UNIVERSIDAD SAN FRANCISCO DE QUITO

Módulo para Clasificación Automática y Temática de Páginas Web

María del Cisne García Muñoz

Tesis de Grado presentada como requisito para la obtención del título de Ingeniera en Sistemas

Quito, Mayo del 2012

Universidad San Francisco de Quito
Colegio Politécnico

HOJA DE APROBACION DE TESIS

Módulo para Clasificación Automática y Temática de Páginas Web

María del Cisne García Muñoz

Enrique Vinicio Carrera, PhD ----------
Director de Tesis y
Miembro del Comité de Tesis

Fausto Pasmay, MA ----------
Coordinador de Ingeniería de Sistemas y
Miembro del Comité de Tesis

Santiago Gangotena, PhD ----------
Decano del Colegio Politécnico

Este trabajo está dedicado a mis padres por su incondicional apoyo y cariño.

Una dedicatoria especial a mi hermana, porque con sus palabras de aliento me dio el empuje final para culminar este proyecto.

AGRADECIMIENTOS

A mi madre por su apoyo constante durante toda mi carrera. Por estar ahí siempre con una taza de café o simplemente preguntándome si necesito algo. Por ser la persona que me recordó lo importante de finalizar los estudios aunque el trabajo haya ocupado gran parte del tiempo disponible.

A mi padre por su cariño y sus palabras de aliento, que desde siempre me han acompañado y me han ayudado para concluir este proceso.

A mi hermana que me acompañó en esta aventura y que de forma incondicional, entendió mis ausencias y mis malos momentos.

A mis profesores Fausto Pasmay y Vinicio Carrera, por todo el tiempo que me han dado y por sus sugerencias e ideas de las que tanto provecho he sacado.

Gracias también a mis queridos compañeros, que me apoyaron y me permitieron entrar en su vida durante estos años. Muchas gracias por sus preguntas que me han hecho crecer en conocimiento, pero más que nada les agradezco por su amistad.

Y finalmente quiero agradecer a todas aquellas personas que de una u otra forma, colaboraron o participaron en la realización de este proyecto, entregándome una sonrisa o una palabra de aliento.

RESUMEN

La web se ha transformado en uno de los medios de comunicación más utilizados en la actualidad, la mayoría de diseñadores, programadores y usuarios, trabajan con la información que se encuentra en la web. Por ello es de vital importancia la mejor utilización de los recursos disponibles que solamente se logra al contar con algoritmos que resuelvan las necesidades en el menor tiempo posible. Es aquí donde los algoritmos de clasificación juegan un papel muy importante, ya que no solo pueden mejorar la calidad de las búsquedas que se realizan, sino que también permiten optimizar los recursos que ahora se concentran en el tema adecuado. Por ello, el presente proyecto propone, mediante la utilización de coeficientes TFIDF y la técnica de embolsamiento, construir un prototipo de módulo de clasificación automática, temática, simple y eficiente de páginas web, para la integración con el sistema de búsquedas PSearch. Se busca un balance entre exactitud y tiempo de respuesta, para permitir que el sistema PSearch entregue mejores resultados a sus usuarios. Mediante la selección de técnicas de pre-procesamiento simples se quiere extraer información crítica de cada uno de los documentos HTML y posteriormente discriminarlos con una exactitud superior al 90%.

ABSTRACT

The web has become one of the most commonly used media today, most designers, programmers and users, work with the information found on the web. It is therefore vital to use the available resources in the best way that we can. This can only be achieved by having algorithms that solve the needs in the shortest time possible. This is where classification algorithms play an important role, because not only they can improve the quality of searches performed, but also to optimize the resources, that now are concentrated in the appropriate topic.

This is the reason why, this project proposes to build a Simple and efficient prototype of automatic classification, using TFIDF coefficients and bagging techniques for integration with Psearch system. The objective is to find a balance between accuracy and response time to allow the system to deliver better results to the Psearch users. The module uses simple pre-processing techniques to extract vital information from each of the HTML document and then classify them with accuracy above 90%.

CONTENIDO

LISTA DE DIAGRAMAS

LISTA DE TABLAS

CAPÍTULO 1

INTRODUCCIÓN

1.1. ANTECEDENTES

En la actualidad, el avance de la tecnología ha permitido que las personas se comuniquen, investiguen y realicen negocios sin salir de sus casas. Es por ello que cada día más y más personas utilizan el internet como el único medio para capturar clientes, obtener información, comunicarse con sus seres queridos, entre otros.

Millones de cibernautas utilizan esta gigantesca red a diario, debido a que la interfaz web presenta características que la hacen amigable, intuitiva, e independiente de arquitectura. Sin embargo al existir tal cantidad de información, el hallar aquella que se está buscando constituye un gran reto, para los usuarios. Sin embargo este reto no solo se limita a los usuarios, sino que también incluye a diseñadores de páginas web así como a desarrolladores de buscadores.

Diseñadores y Desarrolladores trabajan para lograr que el usuario final encuentre aquello que busca, sin embargo entre millones de páginas es de vital importancia definir claramente los términos que logren describir su contenido, de tal manera que la información contenida no se pierda en el ciberespacio.

Pero la definición de términos constituye el primer paso, ya que no sería útil tener términos definidos si no existe una herramienta que los utilice para entregar los resultados deseados.

De esta manera se hace imprescindible la búsqueda de un algoritmo eficiente que permita clasificar la información para facilitar las búsquedas y minimizar la utilización de recursos.

Debido a la naturaleza impredecible de los usuarios de las páginas web, es muy importante que el algoritmo sea dinámico, y que tenga una alta precisión que maximice la experiencia del usuario que en una era globalizada no tienen tiempo para errores y correcciones.

Si se encuentra un algoritmo de clasificación eficiente, se solucionaría tres problemas a la vez, ya que este ayudaría a que el usuario encuentre lo que buscaba con mayor rapidez, a que los diseñadores incluyan términos en sus encabezados y títulos que corresponden al tipo de usuarios que desean ingresen a esta páginas así como a los buscadores que van a utilizar menor cantidad de recursos al contar con algoritmos especializados e información más organizada.

Se ha desarrollado en esta área de buscadores personalizados un sistema conocido como PSearch Engine el cual cuenta con opciones para la creación de perfiles de usuario que permitan entregar información más relevante. En este sistema se cuenta con un algoritmo de clasificación, sin embargo la precisión del mismo no es la necesaria para la completa satisfacción del usuario. De esta manera un módulo de clasificación temático, simple y eficiente, se hace indispensable para cualquier buscador, en especial el antes mencionado, que trata de optimizar los recursos para ofrecer información filtrada y personalizada a sus usuarios.

1.2. IMPORTANCIA DEL PROYECTO

Debido a que la web se ha transformado en uno de los medios de comunicación más utilizados en la actualidad, la mayoría de diseñadores busca crear una interfaz que les permita no solo mostrar una información agradable al usuario, sino que permita que la velocidad a la que obtienen lo que buscan sea la mayor posible.

Sin embargo dada la cantidad y la variedad de usuarios que acceden a las páginas web, en ciertas ocasiones no es posible brindarles el tiempo de respuesta buscado. Por ello es de vital importancia la mejor utilización de los recursos disponibles que solamente se logra al contar con algoritmos que resuelvan las necesidades en el menor tiempo posible.

La mayoría de usuarios obtienen la información a través de buscadores, que presentan varias opciones de donde se tiene la libertad de escoger lo que se necesita. Sin embargo esta información presentada debe ser relevante a la búsqueda realizada y además debe presentarse en un tiempo razonable para el usuario. Es aquí donde los algoritmos de clasificación juegan un papel muy importante, ya que no solo pueden mejorar la calidad de las búsquedas sino que también permiten optimizar los recursos que ahora se concentran en el tema adecuado y no en ruido (publicidad, animaciones, etc.) que generalmente es el principal obstáculo que se tiene en las búsquedas en páginas web. Sin embargo los algoritmos de clasificación no son solo útiles para búsquedas, también se utilizan en procesos administrativos o de recuperación de información.

De esta manera al contar con un algoritmo simple y eficiente de clasificación, se trata de satisfacer la necesidad de precisión y tiempo de respuesta en un ambiente cambiante e impredecible.

1.3. OBJETIVO FINAL DEL PROYECTO

Mediante la utilización de coeficientes TFIDF y la técnica de embolsamiento, construir un prototipo del módulo de clasificación automática, temática, simple y eficiente de páginas web, para la integración con el sistema PSearch.

1.4. OBJETIVOS ESPECÍFICOS DEL PROYECTO

- Obtener información acerca de las diferentes categorías utilizadas en directorios reconocidos como Google, Bing, Yahoo Categories.
- Obtener los términos más comunes encontrados en diferentes categorías de directorios reconocidos como Google, Bing, Yahoo Categories.
- Analizar los términos obtenidos para cada categoría para obtener un conjunto pequeño que describa cada clase, basado en el coeficiente TFIDF.
- Obtener un conjunto de datos de entrenamiento constituido por enlaces de páginas Web que correspondan a las diferentes categorías definidas.
- Obtener un conjunto de datos de prueba constituido por enlaces de páginas Web que correspondan a las diferentes categorías definidas.
- Utilizar algoritmos probados como redes neuronales y algoritmos bayesianos como parte del módulo de clasificación para incrementar la eficiencia del mismo.
- Integrar el módulo de clasificación al sistema PSearch Engine, en lugar del algoritmo de clasificación simple que se encuentra utilizando actualmente.

CAPÍTULO 2

MARCO TEÓRICO

2.1. QUÉ ES MINERÍA DE DATOS (DATA MINING)?

El nombre de Minería de Datos deriva de las similitudes entre buscar valiosa información de negocios en grandes bases de datos y minar una montaña para encontrar una veta de metales valiosos. [1]

La Minería de Datos prepara, sondea y explora los datos para sacar la información oculta en ellos. Es un mecanismo de explotación con varias ventajas: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios; por otro, ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. Las herramientas de Minería de Datos integran un conjunto de áreas que tienen como propósito predecir futuras tendencias y comportamientos. En general se puede dividir a la Minería de Datos en dos grandes categorías Minería de Datos Predictiva y Minería de Datos para descubrimiento de conocimiento. [2]

Los algoritmos supervisados o predictivos predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos). Cuando una aplicación no

es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos no supervisados o del descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas. En el caso del presente proyecto, se trata de un aprendizaje predictivo, ya que el módulo de clasificación asignará la etiqueta más adecuada a la búsqueda realizada. [3][4]

2.2. QUÉ ES MINERÍA WEB (WEB MINING)?

Minería web es el proceso de descubrimiento y análisis de información relevante de los documentos de la Web. Involucra el uso de técnicas y acercamientos basados en la minería de datos (Data Mining). La minería web trata de identificar y extraer la información más relevante y específica en grandes volúmenes de datos. Busca aquella información que tiene relación entre sí. [7]

Las técnicas de minería web pueden aplicarse a diferentes estructuras de datos, desde datos completamente estructurados como tablas de una base de datos hasta datos no estructurados como texto libre, por ello constituye una invaluable ayuda en la creación de nueva información a partir de información existente, en la personalización de información así como en el aprendizaje del comportamiento de los usuarios de la web. [8]

De acuerdo con Etzioni [5] el proceso general de la minería web consiste de cuatro fases:

- *Recuperación de Información (IR)*

Primera fase de la minería web, constituye el proceso del descubrimiento automático de documentos relevantes de acuerdo a una cierta búsqueda. Los documentos relevantes incluyen, documentos disponibles en la web tales como noticias electrónicas, newsgroups, newswires, contenido de las html, etc.

- *Extracción de Información (IE)*

Segunda fase de la minería web, tiene como objetivo transformar los documentos extraídos en el proceso de recuperación de información, en documentos que sean fáciles de leer y de analizar.

- *Generalización*

Tercera fase de la minería web, incluye el reconocimiento de patrones generales de una página en particular o de diferentes páginas.

- *Análisis*

Fase final de la minería web, consiste en la interpretación de los patrones encontrados. Una vez que los patrones han sido identificados, la parte humana juega un papel importante haciendo uso de herramientas adecuadas para entender y visualizar los patrones, de tal forma que se obtenga información útil.

2.3. CLASIFICACIÓN DE LA MINERÍA WEB (WEB MINING)

Hay tres dominios de conocimiento que pertenecen a la Minería Web, ellos son: Minería de contenido web (Web Content Mining), Minería de la estructura web (Web Structure Mining) y la Minería del uso de la Web (Web Usage Mining). [7]

2.3.1. MINERÍA DE CONTENIDO WEB (WEB CONTENT MINING)

En años recientes el crecimiento de la web ha excedido todas las expectativas. Hoy en día hay billones de documentos HTML, imágenes y otros archivos multimedia disponibles en el internet, y el número sigue aumentando. Pero considerando la impresionante variedad de la web, obtener contenido valioso se ha vuelto una tarea muy difícil. El uso de la web como proveedor de información es desafortunadamente más complejo que trabajar con bases de datos estáticas. Por la naturaleza dinámica y su vasto número de documentos, las búsquedas retornan miles de resultados, por lo que es necesario tener métodos para presentar estos resultados y ayudar al usuario a seleccionar el contenido más interesante. [5].

La minería de contenido web, entonces constituye un método automático para descubrir información útil a partir del contenido de una página web o de una red de páginas. Se trabaja con todo tipo de datos (imagen, audio, video o texto).

La minería de contenido web usa las ideas y principios de la minería de datos (DM Data Mining) y del procesamiento natural de lenguaje conocida por sus siglas en inglés (NLP Natural Language Processing) y recuperación de información (IR Information Retrieval). Los principales usos para este tipo de minería de datos son: reunir, categorizar, organizar

y proveer la mejor información disponible en la red para el usuario que pide la información. [11]

La minería de contenido web tiene una aplicación muy importante en el mundo de los negocios; permite a las empresas estructurar sus páginas web de tal manera que los usuarios encuentren la información que buscan sin la necesidad de recorrer la página completa para encontrarla. También es utilizado como herramienta en el área de marketing, para determinar las palabras clave que ofrecerán mayor relevancia en búsquedas generales. [9]

2.3.2. Minería de Estructura Web (Web Structure Mining)

La Web puede revelar más información que la contenida en sus documentos. Podemos obtener información acerca de si los usuarios encuentran la información, si la estructura de un sitio es demasiado ancha o demasiado profunda, si los elementos están colocados en los lugares adecuados dentro de la página. Es por ello que la minería de estructura web intenta descubrir la arquitectura de la información, el modelo de la estructura de los links de una página web. Este modelo entonces puede utilizarse para categorizar las páginas, además nos permitiría identificar páginas similares así como páginas relacionadas. Usa la teoría de grafos, para formar representaciones gráficas de las distintas arquitecturas. [5]

Sin embargo, a través de la minería de estructura web, no solo podemos conocer la estructura de una página web, sino de los enlaces de está con otras páginas así como de la web en general. Al igual que para la arquitectura de una página web, para la estructura general de la web también se utiliza la representación de un grafo.

