 623-630.

NIGAM, K., MCCALLUM, A. K., THRUN, S., & MITCHELL, T. (2000a) Text classification from labeled and unlabeled documents using em. **Machine Learning**, 39(pp. 103–134.

NIGAM, K. A. R. G. (2000b) Analyzing the effectiveness and applicability of co-training. **Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM 2000)**, pp. 86–93.

ONODA, T., HIROSHI MURATA AND SEIJI YAMADA (2004) Relevance Feedback Document Retrieval using Non-Relevant Documents. pp.

ONODA, T., HIROSHI MURATA AND SEIJI YAMADA (2005) Relevance Feedback Document Retrieval Using Support Vector Machines. **Springer-Verlag Berlin Heidelberg 2005**, pp. 59–73.

ONODA, T., HIROSHI MURATA AND SEIJI YAMADA (2006) Non-relevance feedback document retrieval based on one class SVM and SVDD. **IEEE International Joint Conference on Neural Networks**, pp. 1212–1219.

ONODA, T. A. H. M. (2002) Interactive Document Retrieval with Active Learning. pp.

OSBORNE, M. A. J. B. (2004) Ensemble-based active learning for parse selection. **Proceedings of Human Language Technology Conference – the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL 2004)**, pp. 89–96.

OSUGI, T., KUN, D., AND SCOTT, S. (2005) Balancing exploration and exploitation: A new algorithm for active machine learning. **Proc. of ICDM '05**, pp. 330-337.

PAASS, G. A. J. K. (1995) Bayesian query construction for neural network models. **In Advances in Neural Information Processing Systems (NIPS)**, 7(pp. 443–450.

PATAN-E, G. A. M. R. (2001) The enhanced LBG algorithm. **IEEE Transactions on Neural Networks**, 14(9) pp. 1219-1237.

PEÑA, R., BAEZA-YATES, R. AND RODRIGUEZ, J. V. (2003) Gestión Digital de la Información. **Alfaomega Grupo Editor**, pp.

PIRAS, L. (2011) Interactive search techniques for content-based retrieval from archives of images. *Department of Electrical and Electronic Engineering.* University of Cagliari.

PORTER (1980) An algorithm for suffix stripping. 14(3) pp. 130-137.

PURPURA, S., CLAIRE CARDIE AND JESSE SIMONS (2008) Active Learning for e-Rulemaking: Public Comment Categorization. pp.

RAHMANI, R. A. S. A. G. (2006) MISSL:Multiple-instance semi-supervised learning. **In Proceedings of the International Conference on Machine Learning (ICML)**, pp. 705–712.

RAY, S. A. M. C. (2005) Supervised versus multiple instance learning: An empirical comparison. **In Proceedings of the International Conference on Machine Learning (ICML)**, pp. 697–704.

REICHART, R. A. A. R. (2007) An ensemble method for selection of high quality parses. **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)**, pp. 408–415.

RINGGER, E., PETER MCCLANAHAN, ROBBIE HAERTEL, GEORGE BUSBY, MARC CARMEN, JAMES CARROLL, KEVIN SEPPI AND DERYLE LONSDALE (2007) Active learning for part-of-speech tagging: Accelerating corpus annotation. **Proceedings of the Linguistic Annotation Workshop**, pp. 101–108.

ROBERTSON, S. E., H. ZARAGOZA AND M. TAYLOR (2003) Microsoft Cambridge at TREC-12: HARD track. **Proceedings of TREC 2003**, pp. 500-255.

ROBERTSON, S. E. A. K. S. J. (1977) Relevance weighting of search terms. **Journal of the American Society for Informaiton Science**, 27(pp.

ROCCHIO, J. J. (1971) Relevance feedback in information retrieval. **In The SMART Retrieval System: Experiments in Automatic Document Processing**, pp. 313-323.

ROY, N. A. A. M. (2001) Toward optimal active learning through sampling estimation of error reduction. **Proceedings of the International Conference on Machine Learning (ICML)**, pp. 441–448.

RUI, Y., T. S. HUANG, M. ORTEGA AND S. MEHROTRA (1998) Relevance feedback: A power tool for interactive content-based image retrieval. **IEEE Trans. Circuits Syst. Video Technol.**, 8(pp. 644–655.