Algunos estudios han llegado a la conclusión de que dicho grafo tiene la forma de un corbatín, con un núcleo muy bien conectado en el centro y dos componentes laterales. El primer componente conocido como el conjunto de entrada (IN set) que se compone de páginas desde donde se puede llegar al núcleo y el conjunto de salida (OUT set) que se compone de páginas a las que se puede llegar desde el núcleo.

Otros estudios descubrieron en cambio otras relaciones importantes entre páginas web, sobre todo útiles en búsquedas. Existen dos tipos de páginas relevantes para una búsqueda: autoridades y concentradores. Las autoridades son páginas que contienen información relevante a la búsqueda, mientras que los concentradores son páginas que apuntan a buenas fuentes de información. Obviamente ambos tipos de páginas están conectados, los buenos concentradores apuntan a buenas autoridades, y las buenas autoridades son apuntadas a su vez por varios concentradores. [6]

Muchos de los buscadores se han beneficiado de esta representación particular de la web, de tal forma que el desarrollo de nuevas técnicas de la minería de estructura web resulta en el mejoramiento de la información que se le entrega al usuario que hace una búsqueda. [13]

2.3.3. Minería de Uso Web (Web Usage Mining)

La minería de uso web busca encontrar conocimiento útil de los datos secundarios obtenidos de la interacción de los usuarios con la web. Entre las fuentes más importantes donde se almacena esta interacción tenemos a los web logs, proxy logs; los cuales pueden almacenarse en diversos formatos como: Common Log, Extended log formats, etc. [5] [6]

La minería de uso web consiste de tres fases: Pre-procesamiento, descubrimiento de patrones y análisis de los mismos. [14]

- *Pre-procesamiento*

Los datos obtenidos de logs, no son aptos para la aplicación directa de algoritmos de minería de datos. Estos datos deben ser limpiados es decir debe removerse los datos irrelevantes como accesos de Web Crawlers, pedidos fallidos entre otros.

- *Descubrimiento de patrones*

Aplicando métodos estadísticos y de minería de datos a los web logs, pueden ser identificados los patrones interesantes concernientes al comportamiento de los usuarios en la navegación. Puede abordar la búsqueda de patrones de acceso general para analizar el tráfico de información en la Web.

- *Análisis de patrones*

Consiste en entender los patrones de acceso, el comportamiento y las tendencias de los usuarios, ello con el fin de reestructurar contenidos de los sitios, ubicándolos de forma más accesible o para dirigir a los usuarios a lugares concretos durante la navegación. También se pueden realizar búsquedas de uso personalizado donde se analizan las tendencias individuales de cada visitante para adaptar dinámicamente la información a partir de un perfil de usuario.

La minería de uso web se ha vuelto crítica para el manejo de las páginas web, ya que permite la creación de páginas adaptativas, personalización, servicios de soporte, etc. [15]

2.4. FILTRADO DE DATOS.

En años recientes, las búsquedas en la web se han convertido en la forma más conveniente de encontrar la información que necesitamos. Tradicionalmente los buscadores solamente trataban con las de los usuarios directamente, sin embargo, es muy común que quien busca la información no pueda especificar de forma precisa y exacta lo que necesita, esto debido a falta de entrenamiento o desconocimiento de los datos sobre los cuales se va a realizar la búsqueda.

Los términos que se ingresan en los buscadores, son generalmente muy limitados y contienen muy pocas palabras clave. Estos conjuntos de términos no son capaces de reflejar lo que el usuario busca realmente. Como resultado de ello, casi siempre lleva a problemas de sobre carga de información o a ofrecerle al usuario información que no tiene nada que ver con la búsqueda realizada. Si se le ofrece al usuario grandes cantidades de información que debe revisar manualmente, es muy probable que termine frustrado y no encuentre algo útil en la búsqueda.

Es por ello que se han implementado métodos como el filtrado de datos, para desarrollar sistemas eficientes que le den al usuario lo que busca. El filtrado de datos, ayuda a los usuarios eliminando la información irrelevante y atrayendo su atención a la información relevante. Los filtros son mediadores entre las fuentes de información y el usuario final.

El contexto en el que se realiza el filtrado de datos es dinámico, por lo que un sistema de filtrado debe ser capaz de ajustarse a los cambios de tal forma que le ofrezca al usuario la información que necesita de una manera consistente y oportuna.

El filtrado de datos implica repetidas interacciones a través de varias sesiones con los usuarios. El sistema debe recordar al usuario e individualizar su desempeño para dicho usuario. Los sistemas de filtrado de datos mantienen perfiles que son representaciones de los intereses de los usuarios. Cada vez que el usuario usa el sistema, este debe mejorar en proporcionar la información que desea, de tal forma en que las necesidades del usuario sean predecibles y consistentes. Tres enfoques han sido identificados dependiendo de la manera en que los documentos son seleccionados en el sistema. Así tenemos:

- *Sistemas Cognitivos*

  Los documentos se seleccionan basados en las características de su contenido.

- *Sistemas Sociales*

  Los documentos se seleccionan basados en las recomendaciones y anotaciones de otros usuarios.

- *Sistemas Económicos*

  Los documentos son seleccionados basados en cálculos costo-beneficio y mecanismos de fijación de precios.

El filtrado de datos se ha convertido en una herramienta de vital importancia en la web. Uno de los aspectos que más se toma en cuenta para el filtrado de datos es la relevancia de las páginas que se van a filtrar. Una alta relevancia para la búsqueda puede determinarse de acuerdo a diferentes criterios:

- *Filtrado Algorítmico*

Se utiliza principalmente en búsquedas en la web. Una vez que el usuario ingresa sus términos de búsqueda, el algoritmo le muestra otras opciones que podrían ser de su interés. Google utiliza este enfoque en sus búsquedas.

- *Filtrado Social*

Utilizado por varias aplicaciones y redes sociales como Facebook. Sugiere al usuario opciones que sus amigos prefieren. El factor predominante en este enfoque es el factor de afinidad entre el usuario y la fuente de información.

- *Filtrado Humano*

Muchas personas recomiendan páginas web como un hábito; este enfoque se basa en dichas recomendaciones, los usuarios ahora confían en dichas sugerencias como nuevas fuentes de información. [16]

2.5. CLASIFICACIÓN AUTOMÁTICA.

Clasificación es una técnica de minería de datos utilizada para predecir la pertenencia de una instancia de datos a un grupo. Se encarga de predecir etiquetas categóricas, a través de la construcción de un modelo. Por estas características la clasificación de páginas web se ha vuelto esencial en muchas tareas en la extracción de información en la web.

La clasificación de datos es un proceso que consta de dos pasos. En el primer paso (paso de aprendizaje) se construye un modelo a partir de un conjunto de entrenamiento, que contiene las tuplas y sus etiquetas asociadas. Dado que las etiquetas ya han sido establecidas, este paso también se conoce como aprendizaje supervisado. Si las etiquetas no se conocen se denomina

como aprendizaje no supervisado. El segundo paso consiste en la utilización del modelo para clasificar. En este paso se mide la exactitud del modelo, al probarlo con un conjunto de prueba. El conjunto de prueba está formado por tuplas que tienen ya sus etiquetas asignadas, estos datos no se utilizaron en el entrenamiento del modelo o clasificador. Una vez que el conjunto de datos fue clasificado por el modelo, se compara la etiqueta asignada por el clasificador con la etiqueta predeterminada. La exactitud del clasificador constituye el porcentaje de tuplas correctamente clasificadas. Además que la exactitud del modelo, se deben tomar en cuenta otros aspectos para evaluar un clasificador. Entre estos podemos mencionar: velocidad y escalabilidad, robustez (capacidad del modelo para manejar ruido y datos faltantes), etc.

Existen algunos enfoques para la clasificación supervisada, entre los cuales podemos mencionar:

- *Árboles de Decisión*

  Se basa en la construcción de un diagrama de flujo cuya estructura se parece a la de un árbol. Los nodos internos denotan un atributo a ser probado y las ramas representan un resultado de la prueba. Los nodos hoja representan las etiquetas de las clases. Los árboles de decisión se construyen en dos pasos. En el primer paso, todos los ejemplos se ubican en la raíz, recursivamente se va repartiendo los nodos de acuerdo a los atributos que se hayan identificado. En este momento se lleva a cabo el segundo paso que consiste en identificar y remover ramas con ruido o elementos atípicos. Finalmente se prueba el árbol con el conjunto de prueba.

- *Clasificador Bayesiano*

Estos son clasificadores estadísticos. Pueden predecir la probabilidad de que una tupla pertenezca a una clase. “El aprendizaje bayesiano calcula la probabilidad de cada hipótesis de los datos y realiza predicciones sobre estas bases. Es un aprendizaje casi óptimo, pero requiere grandes cantidades de cálculo debido a que el espacio de hipótesis es normalmente muy grande” [21]. El clasificador Bayesiano simple supone un tamaño de la muestra asintóticamente infinito e independencia estadística entre variables. Con estas condiciones, se puede calcular las distribuciones de probabilidad de cada clase para establecer la relación entre los atributos (variables independientes) y la clase (variable dependiente) [22]. A pesar de que este clasificador no considera las dependencias naturales que existen entre los diferentes atributos, su desempeño es bastante bueno y se asemeja al de un árbol de decisión o incluso con algunas redes neuronales. También tenemos otro tipo de clasificador bayesiano que considera la existencia de posibles dependencias entre las variables; las redes Bayesianas son gráficos acíclicos dirigidos cuyos nodos representan variables y los arcos que los unen codifican dependencias condicionales entre las variables. Los nodos pueden representar cualquier tipo de variable, ya sea un parámetro medible (o medido), una variable latente o una hipótesis [23].

- *Clasificación basada en reglas*

El modelo en este tipo de clasificadores está representado por una serie de reglas Si – Entonces. Dichas reglas pueden ser generadas de un árbol de decisión o directamente de los datos usando un algoritmo.

- *Clasificación por retro propagación*

En este tipo de clasificación se utilizan los algoritmos de redes neuronales. Una red neuronal consiste de una serie de unidades de entrada y salida en las cuales cada conexión tiene un peso asociado a ella. Durante la fase de aprendizaje la red aprende ajustando los pesos, de tal forma que sea capaz de predecir la etiqueta correcta dada la tupla de entrada. El entrenamiento de este tipo de redes toma mucho tiempo, sin embargo incluye gran tolerancia a ruido y tiene una gran habilidad para clasificar datos para los que no fue entrenado.

En cuanto a clasificación no supervisada, el método más utilizado es el Clustering. El proceso que consiste en agrupar un conjunto de objetos físicos o abstractos en clases de objetos similares se conoce como Clustering. El primer paso para utilizar este método es dividir el conjunto de datos en partes más pequeñas que contengan datos similares, para después asignar las etiquetas a estos grupos más pequeños. Este tipo de clustering tiene la ventaja de ser un proceso adaptable a los cambios.

En general la mayor parte de métodos de Clustering pueden clasificarse en las siguientes categorías:

- *Métodos de partición*

Dado un conjunto de datos con n objetos o tuplas, los métodos de partición construye k particiones de los datos, donde cada partición representa un cluster y k<n. Las particiones se construyen de forma iterativa, se inicia con una y las demás se generan a partir de esta, moviendo objetos de una partición a otra. Cada partición debe tener al menos un objeto y cada uno de los objetos debe pertenecer a una sola de las particiones.

- *Métodos Jerárquicos*

Estos métodos crean una descomposición jerárquica de un conjunto de datos dado. Dependiendo del tipo de descomposición pueden ser acumulativos (bottom-up) o divisivos (top-down). Para el enfoque acumulativo cada objeto inicia en un grupo diferente, que con cada paso puede unirse a otro cercano, hasta llegar a tener un solo grupo o hasta llegar a una condición que detenga el proceso. Para el enfoque divisivo por el contrario todos los objetos inician en el mismo grupo y con cada paso se van separando hasta que cada objeto termine en un grupo o hasta llegar a una condición que termine el proceso. Ambos enfoques aunque diferentes se parecen en que una vez que se llevo a cabo un paso no puede deshacerse.

- *Métodos basados en densidad*

Hacen que los clusters crezcan mientras el número de objetos en la vecindad sea mayor a cierto umbral. Estos métodos se utilizan para descubrir ruido en los datos o clusters no esféricos.

- *Métodos basados en grillas*

Se genera una estructura con un determinado número de celdas, que permite formar los clusters. La ventaja de estos métodos radica en que son muy veloces, el tiempo de procesamiento no depende del número de objetos, sino del número de celdas.

- *Métodos basados en modelos*

Se formula un modelo hipotético para cada cluster y se busca el mejor ajuste de dicho modelo.

- *Métodos basados en restricciones*

En estos métodos se incorpora restricciones (expectativas de los usuarios o propiedades) al resultado de los clusters. [17]

2.5.1. COEFICIENTE TFIDF

El coeficiente TFIDF por sus siglas en inglés (Term Fre-quency – Inverse Document Frequency) es una medida estadística propuesta por Gerard Salton, usada para evaluar la importancia de una palabra para el ámbito de cada uno de los documentos en particular y para el de la colección en su conjunto. Su análisis puede dividirse en dos partes esenciales TF e IDF.

La importancia de cada palabra TF (Term Frequency) incrementa proporcionalmente de acuerdo al número de veces que la misma aparece en el documento, sin embargo debemos tomar en cuenta el tamaño del documento, ya que esto afectará al peso final de cada palabra. Para evitar este problema, se normaliza el valor tomando en cuenta también la cantidad de documentos en los que aparece dicha palabra, para tratar de nivelar todas las palabras que aparecerán en la colección. [18]

De esta forma la importancia del término $\mathbf{t_i}$ en el documento $\mathbf{d_j}$ está dado por:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Donde $n_{i,j}$ es el número de ocurrencias del término considerado en el documento d_j. El denominador constituye el número de ocurrencias de todos los términos en el documento d_j.

La frecuencia inversa de los documentos (IDF) es una medida de la importancia general del término y se calcula mediante:

$$IDF_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

Donde el numerador lo constituye el número total de documentos en el cuerpo y el denominador es el número de documentos donde el término t_i aparece $(i.e., n_{i,j} \neq 0)$

Así el coeficiente TFIDF para el término t_i en el documento d_j es:

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i$$

Un valor TFIDF alto es alcanzado por un término con alta frecuencia en el documento considerado, pero baja frecuencia en la colección de documentos. De esta manera, el coeficiente tiende a filtrar términos comunes. [19]

2.5.2. EMBOLSAMIENTO

El embolsamiento es un concepto de la minería de datos, que consiste en la combinación de la predicción de múltiples clasificadores o predictores individuales (mediante un simple mecanismo de votación), implementados muchas veces por técnicas de

modelamiento distintas, pero entrenados bajo un mismo conjunto de datos, para producir un único clasificador.

Muchos investigadores (Breiman, Clemen, Wolpert) han mostrado que el embolsamiento es principalmente efectivo en algoritmos de aprendizaje “inestables” como redes neuronales y árboles de decisión, donde pequeños cambios en el conjunto de entrenamiento producen grandes variaciones en su predicción. Un clasificador basado en embolsamiento tiene normalmente una mayor exactitud que cualquier técnica de clasificación individual, porque ayuda a contrarrestar las deficiencias de cada uno de los clasificadores individuales, además de reducir la inestabilidad inherente de los resultados cuando se tienen modelos complejos aplicados a conjuntos de datos pequeños. [20]

También debemos considerar que el embolsamiento ayuda a incrementar la robustez del clasificador ante el ingreso de datos ruidosos, ya que el modelo compuesto reduce la varianza de los clasificadores individuales. [19]

CAPÍTULO 3

DISEÑO DEL MÓDULO DE CLASIFICACIÓN

3.1. DEFINICIÓN DE REQUERIMIENTOS

3.1.1. INTRODUCCIÓN

El sistema PSearch para la personalización de búsquedas, necesita mejorar la precisión de los resultados que presenta al usuario. De esta manera, se necesita un algoritmo de clasificación que ayude al mejoramiento del mencionado sistema. Este proyecto busca la elaboración de un módulo de clasificación que cumpla con los requerimientos del sistema PSearch.

3.1.2. REQUERIMIENTOS ESPECÍFICOS

Una vez que se analizó el sistema Psearch, tanto su código como su arquitectura, se determinó los puntos en los cuales debería mejorarse el algoritmo de clasificación que utiliza. Entonces se definieron los siguientes requerimientos para ser implementados en el módulo de clasificación:

R-1.- El sistema de búsquedas personalizadas Psearch requiere un módulo de clasificación temática de páginas web.
R-1.1.- El módulo debe permitir la clasificación por categorías.
R-1.2.- La clasificación debe ser excluyente, no se permite que una página pertenezca a más de una categoría.
R-1.3.- Cada categoría será descrita por un conjunto pequeño de términos.
R-2.- El módulo de clasificación debe ser simple.
R-2.1.- La clasificación debe realizarse utilizando información contenida en las páginas.

R-2.2.- El módulo debe incluir pre procesamiento de los datos de las páginas que se van a clasificar de tal forma que la clasificación sea más sencilla.
R-3.- El módulo de clasificación debe ser eficiente.
R-3.1.- Debe usar la menor cantidad de recursos computacionales.
R-4.- El módulo debe buscar exactitud.
R-4.1.- El módulo debe incluir más de un algoritmo.
R-5.- El módulo debe tener una implementación modular.
R-5.1.- Se definen tres etapas claramente definidas: preprocesamiento, clasificación y entrenamiento.
R-6.- Las páginas clasificadas deben ser finalmente almacenadas para su posterior uso.
R-6.1.- Se requiere el uso y diseño de un sistema de base de datos para almacenar la información de las páginas que se van a clasificar, para que puedan utilizarse en búsquedas posteriores, haciendo más rápido al sistema.

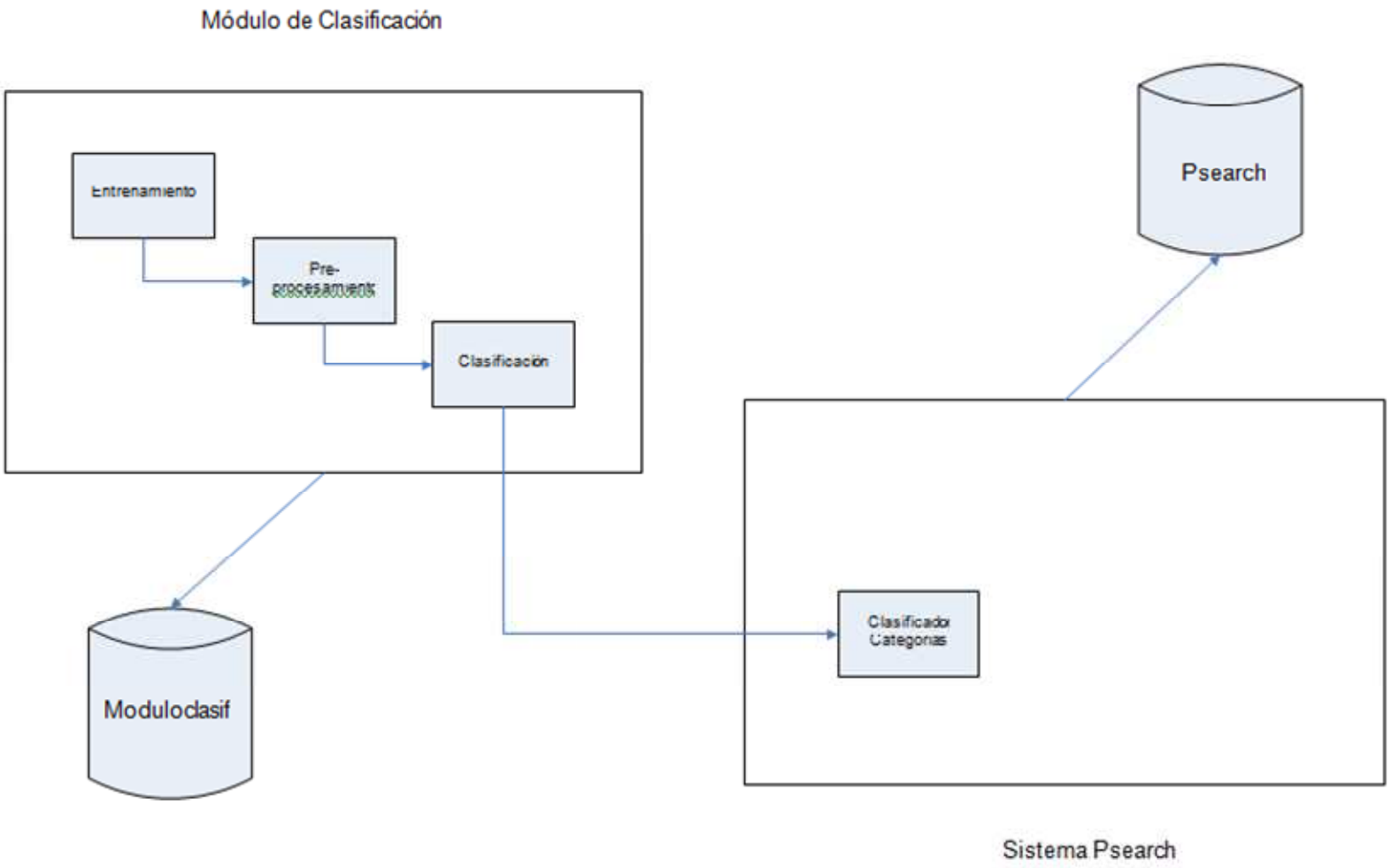
3.2. ARQUITECTURA DEL MÓDULO DE CLASIFICACIÓN

Una vez que se han definido los requerimientos para el módulo de clasificación, también se definió la arquitectura del mismo:

El módulo de clasificación contará con tres componentes principales: Entrenamiento, Pre-procesamiento y Clasificación. En el diagrama de arquitectura se puede observar estos componentes y la comunicación que existe entre ellos.
El módulo de clasificación finalmente entregará un resultado que el sistema PSearch pueda utilizar en sus cálculos finales para entregar al usuario resultados mejores.

La mencionada arquitectura puede observarse en el diagrama 1.- Arquitectura del módulo de clasificación.

Diagrama 1.- Arquitectura del Módulo de Clasificación



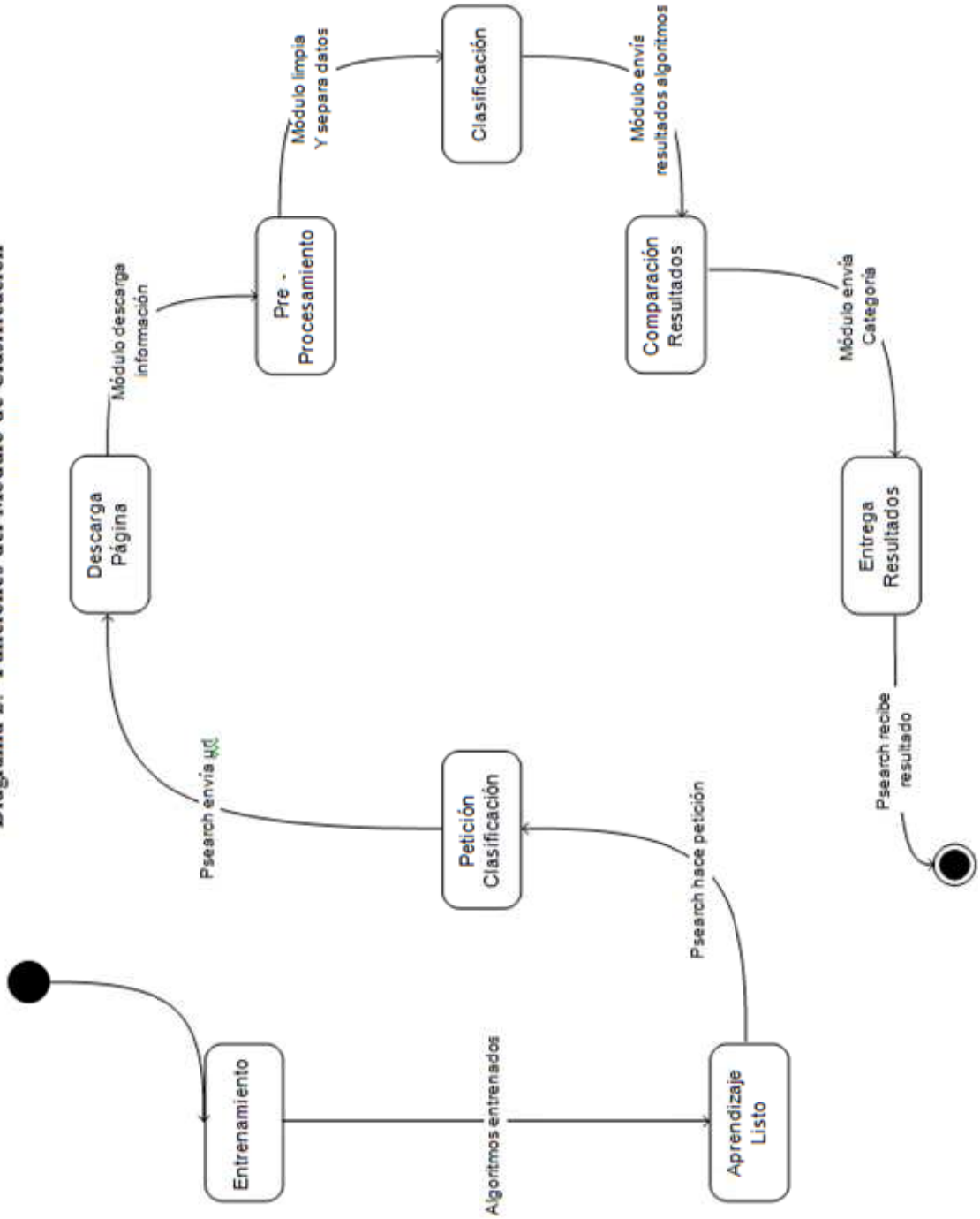
3.3. Funciones y Actores del Módulo

Las funciones del módulo de clasificación y sus actores se describen de manera detallada en el diagrama de transición de estados y en los casos de uso de las siguientes secciones.

3.3.1. Diagrama de Transición de Estados

Las transiciones se producen como consecuencia de eventos y pueden tener un procesamiento asociado. El módulo de clasificación, inicia en el estado entrenamiento, donde los algoritmos que forman parte del mismo reciben los datos necesarios para poder realizar una clasificación exitosa cuando sea necesario. Una vez que el entrenamiento finaliza, el módulo pasa al estado aprendizaje listo, en donde espera por una petición del sistema PSearch. Una vez que llega dicha petición, el módulo cambia al estado Petición de clasificación, en donde inicia el proceso más importante para el módulo. Se recibe el url de la página que se desea clasificar, en este momento el módulo pasa al estado Descarga página, en este estado el módulo busca la página web y descarga el título, metadatos y body. Al finalizar la descarga, el módulo entra al estado Pre-procesamiento, donde se limpia los datos de la página, que son enviados para su clasificación. En el estado clasificación, los algoritmos que conforman el módulo son ejecutados y finalmente envían sus resultados para que sean comparados, eso lleva al módulo al estado Comparación resultados, donde finalmente se llega al estado final Entrega de resultados en el que el sistema PSearch recibe el resultado final.

Diagrama 2.- Funciones del Módulo de Clasificación



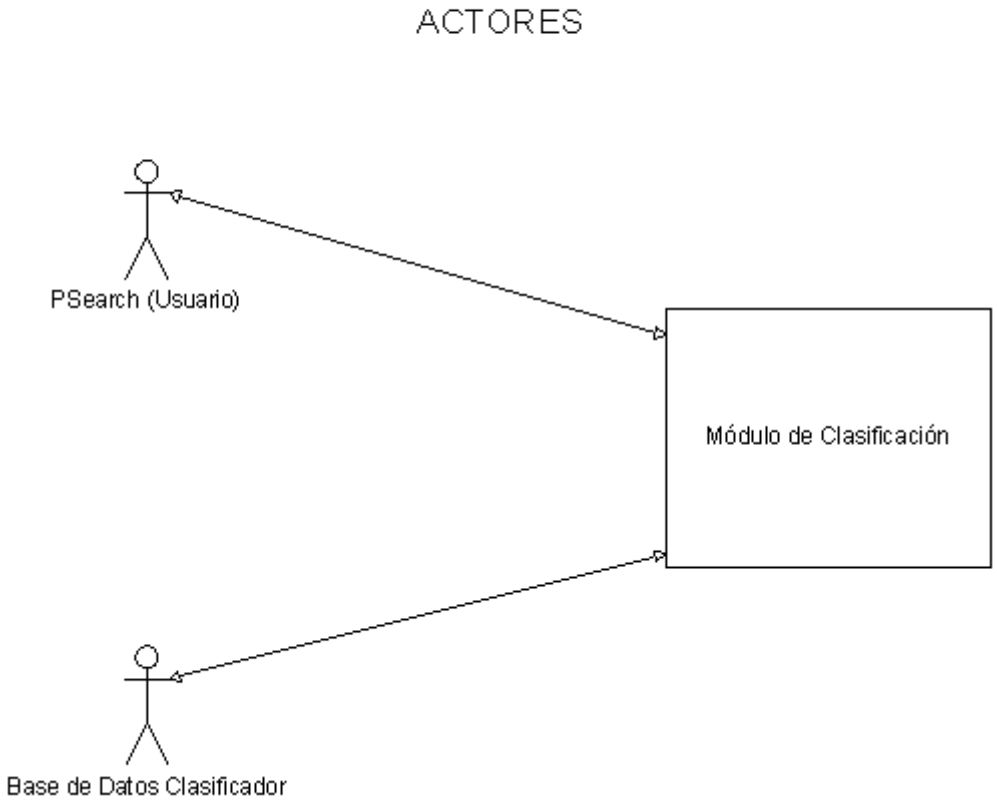
3.3.2. ACTORES

El módulo de clasificación se relaciona con dos entidades externas principalmente. A continuación se describe a dichas entidades:

ACTOR	PSearch App (Usuario)
CASOS DE USO	Pre-procesar Clasificar
TIPO	Primario
DESCRIPCIÓN	Es el actor principal que utiliza el módulo de clasificación. Realiza los pedidos de clasificación de páginas web.