SALTON, G., AND BUCKLEY, C. (1990) Improving retrieval performance by relevance feedback. **Journal of the American Society for Information Science**, 41(pp. 288-297.

SALTON, G. E. (1971) The SMART Retrieval System: Experiments in Automatic Document Processing. **Prentice Hall In. Englewood Cliffs, NJ**, pp.

SALTON, G. Y. M. J. M. (1983) Introduction to Modern Information Retrieval. **McGraw-Hill**, pp.

SASSANO, M. (2002) An empirical study of active learning with support vector machines for Japanese word segmentation. **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 505–512.

SCHEFFER, T., CHRISTIAN DECOMAIN AND STEFAN WROBEL (2001) Active hidden Markov models for information extraction. **Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis (IDA-2001)**, pp. 309–318.

SCHEIN, A. I. A. L. H. U. (2007) Active learning for logistic regression: An evaluation. **Machine Learning**, 3(pp. 235–265.

SCHOHN, G. A. D. C. (2000) Less is more: Active learning with support vector machines. **Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)**, pp. 839–846.

SEBASTIANI, F. (2002) Machine Learning in Automated Text Categorization. **ACM Computing Surveys**, 34(pp. 1-47.

SEGAL, R., TED MARKOWITZ AND WILLIAM ARNOLD (1994) Fast Uncertainty Sampling for Labeling Large E-mail Corpora. pp.

SETTLES, B. (2010) Active Learning Literature Survey. University of Wisconsin–Madison.
SETTLES, B., MARK CRAVEN AND SOUMYA RAY (2008a) Multiple-instance active learning. **In Advances in Neural Information Processing Systems (NIPS)**, 20(pp. 1289–1296.
SETTLES, B. A. M. C. (2008b) An analysis of active learning strategies for sequence labeling tasks. **Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1069–1078.
SEUNG, H. S., M. OPPER AND H. SOMPOLINSKY (1992) Query by committee. **Proceedings of the ACMWorkshop on Computational Learning Theory**, pp. 287–294.
SHEN, D., JIE ZHANG, JIAN SU, GUODONG ZHOU AND CHEW-LIM TAN (2004) Multi-criteria-based active learning for named entity recognition. **Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)**, pp. 589–596.
SHEN, X. (2007) User-Centered Adaptive Information Retrieval. Urbana, Illinois, University of Illinois at Urbana-Champaign.
SHEN, X. A. C. Z. (2003) Active feedback–UIUC TREC2003 HARD experiments. **Proceedings of TREC 2003**, pp.
SHEN, X. A. C. Z. (2005) Active feedback in ad hoc information retrieval. **Proc. ACM SIGIR'05**, pp. 59-66.
SIEGELMANN, H. A. O. L. (2005) Introducing an Active Cluster-Based Information Retrieval Paradigm. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY**, 10(pp. 1024–1030.
SIEGELMANN, H. A. T. J. (2001) Active information retrieval. **Proceedings of NIPS 2001**, 14(pp. 777-784.
STECK, H. (2007) Hinge rank loss and the area under the ROC curve. **Proceedings of the European Conference on Machine Learning (ECML'07)**, pp. 347–358.
STEEDMAN, M., REBECCA HWA, STEPHEN CLARK, MILES OSBORNE, ANOOP SARKAR, JULIA HOCKENMAIER, PAUL RUHLEN, STEVEN BAKER AND JEREMIAH CRIM (2003) Example selection for bootstrapping statistical parsers. **Proceedings of Human Language Technology Conference – North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL 2003)**, pp. 157–164.
SUÁREZ, A. (2005) Aprendizaje Automático.
TANG, M., XIAOQIANG LUO AND SALIM ROUKOS (2002) Active learning for statistical natural language parsing. **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)**, pp. 120–127.
THOMPSON, C. A., MARY ELAINE CALIFF AND RAYMOND J. MOONEY (1999) Active learning for natural language parsing and information extraction. **Proceedings of the Sixteenth International Machine Learning Conference (ICML-99)**, pp. 406–414.
TIAN, A. A. M. L. (2011) Active Learning to Maximize Accuracy vs. Effort in Interactive Information Retrieval. **SIGIR'11**, pp.
TOLOSA, G. H. Y. F. R. A. B. Introducción a la Recuperación de Información Conceptos, modelos y algoritmos básicos. Universidad Nacional de Luján. Argentina.
TONG, S. (2001a) Active Learning: Theory and Applications. *Deparment of Computer Science.* Stanford University.
TONG, S. A. D. K. (2001b) Support vector machine active learning with applications to text classification. **Proceedings of the International Conference on Machine Learning (ICML)**, pp. 999–1006.