ACTOR	Base de datos Clasificador
CASOS DE USO	Clasificar
TIPO	Secundario
DESCRIPCIÓN	Base de datos donde se almacena toda la información sobre la clasificación de las páginas web.

Diagrama 3.- Actores del Módulo de Clasificación



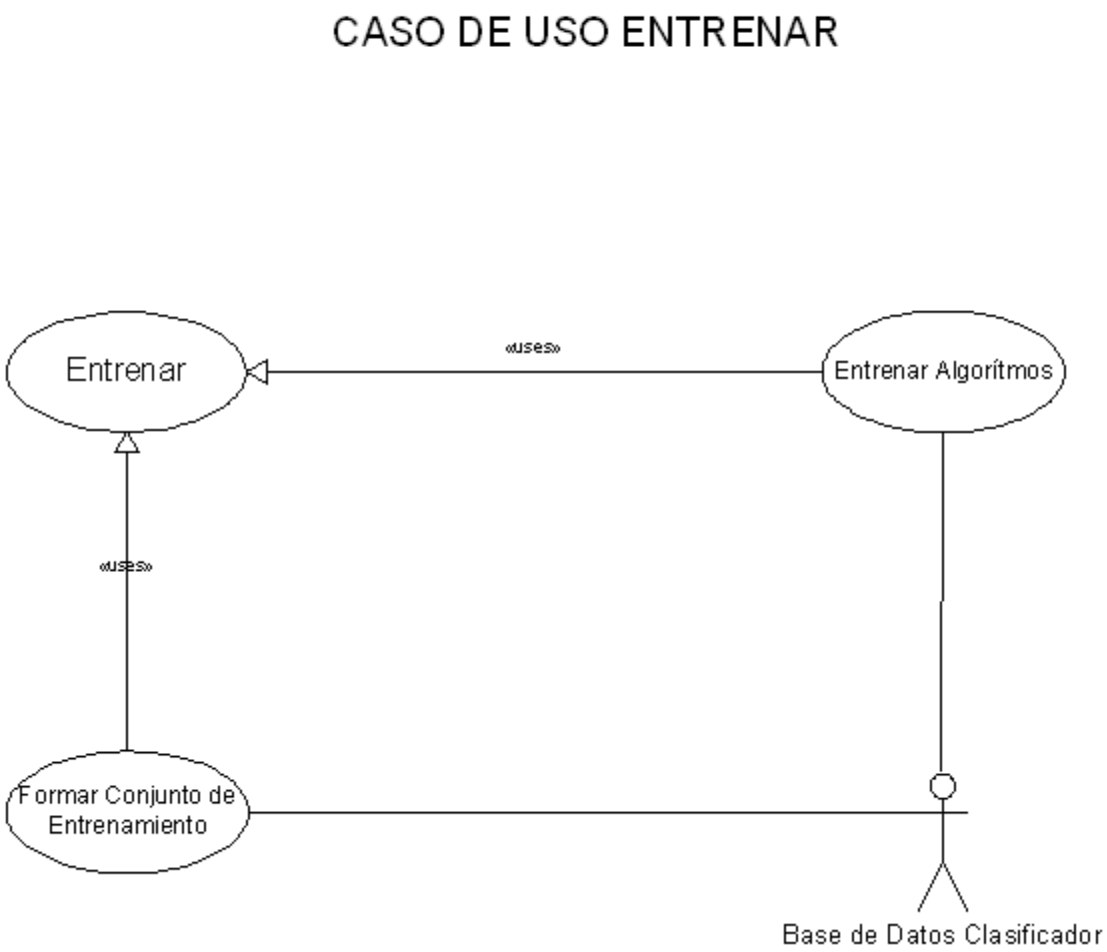
3.3.3. CASOS DE USO

Una vez que conocemos las funciones del módulo, así como los actores que se relacionan con el mismo, debemos definir la manera en que dichos actores intervienen en las funciones del módulo. Estas interacciones las exploramos a través de los casos de uso que se muestran a continuación:

- Caso de Uso Entrenar

CASO DE USO	Entrenar
ACTORES	Base de datos Clasificador
TIPO	Básico
PROPÓSITO	Preparar los algoritmos de clasificación del módulo y determinar los vectores que se utilizarán en clasificaciones posteriores.
RESUMEN	Este caso de uso permite el entrenamiento de los diferentes algoritmos que constituyen el módulo de clasificación, además del almacenamiento de una serie de términos que por su relevancia se utilizaran para clasificar posteriores páginas web.
PRECONDICIONES	Ninguna
FLUJO PRINCIPAL	El conjunto de páginas de entrenamiento pasa por los algoritmos, se determina los términos necesarios y se almacena la información en la base de datos.
SUB FLUJOS	Ninguno
EXCEPCIONES	No exista datos de entrenamiento.

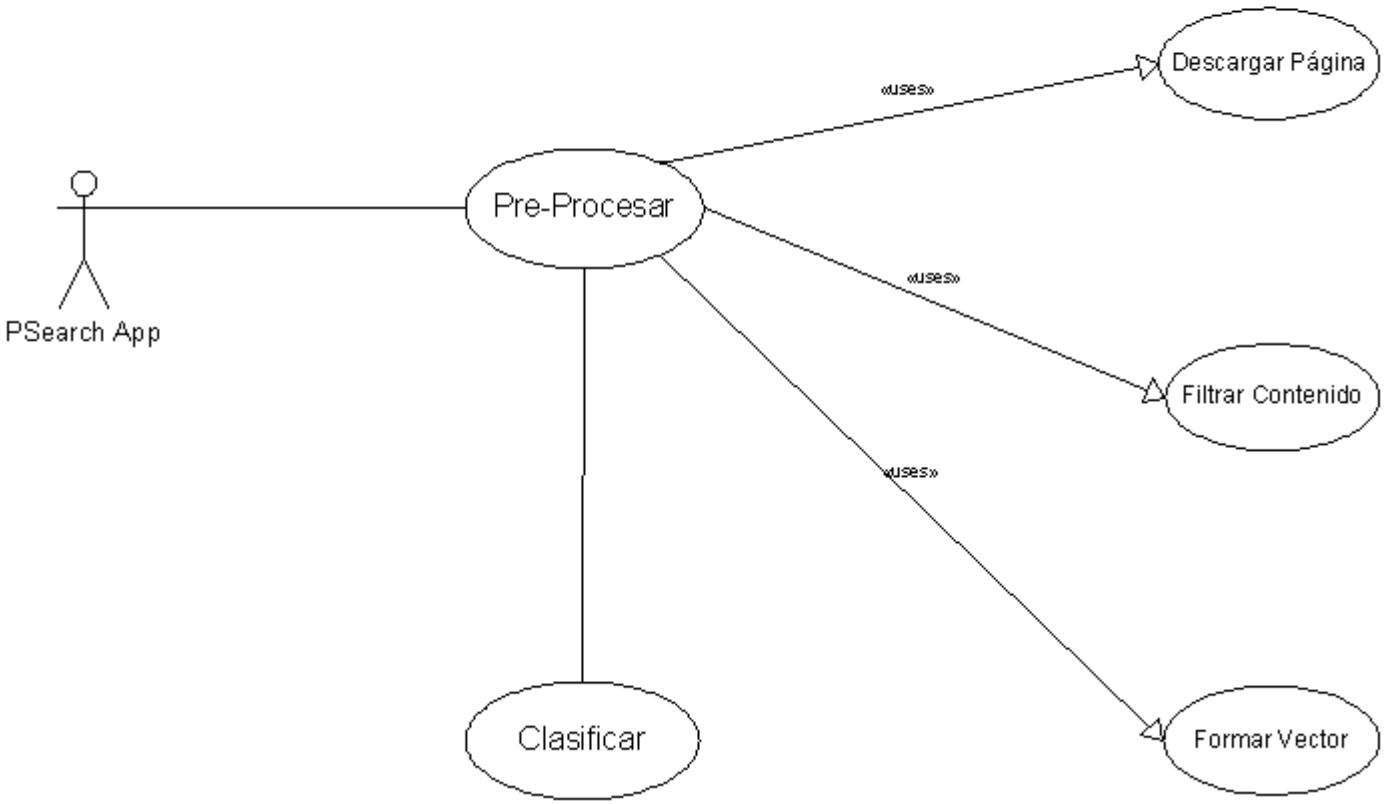
Diagrama 4.- Caso de Uso Entrenar



- Caso de Uso Pre – procesar

CASO DE USO	Pre – Procesar
ACTORES	PSearch App
TIPO	Básico
PROPÓSITO	Eliminar la información innecesaria del contenido de las páginas web para facilitar su clasificación.
RESUMEN	Este caso de uso permite remover todos los tags html y demás información poco relevante para la clasificación, una vez que la información ha sido procesada, se forma un vector que es enviado al caso de uso clasificar.
PRECONDICIONES	La aplicación PSearch debe enviar un pedido de clasificación, así como el url de la página que se desea clasificar.
FLUJO PRINCIPAL	La aplicación PSearch envía el pedido de clasificación, entonces se descarga el contenido de la página y se procede a remover los tags e información innecesaria.
SUB FLUJOS	Ninguno
EXCEPCIONES	No se encuentre la página solicitada.

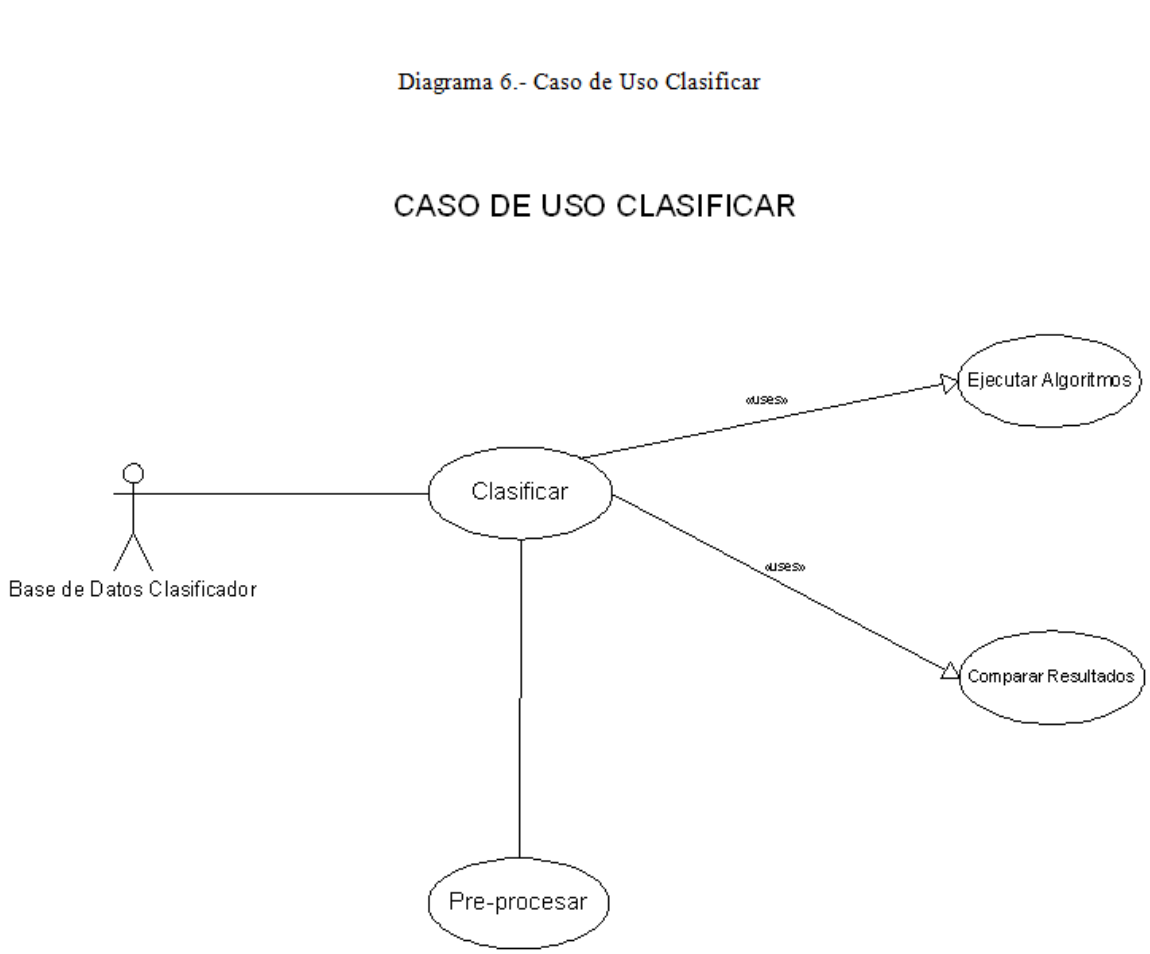
Diagrama 5.- Caso de Uso Pre-Procesar



- Caso de Uso Clasificar

CASO DE USO	Clasificar
ACTORES	PSearch App, Base de datos Clasificador
TIPO	Básico
PROPÓSITO	Determinar la categoría a la que pertenece una página web, enviada por la aplicación PSearch.
RESUMEN	Este caso de uso permite clasificar la página enviada por PSearch. El vector enviado pasa por los algoritmos, se comparan los resultados y se determina la categoría final de la página.
PRECONDICIONES	Haber ejecutado el caso de uso Pre-procesar
FLUJO PRINCIPAL	Una vez ejecutado el caso de uso pre-procesar se ejecutan los diferentes algoritmos, cada uno determina la categoría a la que pertenece la página, estos resultados se comparan, aquella categoría que tenga mayor número de ocurrencias es la que se devuelve como resultado a la aplicación PSearch.
SUB FLUJOS	Ninguno
EXCEPCIONES	No se encuentre la página solicitada.

Diagrama 6.- Caso de Uso Clasificar



3.4. IMPLEMENTACIÓN

3.4.1. INTEGRACIÓN CON EL SISTEMA PSEARCHENGINE

El sistema PSearch, es un sistema de búsquedas web, que mediante el Google Web Search Api devuelve al usuario la información más cercana a sus gustos y preferencias. PSearch trabaja mediante el uso de perfiles, en los que los usuarios escogen cuales son las categorías que les interesan. Es a través de esta información que se puede personalizar los resultados devueltos por este sistema. Es en esta área donde se integra el módulo de clasificación, que será llamado cada vez que el usuario haga una búsqueda que deba filtrarse en base a sus preferencias.

Para el presente trabajo, se desarrolló un prototipo inicial basado en la plataforma Java. Se utilizó el JDK 1.6 ejecutando sobre una máquina Turion 64 x2 con 2MB de memoria RAM y con el sistema operativo Windows Vista.

CAPÍTULO 4

CLASIFICACIÓN AUTOMÁTICA Y TEMÁTICA DE PÁGINAS WEB MEDIANTE LA UTILIZACIÓN DE UN ALGORITMO SIMPLE Y EFICIENTE

4.1. DESCRIPCIÓN GENERAL

La clasificación de páginas Web es esencial en la actualidad; se utiliza en diferentes procesos de administración y recuperación de información, sin embargo su principal uso se caracteriza por el mejoramiento de la calidad de búsquedas mediante el filtrado de contenido. PSearch es un sistema de búsqueda, que trata de brindar al usuario la información más relevante. En este momento, los resultados entregados por el sistema, no son los mejores, es por ello que el módulo de clasificación propuesto, trata de que Psearch alcance el objetivo de entregar al usuario resultados más certeros, que satisfagan su necesidad de información.

El módulo de clasificación trata de mejorar la aplicación existente, es por ello que se debe tener en cuenta dos factores muy importantes como lo son la exactitud y el desempeño. Existen clasificadores muy exactos, pero requieren algoritmos complejos y costosos, reduciendo su desempeño. Por otro lado, clasificadores rápidos no brindan los resultados deseados.

El módulo se basa en la frecuencia de términos, principalmente. El uso de este atributo permite reducir la complejidad del clasificador y alcanzar un buen desempeño. Además de

alto desempeño, el módulo busca la mayor exactitud posible, por ello emplea varios algoritmos como redes neuronales y clasificadores Bayesianos combinados mediante la técnica de embolsamiento. El módulo de clasificación consta de tres etapas claramente definidas: el pre-procesamiento de las páginas, su clasificación y la etapa de entrenamiento de los algoritmos.

- *Pre – procesamiento*

El módulo de clasificación, inicia su trabajo una vez que ha recibido la petición de clasificación del sistema PSearch. Se envía el URL de la página que se desea clasificar. Se verifica entonces si el URL que se envió está completo y si es necesario se añade los términos http:// o un "/" al final de tal forma que pueda conectarse sin problemas a la página deseada y poder extraer el contenido de la misma para su clasificación. Debemos recordar que ya que se trata de una aplicación clasificadora de páginas web, debe establecerse con anterioridad una conexión a internet, de lo contrario se lanza una excepción y no se lleva a cabo el proceso de clasificación. Una vez que se ha validado el URL, se procede a obtener el contenido de la página. La aplicación entonces establece la conexión con la página y extrae el contenido. Para la extracción del contenido utiliza la clase de java URL, que ofrece métodos específicos para este propósito. Una vez que la aplicación se conecta a la página, extrae el contenido de la misma el cual es guardado en un buffer para su posterior procesamiento. Cuando contamos con estos datos en el buffer, se los transforma a texto. Este texto es entonces dividido en tres partes: Título, meta datos y cuerpo. Estas tres secciones, son almacenadas para su posterior análisis. La división está justificada por la alta relevancia que tienen elementos como el título y los meta datos en la definición del tema principal de una página web.

Las tres secciones (título, meta datos y cuerpo) son entonces filtradas de manera independiente para eliminar todas sus etiquetas HTML. Cada sección entonces se convierte en un grupo de palabras sin ningún tipo de formato.

El siguiente paso consiste en la identificación del lenguaje de la página web que se desea clasificar, cada lenguaje tiene sus particularidades que deben ser manejadas de diferentes maneras. El módulo de clasificación puede manejar páginas web tanto en inglés como en español. Cada una de las secciones almacenadas, pasa por un algoritmo para retirar las palabras comunes, que no aportan ninguna información sobre el contenido de la página como adverbios, preposiciones, artículos, pronombres, entre otros. Posteriormente, cada conjunto de palabras es procesado mediante el algoritmo de Paice/Husk [31] para encontrar los lexemas o raíces comunes de cada palabra. El uso de lexemas en lugar de sus variantes morfológicas tiene la ventaja de incrementar la tasa de asociación al agrupar términos con igual raíz a pesar de sus terminaciones diferentes. Estas versiones procesadas del contenido del título, meta datos y cuerpo, son almacenadas. Entonces se pasa al cálculo del coeficiente TFIDF para cada sección. En base a los valores más altos resultantes de este cálculo se forma el conjunto de términos relevantes para la página. Este conjunto de términos forma un vector, que es enviado a los diferentes algoritmos para su posterior clasificación.

- *Clasificación*

La fase de clasificación inicia cuando se envía el vector obtenido en la fase de pre-procesamiento a cada uno de los algoritmos que forman parte del módulo. En ese momento cada algoritmo previamente entrenado clasificará a la página y emitirá un resultado. Posterior a la clasificación de cada una de los algoritmos, se debe determinar la categoría definitiva a

la que pertenece la página. Para ello se hace una comparación entre los resultados obtenidos. Se define la clase por mayoría simple, sin embargo si todos son diferentes, se definió uno de los algoritmos para que cuente con el voto de confianza y determine la categoría a la que pertenece la página. Esta categoría es almacenada y enviada finalmente al sistema PSearch.

- *Entrenamiento*

Esta es una tarea previa, que se realiza una sola vez, fuera de línea y consiste en el entrenamiento individual de cada uno de los algoritmos que forman parte del clasificador. Es una tarea que lleva tiempo y depende directamente del número de páginas Web de entrenamiento seleccionadas. De cualquier forma, el proceso de entrenamiento trata de obtener la información necesaria para que cada uno de los algoritmos del clasificador pueda cumplir con su tarea. Se determina el conjunto de lexemas más relevantes a cada categoría. Para ello, cada página del conjunto de entrenamiento es pre –procesada hasta convertirla en un conjunto de lexemas. Una vez que todas las páginas se han convertido en conjuntos de lexemas, se procede a calcular los coeficientes TFIDF de cada lexema existente en los documentos de entrenamiento. Para finalizar esta primera parte, se ordenan los valores TFIDF dentro de cada categoría y se escogen los lexemas correspondientes a los valores TFIDF más altos. El número de lexemas escogidos para cada categoría es un parámetro variable que sin duda generará resultados diferentes durante el proceso de clasificación.

Estos términos definen a cada categoría y son los que nos servirán para generar los vectores que sirven para entrenar a los algoritmos del clasificador. Al terminar la fase de entrenamiento, el módulo queda listo para recibir las páginas y clasificarlas.

4.2. ALGORITMOS UTILIZADOS

El módulo de clasificación está compuesto por tres algoritmos agrupados mediante embolsamiento. A continuación se describe cada uno de ellos.

4.2.1. CLASIFICADOR POR FRECUENCIA DE TÉRMINOS

Es un clasificador basado en el análisis de términos del contenido de la página web. Este algoritmo contiene por categoría una serie de términos que si son encontrados, significa que dicha página pertenece a esa categoría. El algoritmo es exclusivo, es decir que una página solamente puede pertenecer a una sola categoría. Por ello los términos que se definieron tratan de diferenciar claramente las categorías que maneja la aplicación. El algoritmo no requiere entrenamiento previo, sin embargo para su desarrollo se realizó un análisis detenido de las páginas de entrenamiento para otros algoritmos, de tal manera que se pueda determinar los términos comunes que pertenecían a cierta categoría.

El algoritmo empieza su ejecución al recibir el vector que contiene los lexemas con coeficiente TFIDF más altos. Entonces se analiza el URL, título, meta datos y cuerpo, de tal forma que se compara, los lexemas que se recibieron como entrada, con los que se definieron para cada categoría. De esta comparación, se determina la categoría a la que pertenece, que será aquella a la que mayor cantidad de lexemas coincidan.

4.2.2. CLASIFICADOR BAYESIANO

Los clasificadores Bayesianos, son clasificadores estadísticos que tratan de predecir las probabilidades de pertenencia a una clase. Para el módulo de clasificación, se seleccionó

un clasificador Bayesiano simple o “inocente”, para predecir si una tupla pertenece a una clase en particular. El clasificador bayesiano simple o “inocente” implementado asume que el efecto del valor de un atributo en una clase dada es independiente de los valores de los demás atributos. Este supuesto se conoce como “Independencia condicional de Clase”, el cual simplifica los cálculos que deben hacerse (Es de este supuesto de donde se deriva el nombre de “inocente”). Sin embargo a pesar de su simplicidad, este algoritmo puede superar el desempeño de métodos de clasificación mucho más sofisticados. [24]. Estudios realizados para comparar diferentes algoritmos de clasificación, encontraron que el clasificador Bayesiano conocido como ‘’inocente” es comparable en desempeño con árboles de decisión y algunas redes neuronales. Además se ha mostrado gran exactitud y excelente desempeño al trabajar con grandes bases de datos. [25] [26]

En la práctica, este algoritmo ha tenido un notable éxito, motivo por el cual forma parte de esta implementación. El clasificador bayesiano funciona de la siguiente manera:

Sea D, un conjunto de tuplas y sus clases asociadas. Cada tupla está representada por un vector de atributos X=$(x_1, x_2, \ldots . x_{14})$.

Existen 13 clases, $C_1, C_2, \ldots . C_{14}$. Dada una tupla, X,el clasificador predecirá si X pertenece a la clase cuya probabilidad posterior sea la más alta. Esto es, el clasificador Bayesiano predecirá que la tupla X pertenece a la clase C_i solamente si

$$P(C_i|X) > P(C_j|X) \text{ para } 1 \leq j \leq m, j \neq i$$

De tal forma maximizamos $P(C_i|X)$. La clase C_i para la cual $P(C_i|X)$ es maximizada se conoce como la “hipótesis de maximización posterior”. Por el teorema de Bayes tenemos:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Dado que $P(X)$ es constante para todas las clases, solamente necesitamos maximizar $P(X|C_i)P(C_i)$. Si no se conoce suficiente información de las probabilidades de las clases, comúnmente se asume que las clases son igualmente probables, $P(C_1) = P(C_2) = \cdots = P(C_{14})$. Dados los conjuntos de datos con muchos atributos, sería demasiado costoso computacionalmente calcular $P(X|C_i)$. Para reducir este tiempo, hacemos la asunciónde independencia condicional entre las clases. Esto significa que no existen relaciones de dependencia entre los atributos. De tal forma,

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i)$$

$$= P(X_1|C_i) \cdot P(X_2|C_i) \cdot .... \cdot P(X_n|C_i)$$

Estimar entonces las probabilidades $P(X_1|C_i)$, $P(X_2|C_i)$,…, $P(X_n|C_i)$ de las tuplas de entrenamiento se vuelve un proceso muy sencillo. En este caso $X_k$hace referencia a los valores de los atributos en cada tupla.Para cada atributo, queremos saber, si el atributo es categórico o continuo, de tal forma que se pueda realizar correctamente el cálculo de $P(X|C_i)$. En el caso de los datos para el clasificador, los valores de los atributos son categóricos. De donde $P(X_k|C_i)$ es el número de tuplas en la clase C_i en D, teniendo el valor de X_k para cada atributo, dividido para $|C_{i,D}|$, el número de tuplas en la clase C_i en D. [27]

4.2.3. RED NEURONAL

La red neuronal usa un esquema de recordación sin re-alimentación que fue entrenado mediante el algoritmo de retro-propagación del error. Dicho algoritmo es un método común de aprendizaje supervisado basado en el método del gradiente descendiente. [28]

La red neuronal utilizada en el módulo de clasificación se caracteriza por tener una arquitectura en niveles y conexiones estrictamente hacia adelante entre las neuronas. Se encarga del aprendizaje de un conjunto predefinido de pares de entradas-salidas. [29]

El algoritmo requiere entradas y salidas que sean únicamente 0 o 1, por lo que se codificó los términos clave que determinan la clase a la que pertenece la página. De tal forma, antes de llamar al algoritmo se ejecuta el método Analizar de la clase BaggedClassifier, este método se encarga de analizar el URL, titulo, meta datos y cuerpo dependiendo de los parámetros que reciba. Como resultado este método almacena en la página que recibe como parámetro, un código de 0 y 1 que identifica los términos encontrados en el contenido de la página. Una vez que se cuenta con estos datos y la salida, que de igual manera es un código de 4 bits, que es conocida para el entrenamiento, se procede a ejecutar el algoritmo. Comienza alimentando los valores de la entrada de acuerdo a las siguientes ecuaciones:

$$r_i = \sum_{j \in A} O_j W_{ji} \quad \forall\, i : i \in B$$

Donde A es el grupo de neuronas en una capa, y B constituye la siguiente capa. O_j es la activación para la neurona J, y W_{ij} son los pesos asignados a la conexión entre las neuronas j e i. En la ecuación anterior, se toman los valores de salida y se alimenta a la siguiente capa a través de los pesos. Esta operación se realiza para cada neurona en la

siguiente capa, produciendo un valor de red. Este valor es la suma de todos los valores de activación en las neuronas de la capa anterior, y cada valor de red es aplicado ahora a la siguiente ecuación, conocida como función de activación:

$$O_i = f(r_i) = \frac{1}{1 + e^{-r_i}}$$

Cuando todas las neuronas tienen un valor de activación asociado a un patrón de valores de entrada, el algoritmo sigue buscando errores en cada neurona que no es entrada. Los errores encontrados para las neuronas de salida, son propagados hacia atrás, a la capa anterior para que puedan ser asignados a neuronas de las capas escondidas. Se obtiene de la siguiente manera:

$$\delta_i = f'(r_i) \sum_{j \in \varepsilon} \delta_j W_{ij} \ \forall i : i \in D$$

Donde D es el grupo de neuronas en una capa que no es de entrada y ε es el grupo de neuronas de la siguiente capa. Este cálculo se repite para cada capa escondida en la red. Después de que se ha encontrado la activación y el error asociado a cada grupo de neuronas, los pesos se actualizan, primero encontrando el valor de cada peso de la siguiente manera:

$$\Delta W_{ij} = C O_i \delta_j \ \forall i,j : i \in A, j \in B$$

Donde C, conocida como la razón de aprendizaje, es una constante que controla el valor del cambio de los pesos y W_{ij} es el cambio de los pesos entre la neurona i y j. Finalmente y de ser necesario, el peso es cambiado evaluando:

$$W_{ij_{t+1}} = W_{ijt} + \Delta W_{ij}$$

Como podemos observar este algoritmo puede auto adaptar los pesos de las neuronas de las capas intermedias para aprender la relación que existe entre un conjunto de patrones dados como ejemplo y sus salidas correspondientes. Ya entrenada la red se podrá aplicar esa misma relación, a los nuevos vectores de las páginas con ruido o incompletos, dando una salida activa si la nueva entrada es parecida a las presentadas durante el aprendizaje [30].

4.3. CATEGORÍAS DEL CLASIFICADOR

4.3.1. DEFINICIÓN DE CATEGORÍAS

Las categorías que utiliza el módulo de clasificación, fueron definidas, en base a los dos más grandes directorios existentes en la actualidad Yahoo y Google. Además se tomó en cuenta el idioma de las categorías, ya que existían algunas diferencias dependiendo de si el directorio se navegaba en inglés o en español. El directorio de Yahoo cuenta con 14 categorías, sin embargo solo cuenta con opciones para páginas en inglés. En la tabla 1, encontramos las categorías:

Categorías Directorio Yahoo	
Arts & Humanities	News & Media
Business & Economy	Recreation & Sports
Computers & Internet	Reference
Education	Regional
Entertainment	Science
Government	Social Science
Health	Society & Culture

Tabla 1.- Categorías Directorio Yahoo

El directorio de Google, cuenta con opciones tanto en inglés como en español. Cuenta con 15 categorías en ambos idiomas. En la tabla 2, se muestran las categorías para las dos opciones de idiomas:

Categorías Directorio Google (Inglés)	
Arts	Recreation
Business	Reference
Computers	Regional
Games	Science
Health	Shopping
World	Society
Home	Sports
News	

Categorías Directorio Google (Español)	
Artes	Medios de Comunicación
Ciencia y Tecnología	Negocios
Compras	Referencia
Computadoras	Regional
Deportes	Salud
Educación	Sociedad
Hogar	Tiempo Libre
Juegos	

Tabla 2.- Categorías Directorio Google

Se realizó una comparación entre las categorías para llegar a la definición final. La mayor parte de las categorías coinciden en ambos directorios, sin embargo si existen algunas diferencias.