TONG, S. A. E. C. (2001c) Support vector machine active learning for image retrieval. **Proceedings of the 9th ACM Intl. Conf. on Multimedia**, pp. 107-118.

TUR, G., DILEK HAKKANI-TUR AND ROBERT E. SCHAPIRE (2003) Active learning for spoken language understanding. **Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 276-279.

TUR, G., DILEK HAKKANI-TÜR AND ROBERT E. SCHAPIRE (2005) Combining active and semi-supervised learning for spoken language understanding. **Speech Communication 45**, 2(pp. 171–186

VAN RIJSBERGEN, C. J. (1999) Information Retrieval. **2nd edition. Butterworths**, pp.

VAPNIK, V. A. C., C. (1995) Support-Vector Networks. **Machine Learning**, 20(pp. 273-297.

VILLENA, R., J. (1997) **Sistema de Recuperación de Información**. Valladolid: Departamento Ingeniería Sistemas Telmáticos, Universidad [internet], pp. Disponible en <http://www.mat.upm.es/jmp/doct00/RecupInfo.pdf> [Consultado:4 de febrero de 2011].

VLACHOS, A. (2006) Active annotation. **Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)**, pp. 64–71.

WILCOXON, F. (1945) Individual comparisons by ranking methods. **Biometrics**, 1(pp. 80–83.

WU, H., GUANG QIU, XIAOFEI HE, YUAN SHI, MINGCHENG QU, JING SHEN, JIAJUN BU AND CHUN CHEN (2009) Advertising Keyword Generation Using Active Learning. **WWW 2009 Madrid**, pp.

WU, W.-L., RU-ZHAN LU, JIAN-YONG DUAN, HUI LIU, FENG GAO AND YU-QUAN CHEN (2006) A weakly supervised learning approach for spoken language understanding. **Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)**, pp. 199–207.

XIANG, L. Y. A. Z. C. (2009) Adaptive Relevance Feedback in Information Retrieval. pp.

XIE, H. A. A. O. (2004) An User Preference Information Based Kernel for SVM Active Learning in Content-based Image Retrieval. pp.

XU, Z., KRISTIAN KERSTING, AND THORSTEN JOACHIMS (2010) Fast Active Exploration for Link-Based Preference Learning using Gaussian Processes. pp.

XU, Z., RAM AKELLA AND YI ZHANG (2007) Incorporating diversity and density in active learning for relevance feedback. **Proceedings of the European Conference on IR Research (ECIR)**, pp. 246–257.

XU, Z., XIAOWEI XU, KAI YU, AND VOLKER TRESP (2003) A Hybrid Relevance-Feedback Approach to Text Retrieval. pp.

XU, Z. A. R. A. (2008) Active Relevance Feedback for Difficult Queries. **Proc.of CIKM**, pp.

YU, H. (2005) SVM selective sampling for ranking with application to data retrieval. **Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)**, pp. 354–363.

YU, K., J. BI, AND V. TRESP (2006) Active learning via transductive experimental design. **In ICML'06**, pp.

ZHANG, L., FUZONG LIN AND BO ZHANG (2001) Support Vector Machine Learning for image retrieval. **Proceedings of the International Conference on Image Processing (ICIP 2001)**, 2(pp. 721-724.

ZHANG, T. A. F. J. O. (2000) A probability analysis on the value of unlabeled data for classification problems. **Proceedings of the International Conference on Machine Learning (ICML)**, pp. 1191–1198.