Categorías Directorio Yahoo	Categorías Directorio Google Inglés	Categorías Directorio Google Español
Arts & Humanities	Arts	Artes
Business & Economy	Business	Negocios
Computers & Internet	Computers	Computadoras
Education		Educación
Entertainment		
Government		
Health	Health	Salud
News & Media	News	Medios de Comunicación
Recreation & Sports	Recreation	Deportes
Reference	Reference	Referencia
Regional	Regional	Regional
Science	Science	Ciencia y Tecnología
Social Science		
Society & Culture		Sociedad
		Compras
	Games	Juegos
	Home	Hogar
		Tiempo Libre
	World	

Tabla 3.- Diferencias entre los Directorios

Como podemos observar en la tabla 3, las diferencias las encontramos en las categorías de Educación, Entretenimiento, Gobierno, Ciencias Sociales, Sociedad, Compras, Juegos, Hogar Tiempo Libre y Mundo.

Para cada una de estas categorías se consultó los respectivos directorios con la finalidad entender que tipo de páginas se iba a agrupar en las categorías. De igual manera se analizó las categorías en las que coinciden los directorios.

De tal forma se hizo las siguientes consideraciones para llegar a la lista final de categorías. La categoría Juegos se va a tomar en cuenta en la categoría entretenimiento. La categoría Tiempo libre, va a tomarse en cuenta en la categoría Recreación. Las categorías compras, hogar y mundo, se incorporan en la categoría Sociedad y Cultura. La categoría Gobierno, será considerada en la categoría Regional.

Finalmente se definió la lista de trece categorías que serán utilizadas por el clasificador. Esta lista se muestra en la tabla 4:

Categorías Módulo Inglés	**Categorías Módulo Español**
Arts & Humanities	Artes y Humanidades
Business & Economy	Negocios y Economía
Computers & Internet	Computadoras e Internet
Education	Educación
Entertainment	Entretenimiento
Health	Salud
News & Media	Noticias y Medios
Recreation & Sports	Recreación y Deportes
Reference	Referencia
Regional	Regional
Science	Ciencia
Social Science	Ciencias Sociales
Society & Culture	Sociedad y Cultura

Tabla 4.- Definición de Categorías

A continuación también se define que temas se van a incluir en cada una de las categorías del clasificador, tanto en inglés como en español:

- ***Artes y Humanidades (Arts & Humanities)***.- Abarca animación, arquitectura, artes escénicas, artes gráficas, artes plásticas, artesanía, artistas, cine, cómic, danza, digitales, diseño, fotografía, galerías, graffiti.

- ***Ciencia (Science)***.- Abarca agricultura y ganadería, astronomía, biología, ciencia alternativa, ciencias de la tierra, equipos e instrumentos, física, instituciones, matemáticas, medio ambiente, preguntas a expertos, química, software, tecnología.

- ***Ciencias Sociales (Social Science)***.- Abarca lenguajes, antropología, arqueología, conferencias, estudios étnicos, estudios futurísticos, estudios de género, genealogía, geografía, historia, migración, ciencias políticas, estudios de la mujer.

- ***Computadoras e Internet (Computers & Internet)***.- Abarca ciencias de la computación, código abierto, emuladores, empresas, fuentes, gráficos, hacking, hardware, informática móvil, inteligencia artificial, internet, programación, redes, robótica, seguridad, sistemas, sistemas operativos, software, transmisión de datos, videojuegos.

- ***Educación (Education)***.- Abarca universidades, educación a distancia, formación, recursos, competencias académicas, orientación vocacional, conferencias, asistencia financiera, graduación, programas de estudios, reformas, enseñanza, estándares y pruebas, teoría y métodos.

- ***Entretenimiento (Entertainment)***.- Abarca música, actores, películas, humor, programas de televisión, magia, premios, eventos, juegos, radio, reseñas, superhéroes, trivia, villanos.

- ***Negocios y Economía (Business & Economy)***.- Abarca alimentos y bebidas, asociaciones, automóviles, calzado y curtiduría, comercio electrónico, comercio internacional, construcción, cooperativas, derecho, eléctrica y electrónica, energía, farmacéutica, finanzas, formación, grandes empresas, hotelería, marketing, mayoristas y distribuidores, mercado mobiliario, metales, muebles, navegación, plásticos, productos de consumo, recursos humanos, servicios, telecomunicaciones, textiles, transportes y logística, vidrio.

- ***Noticias y Medios (News & Media)***.- Abarca agencias, bitácoras, buscadores y directorios, digitales, periódicos, periodismo, podcasting, radios, revistas, televisión, blogs, columnas, columnistas, broadcasts.

- ***Recreación y Deportes (Recreation & Sports)***.- Abarca deportes, viajes, parques de diversiones y temáticos, cocina, baile, apuestas, hobbies, motocicletas, aire libre, mascotas, deportes.

- ***Referencia (Reference)***.- Abarca archivos, banderas, bibliografías, bibliotecas, citas célebres, diccionarios, enciclopedias, mapas, museos, publicaciones electrónicas, refranes y dichos, unidades y medidas.

- ***Regional (Regional)***.- Abarca África, América, Antártida, Argentina, Asia, Bolivia, Chile, Colombia, Costa Rica, Cuba, Ecuador, El Salvador, España, Estados Unidos, Europa, Países, gobierno, regiones.

- ***Salud (Health)***.- Abarca adicciones, anatomía, animales, belleza y cosmética, enfermedades, enfermería, fitness, medicina, medicina alternativa, medicina preventiva, salud pública, mujeres, niños, nutrición, odontología, primeros auxilios, salud mental, sexología, seguridad.

- ***Sociedad y Cultura (Society & Culture)***.- Abarca asociaciones, astrología, compras, hogar, consejos, culturas, familia, festividades, filosofía, gente, leyendas urbanas, mitología, moda, muerte, paranormal, religión, relaciones.

4.3.2. Definición de Términos Relevantes para cada categoría

Los términos que caracterizan a cada categoría fueron definidos utilizando las páginas de entrenamiento. Después de calcular los coeficientes TFIDF para todas las secciones de las páginas, y realizar pruebas con diferentes números de términos, se seleccionaron 16 términos (8 en inglés y 8 en español) para cada categoría, ya que este número de términos, es el que mayor exactitud produjo.

En la tabla 5, se muestra cada uno de los términos seleccionados:

Tabla 5.- Términos Relevantes por Categoría

Categoría	Términos ingles	Términos español
Artes y Humanidades (Arts & Humanities)	art, design, animation, architecture, graphic, humanities, photo, gallery	arte, diseño, animación, arquitectura, gráfico, humanidad, fotografía, galería
Ciencia (Science)	science, aeronautic, aerospace, forensic, geography, anthropology, astronomy, biology	ciencia, geografía, biología, química, sistema, ecología, física, meteorología
Ciencias Sociales (Social Sciences)	social, language, study, theory, urban, women, sociology, research	lenguaje, estudio, teoría, filosofía, idioma, social, urbano, mujer
Computadoras e Internet (Computers & Internet)	computer, hardware, software, program, freeware, shareware, information, technology	computadora, programa, información, tecnología, software, hardware, freeware, shareware
Educación (Education)	education, school, university, academic, college, scholarship, institute, teach	educación, escuela, universidad, academia, secundaria, beca, instituto, profesor
Entretenimiento (Entertainment)	entertainment, magic, game, trivia, movie, music, actor, humor	entretenimiento, magia, juego, película, cine, música, actor, parque
Negocios y Economía (Business & Economy)	business, economy, retailer, marketplace, commerce, finance, sell, offer	economía, negocio, vender, comprar, finanzas, oferta, servicio, franquicia
Noticias y Medios (News & Media)	news, media, magazine, newspaper, people, radio, television, broadcast	noticia, revista, periódico, periodismo, medio, radio, televisión, broadcast
Recreación y Deportes (Recreation & Sports)	sport, recreation, auto, travel, outdoor, hobby, pet, map	deporte, recreación, viaje, puntaje, campeonato, torneo, equipo, liga
Referencia (Reference)	reference, phone, number, address, quotation, library, dictionary, encyclopedia	postal,referencia,direccion,libreria,biblioteca,diccionario,enciclopedia,calendario
Regional (Regional)	country, govern, region, asia, europe, america, state, coast	región, país, gobierno, asia, europa, estado, sierra, América
Salud (Health)	health, disease, drug, fitness, nutrition, pharmacy, condition, medicine	salud, enfermedad, medicamento, pastilla, nutrición, farmacia, medicina, doctor
Sociedad y Cultura (Society & Culture)	society, culture, family, myth, folklore, relationship, love, food	sociedad, cultura, folklore, familia, mito, relación, pareja, amor

4.4. CONJUNTO DE DATOS DE PRUEBA

Para el módulo de clasificación, se definió un conjunto de datos de prueba que consta de 13000 páginas web, que corresponden a 1000 páginas por categoría (500 en inglés y 500 en español). De este conjunto el 70% de las páginas (9100) se utilizó para el entrenamiento de los algoritmos y el 30% (3900) para comprobar los resultados entregados por el módulo.

Además de las pruebas realizadas con el módulo de manera independiente, se realizaron varias pruebas para comprobar la mejora en el sistema PSearch. Para ello se definió un conjunto de 100 consultas sobre diversos temas, que abarcan todas las categorías.

4.4.1. UTILIZACIÓN DEL SOFTWARE DESARROLLADO EN LOS DATOS SELECCIONADOS.

Una vez que se desarrolló el módulo de clasificación, y se pasó por todas las pruebas necesarias para su correcta integración al sistema PSearch, se procedió a ejecutar dicho módulo con el conjunto de datos obtenido. Esta ejecución consta de tres fases, dos de las mismas viendo al módulo como un componente independiente y la última como parte del sistema PSearch: entrenamiento del módulo de clasificación, evaluación del módulo de clasificación e integración del módulo de clasificación con el sistema PSearch.

4.4.2. ENTRENAMIENTO DEL MÓDULO DE CLASIFICACIÓN

Una vez que se recopiló el conjunto de datos de prueba, se procedió a extraer el contenido de cada una de las páginas recopiladas. De las 1000 páginas recopiladas por categoría, muchas no permitieron extraer el contenido, o dicho contenido era solamente tags de html, por lo que fueron descartadas.

La tabla 6, muestra el resultado final de la extracción de contenidos de páginas web.

Categoría	Número de Páginas
Artes y Humanidades (Arts & Humanities)	701
Ciencia (Science)	705
Ciencias Sociales (Social Science)	666
Computadoras e Internet (Computers & Internet)	678
Educación (Education)	746
Entretenimiento (Entertainment)	694
Negocios y Economía (Business & Economy)	828
Noticias y Medios (News & Media)	529
Recreación y Deportes (Sports & Recreation)	703
Referencia (Referente)	737
Regional (Regional)	634
Salud (Health)	780
Sociedad y Cultura (Society & Culture)	682

Tabla 6.- Páginas Obtenidas por Categoría

Las páginas entonces pasaron por el pre – procesamiento del módulo de clasificación y se obtuvo un conjunto de términos para cada una de las categorías. La tabla 7, muestra el número de términos obtenidos por categoría:

Categoría	Número de Términos
Artes y Humanidades (Arts & Humanities)	109493
Ciencia (Science)	130963
Ciencias Sociales (Social Science)	144024
Computadoras e Internet (Computers & Internet)	171573
Educación (Education)	127569
Entretenimiento (Entertainment)	102159
Negocios y Economía (Business & Economy)	100500
Noticias y Medios (News & Media)	153492
Recreación y Deportes (Sports & Recreation)	130075
Referencia (Reference)	136618
Regional (Regional)	125056
Salud (Health)	166521
Sociedad y Cultura (Society & Cultura)	145399

Tabla 7.- Términos Obtenidos por Categoría

Una vez que se descompuso el contenido de las páginas en términos individuales, en su raíz primitiva, se procedió a calcular el coeficiente TFIDF de cada uno de los términos, que permitió seleccionar los términos con coeficiente más alto para representar a cada una de las categorías. Como se mencionó en la sección 4.3.2 se seleccionó 16 términos por categoría cuyos coeficientes calculados fueron los más altos (8 en inglés y 8 en español). Se seleccionaron en total 16 términos, ya que con este número se alcanzaron los mejores resultados en cuanto a exactitud y tiempo de respuesta. Un número pequeño de términos, nos daba como resultado una exactitud pobre pero un tiempo de respuesta

muy bajo. Un número de términos alto en cambio nos daba una buena exactitud pero el tiempo de respuesta era muy alto. En la tabla 8, se muestra los resultados obtenidos para los diferentes números de términos:

Número de Términos por Categoría	Aciertos Español (%)	Tiempo de Ejecución (s)	Aciertos Inglés (%)	Tiempo de Ejecución (s)	Aciertos Totales (%)	Tiempo de Ejecución (s)
1	57	0.001	58	0.002	57.5	0.0015
2	56	0.001	57	0.001	56.5	0.001
3	61	0.002	59	0.015	60	0.0085
4	68	0.001	70	0.021	69	0.01075
5	74	0.003	76	0.002	75	0.00225
6	80	0.005	78	0.022	79	0.0135
7	89	0.003	87	0.003	88	0.003
8	**93**	**0.050**	**91**	**0.045**	**92**	**0.0475**
9	92	0.070	93	0.050	92.5	0.06
10	91	0.068	89	0.071	90	0.0695
11	93	0.078	91	0.069	92	0.07325
12	93	0.087	90	0.054	91.5	0.0705
13	94	0.085	91	0.076	92.5	0.0805
14	92	0.097	94	0.086	93	0.0915
15	95	0.099	93	0.098	94	0.0985
16	96	0.100	93	0.110	94.5	0.105

Tabla 8.- Exactitud y Tiempo de Respuesta de acuerdo al número de términos

De esta manera se escogió un número intermedio de términos, en los que esté balanceada la exactitud y tiempo de respuesta.

La tabla 9, muestra el coeficiente TFIDF de cada uno de los términos seleccionados por categoría. Para ver los coeficientes encontrados para cada uno de los términos de las categorías, referirse a los anexos.

Categoría	Rango TFIDF (Términos Español)	Rango TFIDF (Términos Inglés)
Artes y Humanidades (Arts & Humanities)	3.36 – 5.04	3.26 – 4.89
Ciencia (Science)	3.20 – 5.11	3.19 – 5.16
Ciencias Sociales (Social Science)	3.42 – 5.13	3.76 – 6.11
Computadoras e Internet (Computers & Internet)	3.55 – 5.33	5.44 – 8.17
Educación (Education)	3.29 – 4.93	3.58 – 5.37
Entretenimiento (Entertainment)	4.01 – 6.01	3.29 – 4.93
Negocios y Economía (Business & Economy)	5.44 – 8.17	4.73 – 7.09
Noticias y Medios (News & Media)	5.43 – 6.80	5.87 – 8.80
Recreación y Deportes (Sports & Recreation)	3.58 – 5.36	3.58 – 5.37
Referencia (Reference)	4.63 – 6.94	3.54 – 5.31
Regional (Regional)	4.92 – 7.38	5.12 – 7.68
Salud (Health)	3.12 – 5.02	3.33 – 5.00
Sociedad y Cultura (Society & Cultura)	3.23 – 5.09	3.18 – 5.04

Tabla 9.- Coeficientes TFIDF para Términos Seleccionados

Al contar con los términos que caracterizan a cada categoría, se procedió al entrenamiento de los algoritmos que lo necesitaban, como la red neuronal y el algoritmo bayesiano. El algoritmo por frecuencia de términos simplemente recibió como entrada los términos seleccionados.

4.4.3. EVALUACIÓN DEL MÓDULO DE CLASIFICACIÓN

Con la finalidad de evaluar el módulo propuesto, se ingresó el 30% (2724) de los links obtenidos para comprobar los resultados del módulo de clasificación. La tabla 10, muestra los resultados obtenidos con relación a la exactitud del clasificador una vez que las páginas de prueba fueron aplicadas al sistema entrenado. Esta tabla resume los resultados dentro de cada categoría así como el promedio global:

	Aciertos (Español)	**Porcentaje**	**Aciertos (Inglés)**	**Porcentaje**	**Aciertos Totales**	**Porcentaje**
Artes y Humanidades (Arts & Humanities)	96	92	99	95	194	94
Ciencia (Science)	96	92	94	90	189	91
Ciencias Sociales (Social Science)	90	87	93	89	183	88
Computadoras e Internet (Computers & Internet)	100	96	102	98	202	97
Educación (Education)	98	94	96	92	193	93
Entretenimiento (Entertainment)	94	90	95	91	188	91
Negocios y Economía (Business & Economy)	95	91	92	88	186	90
Noticias y Medios (News & Media)	94	90	94	90	187	90
Recreación y Deportes (Recreation & Sports)	93	89	93	89	185	89
Referencia (Reference)	93	89	92	88	184	89
Regional (Regional)	87	84	90	87	178	86
Salud (Health)	95	91	96	92	190	92
Sociedad y Cultura (Society & Cultura)	95	91	92	88	186	90
Promedio		90		91		91

Tabla 10.- Resultados de Exactitud del Módulo de Clasificación

Con relación al tiempo de ejecución, se obtuvieron los resultados que se muestran en la tabla 11:

Categoría	Descarga de Páginas (seg)	Pre - Procesamiento (seg)	Clasificación (seg)	Total (seg)
Artes y Humanidades (Arts & Humanities)	0.9	0.09	0.03	1.02
Ciencia (Science)	0.81	0.08	0.05	0.94
Ciencias Sociales (Social Science)	0.42	0.01	0.01	0.44
Computadoras e Internet (Computers & Internet)	0.5	0.05	0.06	0.61
Educación (Education)	0.45	0.04	0.1	0.59
Entretenimiento (Entertainment)	0.78	0.06	0.06	0.9
Negocios y Economía (Business & Economy)	0.7	0.02	0.08	0.8
Noticias y Medios (News & Media)	0.65	0.02	0.1	0.77
Recreación y Deportes (Recreation & Sports)	0.54	0.07	0.05	0.66
Referencia (Reference)	0.31	0.09	0.05	0.45
Regional (Regional)	0.79	0.05	0.01	0.85
Salud (Health)	0.92	0.01	0.02	0.95
Sociedad y Cultura (Society & Cultura)	0.92	0.02	0.1	1.04
Promedio	0.66	0.046	0.055	0.77

Tabla 11.- Resultados de Desempeño del Módulo de Clasificación

4.4.4. INTEGRACIÓN CON EL SISTEMA PSEARCH

Finalmente se integró el módulo de clasificación al sistema PSearch y se realizaron 100 búsquedas con el objetivo de probar la exactitud y velocidad del sistema integrado. Para realizar las búsquedas, se utilizó las opciones de personalización con que cuenta el sistema PSearch.

Se hicieron cuatro tipos de pruebas para exactitud y en cada una se midió también el tiempo de ejecución.

La primera prueba consistió en seleccionar en las preferencias del usuario una única categoría que esté muy relacionada con la búsqueda. El sistema PSearch posee dos tipos de perfiles, fijo y adaptativo. El perfil fijo se mantiene durante todas las búsquedas, mientras que el perfil adaptativo va modificándose de acuerdo a los resultados seleccionados por el usuario. De tal manera se realizaron dos variaciones para la primera prueba, búsquedas realizadas con perfil fijo y con perfil adaptativo.

La segunda prueba consistió en seleccionar en las preferencias del usuario más de una categoría que esté muy relacionada con la búsqueda. Al igual que en la prueba anterior se realizó la variación con perfil fijo y adaptativo.

La tercera prueba consistió en seleccionar en las preferencias del usuario una única categoría al azar, es decir esta no estaba relacionada directamente con la búsqueda. Al igual que en la prueba anterior se realizó la variación con perfil fijo y adaptativo.

La cuarta prueba consistió en seleccionar en las preferencias del usuario más de una categoría al azar, es decir esta no estaba relacionada directamente con la búsqueda. Al igual que en la prueba anterior se realizó la variación con perfil fijo y adaptativo.

Para todas las pruebas también se seleccionó al azar el idioma, tratando de que haya igual cantidad de búsquedas en español como en inglés.

La tabla 12 muestra los resultados obtenidos para las 100 búsquedas realizadas antes y después de la incorporación del módulo de clasificación:

		Sin Módulo de Clasificación		Con Módulo de Clasificación	
Número de Prueba	**# de Búsquedas/Prueba**	**Exactitud (%)**	**Velocidad (s)**	**Exactitud (%)**	**Velocidad (s)**
Prueba 1	42	84	0.046	90	0.048
Prueba 2	7	74	0.05	92	0.054
Prueba 3	47	74	0.051	89	0.053
Prueba 4	4	75	0.057	94	0.062
Promedio	100	76.75	0.051	91.25	0.05425

Tabla 12.- Tabla resumen de los resultados de Integración del Módulo.

4.4.5. ANÁLISIS DE RESULTADOS

Una vez que el módulo de clasificación fue entrenado, y evaluado independientemente así como integrado al sistema PSearch y probado, podemos observar algunos de los resultados obtenidos con el fin de comprobar si el módulo ayudó o no al sistema PSearch a mejorar los resultados entregados a sus usuarios.

Durante la fase de evaluación del módulo, se conoció que en promedio el número de aciertos fue de un 91%. De los cuales existió mayor cantidad de errores en páginas en español que en inglés, los aciertos en español en promedio fueron de 90% mientras que en inglés fueron de un 91%. Esta diferencia se debe en gran parte a la exactitud de los algoritmos para obtener la raíz primitiva del contenido de la página. El algoritmo para el idioma inglés ha sido mucho más probado y es más exacto que el algoritmo en español.

La categoría en la que menor cantidad de aciertos se encontraron en promedio (86%), fue la categoría Regional, en esta categoría, la elección de los términos que la iban a representar fue difícil, ya que muchos términos tenían el mismo coeficiente TFIDF. Esto

nos indica que las páginas de dicha categoría contienen información sobre temas muy variados, que dificultan la determinación de que es lo que tienen en común todas ellas.

La categoría en la que más aciertos se encontraron en promedio (97%), fue la categoría Computadores e Internet, en esta categoría la selección de términos fue mucho más fácil, dado que en la mayor parte de las páginas analizadas, los mismos términos fueron encontrados al menos una vez. Esto facilita la clasificación ya que las páginas tienen un tema en común que permite identificarlas.

El resultado obtenido entonces para la exactitud del módulo, fue de un 91% de aciertos en promedio, con páginas tanto en español como en inglés.

En cuanto al tiempo de ejecución, se pudo determinar que el mayor costo se encuentra en la descarga de la página Web y no en el procesamiento de la misma. Es decir la aplicación se encuentra limitada por la latencia de la red y no por la velocidad de procesamiento.

En cuanto a la integración del módulo de clasificación al sistema PSearch se realizaron los cuatro tipos de pruebas descritas en la sección anterior. Para la prueba 1 (Perfil fijo o adaptativo y 1 sola categoría relacionada con la búsqueda) se obtuvo una mejora del 6% en cuanto a exactitud y un incremento de 0.0019 s en el tiempo de ejecución.

Para la prueba 2 (Perfil fijo o adaptativo y 2 o más categorías relacionadas con la búsqueda) se obtuvo una mejora del 18% en cuanto a exactitud y un incremento de 0.0042 s en el tiempo de ejecución.

Para la prueba 3 (Perfil fijo o adaptativo y 1 sola categoría aleatoria no relacionada con la búsqueda) se obtuvo una mejora del 15% en cuanto a exactitud y un incremento de 0.0027s en el tiempo de ejecución.

Para la prueba 4 (Perfil fijo o adaptativo y 2 o más categorías aleatorias no relacionadas con la búsqueda) se obtuvo una mejora del 19% en cuanto a exactitud y un incremento de 0.005 s en el tiempo de ejecución.

En la tabla 13, podemos observar los resultados mencionados anteriormente:

		Sin Módulo de Clasificación		**Con Módulo de Clasificación**		**Diferencia**	
Número de Prueba	**# de Búsquedas/Prueba**	**Exactitud (%)**	**Velocidad (s)**	**Exactitud (%)**	**Velocidad (s)**	**Exactitud (%)**	**Velocidad (s)**
Prueba 1	42	84	0.046	90	0.048	6	0.002
Prueba 2	7	74	0.05	92	0.054	18	0.004
Prueba 3	47	74	0.051	89	0.053	15	0.002
Prueba 4	4	75	0.057	94	0.062	19	0.005
Promedio	100	76.75	0.051	91.25	0.05425	14.5	0.0032

Tabla 13.- Diferencia para exactitud y velocidad antes y después de la integración del módulo de clasificación

Como podemos observar un factor importante en la exactitud de los resultados son las preferencias seleccionadas por el usuario. Cuando las categorías seleccionadas se relacionan con los términos de la búsqueda, esta se hace más difícil, ya que debe ser más exacta, mientras que si las categorías no están completamente relacionadas con la búsqueda, se puede encontrar información más fácilmente, y de los resultados devueltos, mayor cantidad de ellos serán relevantes. Prueba de esto la encontramos en el porcentaje de mejora en cuanto a la exactitud ya que para la prueba 1 en la que se escogió una sola

categoría relacionada con la búsqueda la mejora es del 6%, mientras que para las pruebas 2, 3 y 4 la mejora es del 18%, 15% y 19%, pruebas en las que se escogió algunas categorías, que permitieron que mayor cantidad de resultados resulten relevantes.

Sin embargo, encontramos que sin importar las preferencias seleccionadas por el usuario, los resultados entregados por el sistema PSearch son mejores al integrar el módulo de clasificación.

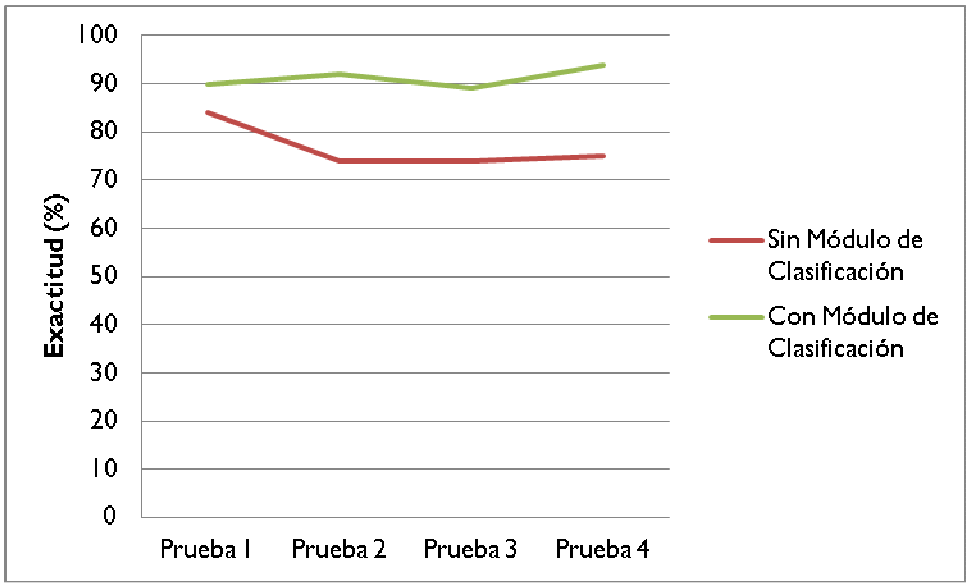


Diagrama 7.- Comparación en Exactitud

La mejora más significativa, se encontró en la Prueba 4, donde el porcentaje de aciertos fue de 94%. Antes de integrar el módulo al sistema, el porcentaje de aciertos era de 75%, pero después de la inclusión del módulo, se encontró una mejora del 19%. Este porcentaje de mejora se debe principalmente, a que las páginas ahora son clasificadas de manera más exacta. De acuerdo a la mejora, podemos decir que los términos seleccionados por categoría lograron definirlas, y se encuentran en las páginas que

corresponden a ellas. En general se obtuvo una mejora promedio total en la exactitud del 14.5%.

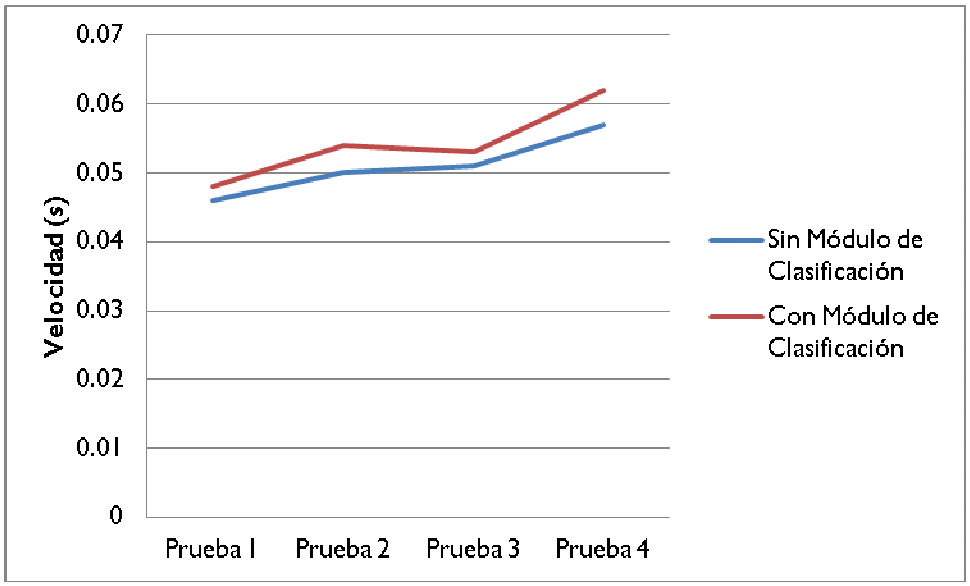


Diagrama 8.- Comparación en Velocidad

En el diagrama 8 encontramos la comparación de los tiempos de procesamiento con y sin el módulo de clasificación integrados. Si bien es cierto el tiempo con el módulo es mayor, el tiempo de ejecución no se vio afectado en gran medida, subió en 0.0019 s. Esta subida de tiempo se justifica, ya que mayor exactitud requiere de algoritmos cuyo procesamiento toma más tiempo, sin embargo se compensa al entregar al usuario resultados que le son más útiles. Si analizamos la subida de tiempos por prueba, podemos observar que mantiene la tendencia, mientras más categorías se seleccionaron en las preferencias del usuario, mayor es el tiempo que tomó la búsqueda. Para la Prueba 1, la subida es la menor, mientras que para la Prueba 4, la subida de tiempo es la mayor.