ZHANG, Y., WEI XU AND JAMIE CALLAN (2003) Exploration and exploitation in adaptive filtering based on Bayesian active learning. **Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)**, pp.

ZHOU, Z.-H., KE-JIA CHEN AND HONG-BIN DAI (2006) Enhancing Relevance Feedback in Image Retrieval Using Unlabeled Data. pp. 1-25.

ZHOU, Z.-H., KE-JIA CHEN AND JIANG, Y. (2004) Exploiting unlabeled data in content-based image retrieval. **Proceedings of the 15th European Conference on Machine Learning. Pisa, Italy**, pp. 525-536.

ZHU, J., HUIZHEN WANG AND EDUARD HOVY (2008a) Learning a stopping criterion for active learning for word sense disambiguation and text classification. **Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)**, pp. 366–372.

ZHU, J., HUIZHEN WANG, TIANSHUN YAO AND BENJAMIN K. TSOU (2008b) Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. **Proceedings of the 22nd International Conference on Computational Linguistics (CoLing)**, pp. 1137-1144.

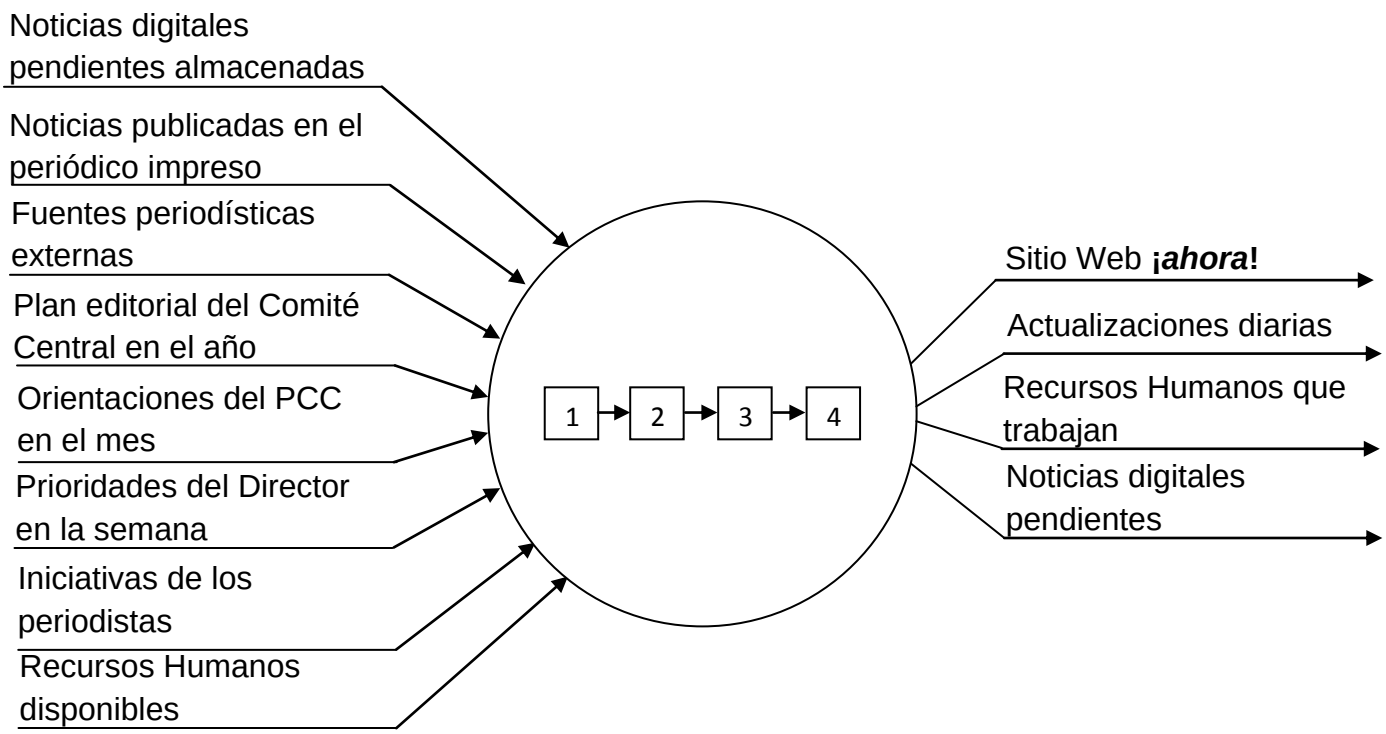
ZHU, J. A. E. H. (2007) Active learning for word sense disambiguation with methods for addressing the class imbalance problem. **Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**, pp. 783–790.

ZHU, X., JOHN LAFFERTY AND ZOUBIN GHAHRAMANI (2003) Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. **Proceedings of the ICML Workshop on the Continuum from Labeled to Unlabeled Data**, pp. 58–65.

ZIPF, G. K. (1949) Human Behaviour and the Principle of Least Effort. **Reading, MA: Addison- Wesley Publishing Co**, pp.

ANEXOS

Anexo 1: Mapa de proceso del flujo editorial del periódico ¡*ahora*! digital



Leyenda:

1. Planificar por parte de la **Editora Web Principal** los temas noticiosos a desarrollar.
2. Asignar por la **Editora Web Principal** las tareas a periodistas y/o fotógrafos.
   a) En ocasiones, la editora requiere antes de asignar buscar en el sitio los periodistas que redactaron noticias análogas a las planificadas.
3. Elaborar los periodistas y/fotógrafos tareas a desarrollar.
   a) Antes de iniciar el desarrollo de la tarea asignada los periodistas tienden a buscar en el sitio trabajos similares (ya publicados) para realizar la tarea en breve tiempo y con una calidad deseada por la **Editora Web Principal**.
4. Revisar y corregir la **Editora Web** (o **Editora Web Principal**) los artículos desarrollados.
5. Publicar la **Editora Web Principal** los textos periodísticos culminados.
   a) De decidirse por los directivos o **Editora Web Principal** cubrir espacios noticiosos libres o cambiar en breve tiempo alguna de las noticias a publicar, la editora busca

la oportuna entre las almacenadas (pendientes), así como las publicadas recientemente en el periódico impreso y fuentes periodística externas.

Anexo 2: Guía para entrevista de diagnóstico de la satisfacción de los entes inmersos en el proceso editorial del periódico ¡*ahora*! digital

(Fuente: elaboración propia)

La presente entrevista tiene como objetivo conocer cómo se sienten los trabajadores inmersos en el proceso editorial del periódico **¡*ahora*!** digital en la Casa Editora **¡*ahora*!**.

Su aporte será vital para diagnosticar cómo se encuentra el proceso editorial del periódico **¡*ahora*!** digital dentro del centro y proponer una solución que permita un correcto funcionamiento del mismo.

1. ¿Se siente usted contento con el trabajo que realiza dentro del proceso editorial del periódico ¡*ahora*!?
2. ¿Cómo es la comunicación y relaciones humanas con sus compañeros de trabajo?
3. ¿Qué parte de su trabajo es la más complicada y quisiera sintetizar?
4. ¿Cómo son las condiciones de trabajo en su puesto laboral?
5. ¿Se siente estimulado con la labor que desempeña?
6. ¿Qué sugiere para mejorar el proceso editorial del periódico **¡*ahora*!** digital en la Casa Editora **¡*ahora*!**?
7. ¿Quisiera argumentar algún otro aspecto de interés relacionado con el proceso editorial?

Anexo 3: Interfaz visual para realizar el proceso de búsqueda de información en el periódico ¡ahora! digital

Anexo 4: Algoritmo de la estrategia de consulta por comité para un protocolo de muestreo selectivo basado en flujo

Entrada: $\epsilon > 0$ – el error de predicción tolerable máximo.

$\delta > 0$ – la confianza deseada.

Gibbs – un ***oracle*** que determina las predicciones Gibbs.

Muestra – un ***oracle*** que genera ejemplos no etiquetados.

Etiqueta – un ***oracle*** que genera la etiqueta correcta de un ejemplo.

1. Llama ***Muestra*** para obtener aleatoriamente un ejemplo no etiquetado x.
2. Llama ***Gibbs*** dos veces, para obtener dos predicciones para la etiqueta de x.
3. **Si** las dos predicciones son iguales, entonces rechaza el ejemplo y retorna al inicio del ciclo (paso 1).
4. **Si no**, llama ***Etiqueta*** para obtener la etiqueta correcta de x, y adiciona el ejemplo etiquetado al conjunto de ejemplos etiquetados.

Los parámetros ϵ y δ son proporcionados al algoritmo de aprendizaje como entrada y usados para determinar el criterio de parada.

Anexo 5: Algoritmo de la estrategia de consulta muestra de incertidumbre para un protocolo basado en fondo

Entrada: $\boldsymbol{T}$ – conjunto de entrenamiento de ejemplos etiquetados.

$\boldsymbol{U}$ – conjunto de ejemplos no etiquetados.

1. Entrena el clasificador utilizando T.
2. Mientras el anotador quiera etiquetar ejemplos
    a) Aplica el clasificador actual a cada ejemplo de U.
    b) Encuentra los $\boldsymbol{b}$ ejemplos para los cuales el clasificador es menos confiable.
    c) Obtiene la etiqueta del anotador de la submuestra de $\boldsymbol{b}$ ejemplos.
    d) Entrena nuevamente el clasificador con todos los ejemplos etiquetados.

Anexo 6: Algoritmo de la estrategia de consulta reducción esperada del error para un protocolo basado en fondo

Entrada: $\boldsymbol{T}$ – conjunto de entrenamiento de ejemplos etiquetados.

$\boldsymbol{U}$ – conjunto de ejemplos no etiquetados.

1. Entrena el clasificador utilizando $\boldsymbol{T}$.
2. Considera cada posible etiqueta $\boldsymbol{y}$ para $\boldsymbol{x}$ ejemplo no etiquetado en el fondo $\boldsymbol{U}$ y adiciona el par $(\boldsymbol{x}, \boldsymbol{y})$ a $\boldsymbol{T}$.
3. Re-entrena el clasificador con el conjunto de entrenamiento aumentado, $\boldsymbol{T}^* = \boldsymbol{T} + (\boldsymbol{x}, \boldsymbol{y})$.
4. Estima la pérdida esperada que resulta como en la ecuación (2.5) ó (2.6).
5. Asigna a $\boldsymbol{x}$ la media de la pérdida esperada para cada posible etiquetamiento $\boldsymbol{y}$, pesado acorde al clasificador actual.
6. Selecciona para el etiquetamiento el ejemplo $\boldsymbol{x}$ no etiquetado que generó el más bajo error esperado en todos los otros ejemplos.

Anexo 7: Resultados experimentales del desempeño de los sistemas de recuperación de información

Tablas obtenidas por el SPSS relacionadas con el Experimento 1

Prueba de Friedman

Rangos

	Rango promedio
SRI_ROCCHIO	4,10
SRI_C	3,87
SRI_QBC_P	3,51
SRI_QBC_F	3,51
SRI_NAIVE	3,02
SRI_UNC	2,99

Estadísticos de contraste(a)

N			83
Chi-cuadrado			23,894
gl			5
Sig. asintót.			,000
Sig. Monte Carlo	Sig.		,000
	Intervalo de confianza de 95%	Límite inferior	,000
		Límite superior	,001

a Prueba de Friedman

Prueba de Wilcoxon

Algoritmos	SRI_ROCCHIO	SRI_C	SRI_QBC_P	SRI_QBC_F	SRI_UNC	SRI_NAIVE
SRI_ROCCHIO						
SRI_C	0,2574					
SRI_QBC_P	0	0,0114				
SRI_QBC_F	0	0,0155	0,4822			
SRI_UNC	0	0,0021	0,021	0,0038		
SRI_NAIVE	0,0004	0,0089	0,039	0,0036	0,2171	

Diferencias significativas.

Diferencias altamente significativas.

Diferencias medianamente significativas.

Tablas obtenidas por el SPSS relacionadas con el Experimento 2

Prueba de Friedman

Rangos

	Rango promedio
SRI_C	3,95
SRI_UNC	3,57
SRI_NAIVE	3,43
SRI_QBC_F	3,40
SRI_QBC_P	3,39
SRI_ROCCHIO	3,27

Estadísticos de contraste(a)

N			83
Chi-cuadrado			15,132
gl			5
Sig. asintót.			,010
Sig. Monte Carlo	Sig.		,008
	Intervalo de confianza de 95%	Límite inferior	,006
		Límite superior	,010

a Prueba de Friedman

Prueba de Wilcoxon

Algoritmos	SRI_C	SRI_UNC	SRI_NAIVE	SRI_QBC_F	SRI_QBC_P	SRI_ROCCHIO
SRI_C						
SRI_UNC	0,172					
SRI_NAIVE	0,3232	0,3662				
SRI_QBC_F	0,1383	0,0723	0,2547			
SRI_QBC_P	0,1383	0,0723	0,2741	0,5018		
SRI_ROCCHIO	0,0002	0,0031	0,0062	0,02	0,0204	

Diferencias significativas.

Diferencias altamente significativas.

Diferencias medianamente significativas.

Tablas obtenidas por el SPSS relacionadas con el Experimento 3

Prueba de Friedman

Rangos

	Rango promedio
SRI_QBC_P	5,66
SRI_NAIVE	4,64
SRI_UNC	4,02
SRI_C	2,49
SRI_ROCCHIO	2,27
SRI_QBC_F	1,92

Estadísticos de contraste(a)

N			80
Chi-cuadrado			262,965
gl			5
Sig. asintót.			,000
Sig. Monte Carlo	Sig.		,000
	Intervalo de confianza de 95%	Límite inferior	,000
		Límite superior	,000

a Prueba de Friedman

Prueba de Wilcoxon

Algoritmos	SRI_QBC_P	SRI_NAIVE	SRI_UNC	SRI_C	SRI_ROCCHIO	SRI_QBC_F
SRI_QBC_P						
SRI_NAIVE	0					
SRI_UNC	0	0				
SRI_C	0	0	0			
SRI_ROCCHIO	0	0	0	0,03		
SRI_QBC_F	0	0	0	0,0176	0,0038	

Diferencias significativas.

Diferencias altamente significativas.

Diferencias medianamente significativas.

Tablas obtenidas por el SPSS relacionadas con el Experimento 4

Prueba de Friedman

Rangos

	Rango promedio
SRI_C	3,99
SRI_ROCCHIO	3,91
SRI_QBC_P	3,55
SRI_QBC_F	3,52
SRI_UNC	3,25
SRI_NAIVE	2,77

Estadísticos de contraste(a)

N			83
Chi-cuadrado			54,242
gl			5
Sig. asintót.			,000
Sig. Monte Carlo	Sig.		,000
	Intervalo de confianza de 95%	Límite inferior	,000
		Límite superior	,000

a Prueba de Friedman

Prueba de Wilcoxon

Algoritmos	SRI_C	SRI_ROCCHIO	SRI_QBC_P	SRI_QBC_F	SRI_UNC	SRI_NAIVE
SRI_C						
SRI_ROCCHIO	0,1896					
SRI_QBC_P	0,1383	0,3484				
SRI_QBC_F	0,1345	0,3473	0,2489			
SRI_UNC	0,0038	0,0896	0,0175	0,0189		
SRI_NAIVE	0	0,0036	0	0	0,0033	

Diferencias significativas.

Diferencias altamente significativas.

Diferencias medianamente significativas.

Tablas obtenidas por el SPSS relacionadas con el Experimento 5

Prueba de Friedman

Rangos

	Rango promedio
SRI_NAIVE	4,53
SRI_QBC_P	4,11
SRI_QBC_F	4,08
SRI_UNC	4,04
SRI_ROCCHIO	2,13
SRI_C	2,11

Estadísticos de contraste(a)

N			80
Chi-cuadrado			173,535
gl			5
Sig. asintót.			,000
Sig. Monte Carlo	Sig.		,000
	Intervalo de confianza de 95%	Límite inferior	,000
		Límite superior	,000

a Prueba de Friedman

Prueba de Wilcoxon

Algoritmos	SRI_NAIVE	SRI_QBC_P	SRI_QBC_F	SRI_UNC	SRI_ROCCHIO	SRI_C
SRI_NAIVE						
SRI_QBC_P	0					
SRI_QBC_F	0	0,4693				
SRI_UNC	0,0001	0,3128	0,339			
SRI_ROCCHIO	0	0	0	0		
SRI_C	0	0	0	0	0,2594	

Diferencias significativas.

Diferencias altamente significativas.

Diferencias medianamente significativas.

Tablas obtenidas por el SPSS relacionadas con el Experimento 6

Prueba de Friedman

Rangos

	Rango promedio
SRI_ UNC	3,76
SRI_ QBC_P	3,74
SRI_NAIVE	3,73
SRI_C	3,53
SRI_ QBC_F	3,39
SRI_ROCCHIO	2,84

Estadísticos de contraste(a)

N			83
Chi-cuadrado			56,803
gl			5
Sig. asintót.			,000
Sig. Monte Carlo	Sig.		,000
	Intervalo de confianza de 95%	Límite inferior	,000
		Límite superior	,000

a Prueba de Friedman

Prueba de Wilcoxon

Algoritmos	SRI_UNC	SRI_QBC_P	SRI_NAIVE	SRI_C	SRI_QBC_F	SRI_ROCCHIO
SRI_UNC						
SRI_QBC_P	0,5432					
SRI_NAIVE	0,0162	0,3747				
SRI_C	0,0011	0,0151	0,0073			
SRI_QBC_F	0,0009	0,0051	0,0033	0,0151		
SRI_ROCCHIO	0,0008	0	0	0,0003	0	

Diferencias significativas.

Diferencias altamente significativas.

Diferencias medianamente significativas.

Tablas obtenidas por el SPSS relacionadas con el Experimento 7

Prueba de Friedman

Rangos

	Rango promedio
SRI_NAIVE	4,82
SRI_UNC	4,56
SRI_QBC_P	3,98
SRI_QBC_F	3,95
SRI_C	1,94
SRI_ROCCHIO	1,75

Estadísticos de contraste(a)

N			79
Chi-cuadrado			252,031
gl			5
Sig. asintót.			,000
Sig. Monte Carlo	Sig.		,000
	Intervalo de confianza de 95%	Límite inferior	,000
		Límite superior	,000

a Prueba de Friedman

Prueba de Wilcoxon

Algoritmos	SRI_NAIVE	SRI_UNC	SRI_QBC_P	SRI_QBC_F	SRI_C	SRI_ROCCHIO
SRI_NAIVE						
SRI_UNC	0,0001					
SRI_QBC_P	0	0				
SRI_QBC_F	0	0	0,3223			
SRI_C	0	0	0	0		
SRI_ROCCHIO	0	0	0	0	0,0276	

Diferencias significativas.

Diferencias altamente significativas.

Diferencias medianamente significativas.

Tablas obtenidas por el SPSS relacionadas con el Experimento 8

Prueba de Friedman

Rangos

	Rango promedio
SRI_QBC_F	3,73
SRI_QBC_P	3,70
SRI_C	3,55
SRI_NAIVE	3,55
SRI_ROCCHIO	3,48
SRI_UNC	2,98

Estadísticos de contraste(a)

N			83
Chi-cuadrado			34,091
gl			5
Sig. asintót.			,000
Sig. Monte Carlo	Sig.		,000
	Intervalo de confianza de 95%	Límite inferior	,000
		Límite superior	,000

a Prueba de Friedman

Prueba de Wilcoxon

Algoritmos	SRI_QBC_F	SRI_QBC_P	SRI_C	SRI_NAIVE	SRI_ROCCHIO	SRI_UNC
SRI_QBC_F						
SRI_QBC_P	0,5022					
SRI_C	0,0412	0,1011				
SRI_NAIVE	0,0412	0,1011	0,1032			
SRI_ROCCHIO	0,0424	0,0665	0,1941	0,1881		
SRI_UNC	0	0,0002	0,0012	0,0012	0,1474	

Diferencias significativas.

Diferencias altamente significativas.

Diferencias medianamente significativas.