CAPÍTULO 5

CONCLUSIONES Y RECOMENDACIONES

1. Las páginas web contienen datos dinámicos acerca de muchos temas por lo que su clasificación no es sencilla.
2. Mediante la utilización de técnicas de pre-procesamiento sencillas, se obtuvo información crítica de cada unas de las páginas web, lo que permitió definir términos para su clasificación.
3. La identificación de una página web de acuerdo a su contenido, requiere la definición de características específicas de la misma que la diferencien del resto de páginas.
4. Los datos obtenidos del contenido de una página web, requieren pre - procesamiento, además deben ser analizados para su completa comprensión.
5. Una vez que se extrajo el contenido de las páginas web, el mismo se transformó para poder ser utilizado por cada uno de los algoritmos. Esta transformación fue necesaria para que cada uno de los algoritmos reciban los parámetros que necesitan para la clasificación. Este paso requirió de un procesamiento extra, que finalmente no afectó al tiempo total de clasificación.
6. El módulo de clasificación analiza por separado tres secciones contenidas en las páginas web, título, meta datos y cuerpo. Esto se debe a que el título y los meta datos contienen información importante para poder determinar a qué categoría pertenece la página que está siendo analizada. Esto le facilita al módulo la caracterización de las páginas, al contar con más información sobre la página.

7. Al incluir el módulo de clasificación en el sistema PSearch, la mejora permitió entregar al usuario información más orientada a sus preferencias en un 10%.
8. La elección de preferencias del usuario influye en la clasificación de las páginas web.
9. Si las preferencias del usuario están completamente relacionadas con la búsqueda realizada, esta se vuelve más compleja, ya que los datos deben ser más precisos y por tanto la exactitud disminuye.
10. Si las preferencias del usuario no están relacionadas con la búsqueda realizada, esta es más sencilla, ya que existe un mayor número de resultados relevantes a la misma, por tanto la exactitud es mayor.
11. El módulo de clasificación logró la mejora del sistema PSearch mediante la extracción de información crítica de cada uno de los documentos HTML, en un 14.5%.
12. Se debe balancear exactitud y desempeño, por ello el módulo de clasificación utiliza algoritmos poco costosos en tiempo de ejecución, para alcanzar una exactitud superior al 90%.
13. El número de páginas utilizado para el entrenamiento fue suficiente, permitiéndonos encontrar los términos que diferencian una clase de otra.
14. A pesar de que se incluyó nuevo código al sistema, el tiempo de respuesta se pudo mantener, siendo la descarga de las páginas, el proceso que más tiempo tomó. Es decir que el tiempo de procesamiento se vio afecto por la latencia de la red, mas no por la inclusión del módulo de clasificación.
15. Incluir un algoritmo más preciso para la determinación de la raíz primitiva de los términos, incrementará la exactitud del módulo de clasificación. Se necesita un algoritmo por idioma que se desee incluir en el buscador. Se realizaron pruebas con búsquedas en

otros idiomas diferentes al inglés y español, por ejemplo en francés, sin embargo los resultados en exactitud son muy bajos, ya que no se cuenta con un algoritmo para manejar este idioma.

16. Incluir mayor número de idiomas al módulo de clasificación, permitirá que mayor número de consultas puedan entregar resultados correctos. El módulo de clasificación determina el idioma, para su pre - procesamiento, si es un idioma no definido, el retirar términos comunes o reducir las palabras a su raíz primitiva se torna en una tarea compleja, ya que esto depende de las particularidades de cada idioma, por lo que los resultados no son buenos.

ANEXOS

OPERACIONES DEL MÓDULO DE CLASIFICACIÓN

Las operaciones básicas del módulo de clasificación se describen a continuación en los diagramas de flujo de datos.

Diagramas de Flujo de Datos

Diagrama de Contexto

El diagrama de contexto constituye la presentación inicial del módulo de clasificación. Se lo muestra como una gran caja negra, con el objetivo de exponer las interacciones del módulo con su entorno, en este caso el sistema PSearch. El diagrama muestra en la parte central, el módulo de clasificación y a los lados las entidades con las cuales se relaciona; el sistema PSearch y la base de datos del módulo. También podemos observar los flujos de datos que van hacía las entidades externas.

DIAGRAMA DE CONTEXTO

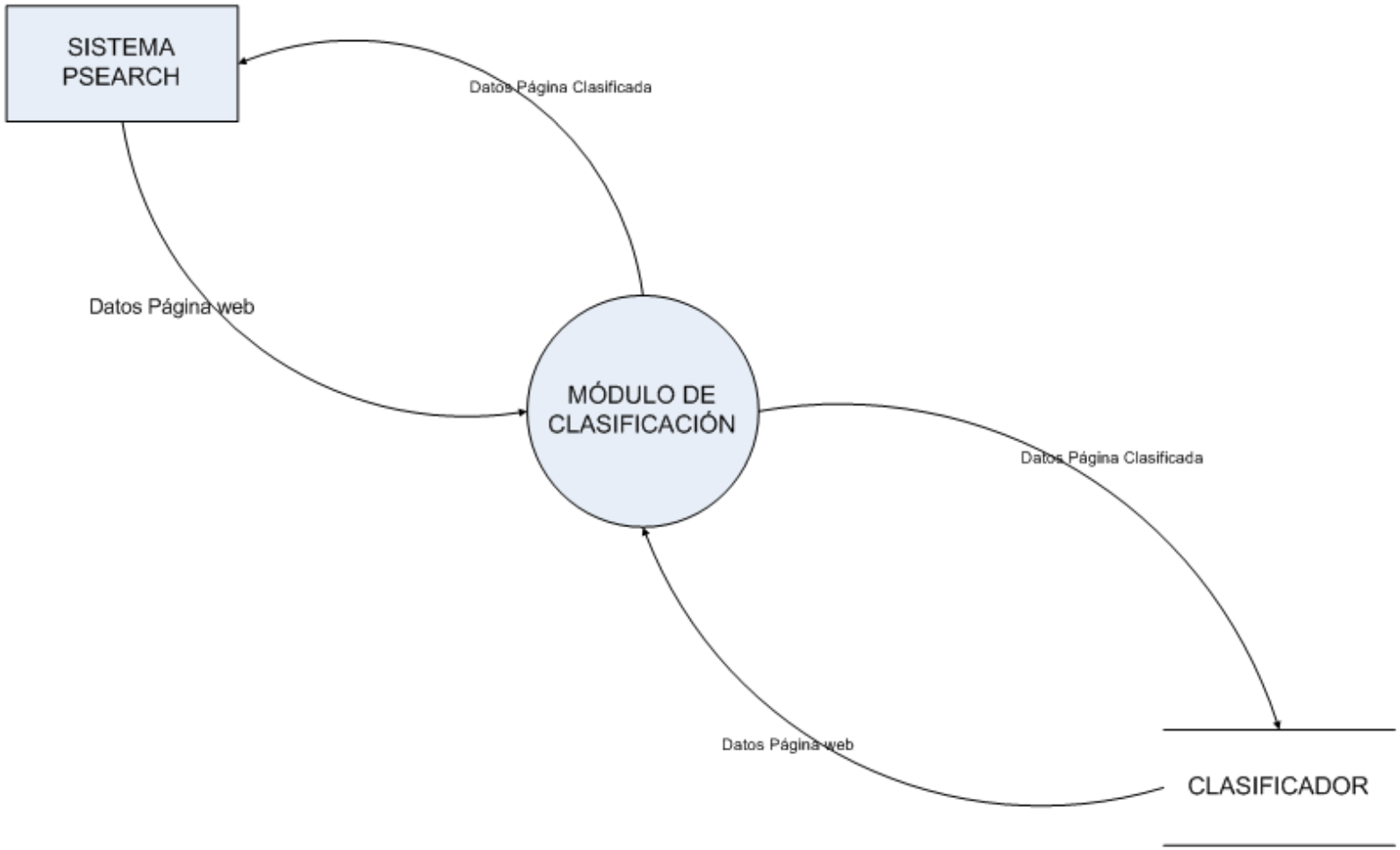


FIGURA 0

En la figura 0, se plasman todos los procesos que describen al proceso principal, además se muestra los flujos de datos y almacenes relacionados. Cada proceso ha sido enumerado para mostrarlos a mayor detalle en diagramas posteriores.

1.- Entrenar: El proceso entrenar recibe como flujo de entrada los urls de entrenamiento y el flujo de salida lo constituyen los términos frecuentes para cada categoría

2.- Pre – procesar: El segundo proceso, recibe como flujo de entrada los datos de una página web y entrega finalmente los datos de la página limpios, listos para clasificarlos.

3.- Clasificar: El proceso clasificar recibe los datos pre procesados y finalmente entrega como salida la clasificación de la página web, es decir el nombre de una de las categorías.

FIGURA 0

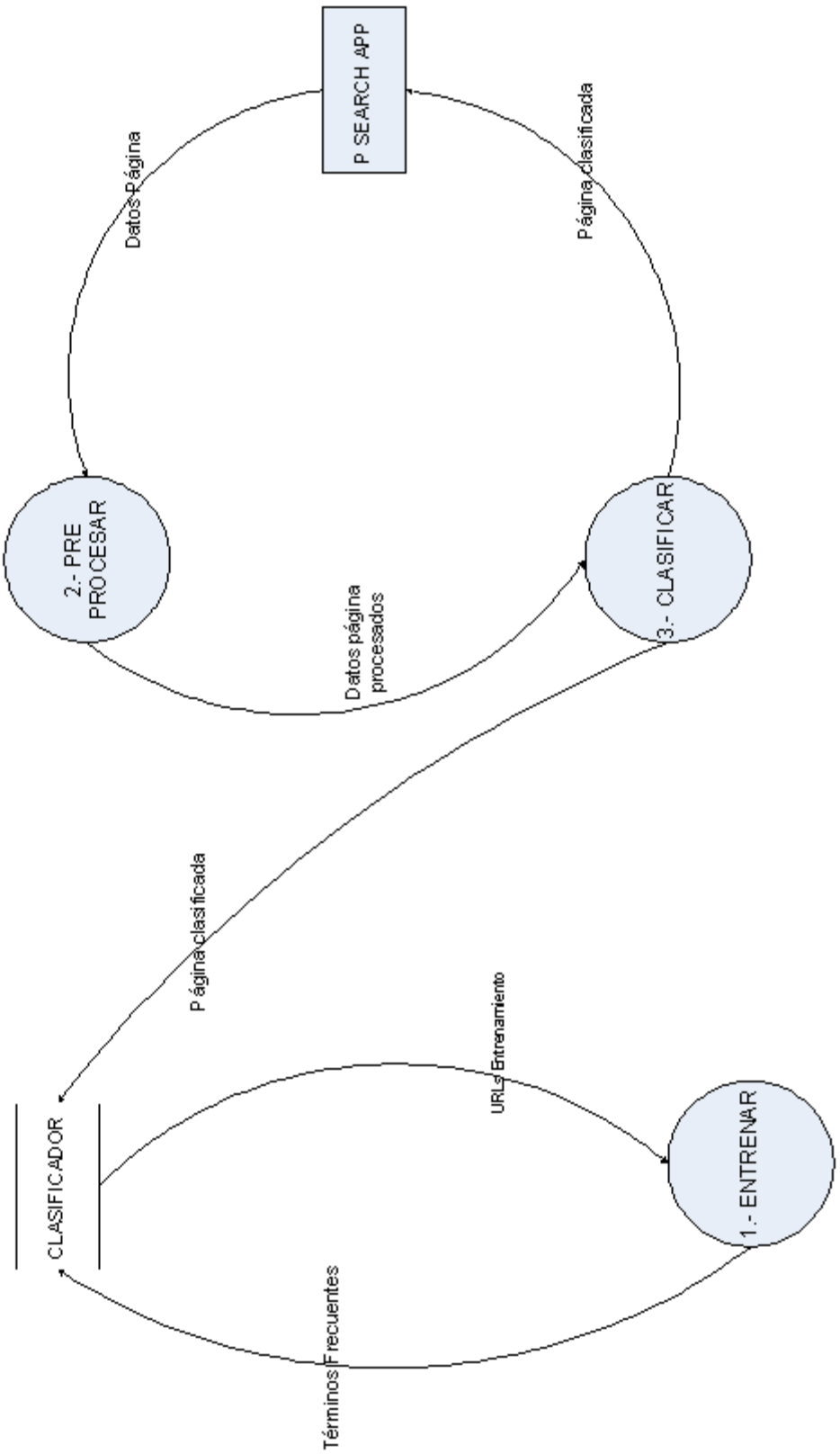

P SEARCH APP
Datos Página
Página clasificada
2.- PRE PROCESAR
3.- CLASIFICAR
Datos página procesados
Página clasificada
CLASIFICADOR
URLs Entrenamiento
1.- ENTRENAR
Términos Frecuentes

FIGURA 1

En la figura 1, se muestran todos los procesos necesarios para llevar a cabo la tarea de entrenamiento.

1.1.- Limpiar Datos: El proceso limpiar datos recibe como flujo de entrada los urls de entrenamiento y el flujo de salida lo constituyen los datos de las páginas procesados.

1.2.- Calcular TFID: El segundo proceso, recibe como flujo de entrada los datos de los urls de entrenamiento procesados. Retorna los valores TFIDF de cada página.

1.3.- Escoger Lexemas: El proceso escoger lexemas recibe los valores TFIDF para entregar el conjunto de lexemas resultado del procesamiento.

1.4.- Formar Vectores: El cuarto proceso necesita como entrada el conjunto de lexemas que resultó del entrenamiento para formar los vectores que definen a las categorías de clasificación.

1.5.- Entrenar Red Neuronal: Este proceso recibe los vectores para entrenamiento, de tal forma que se obtiene como salida los pesos necesarios para que las neuronas clasifiquen una página web.

1.6.- Entrenar Bayesian: Este proceso recibe los vectores para entrenamiento, de tal forma que se obtiene como salida las probabilidades que determinarán la clasificación de una página web.

FIGURA 1

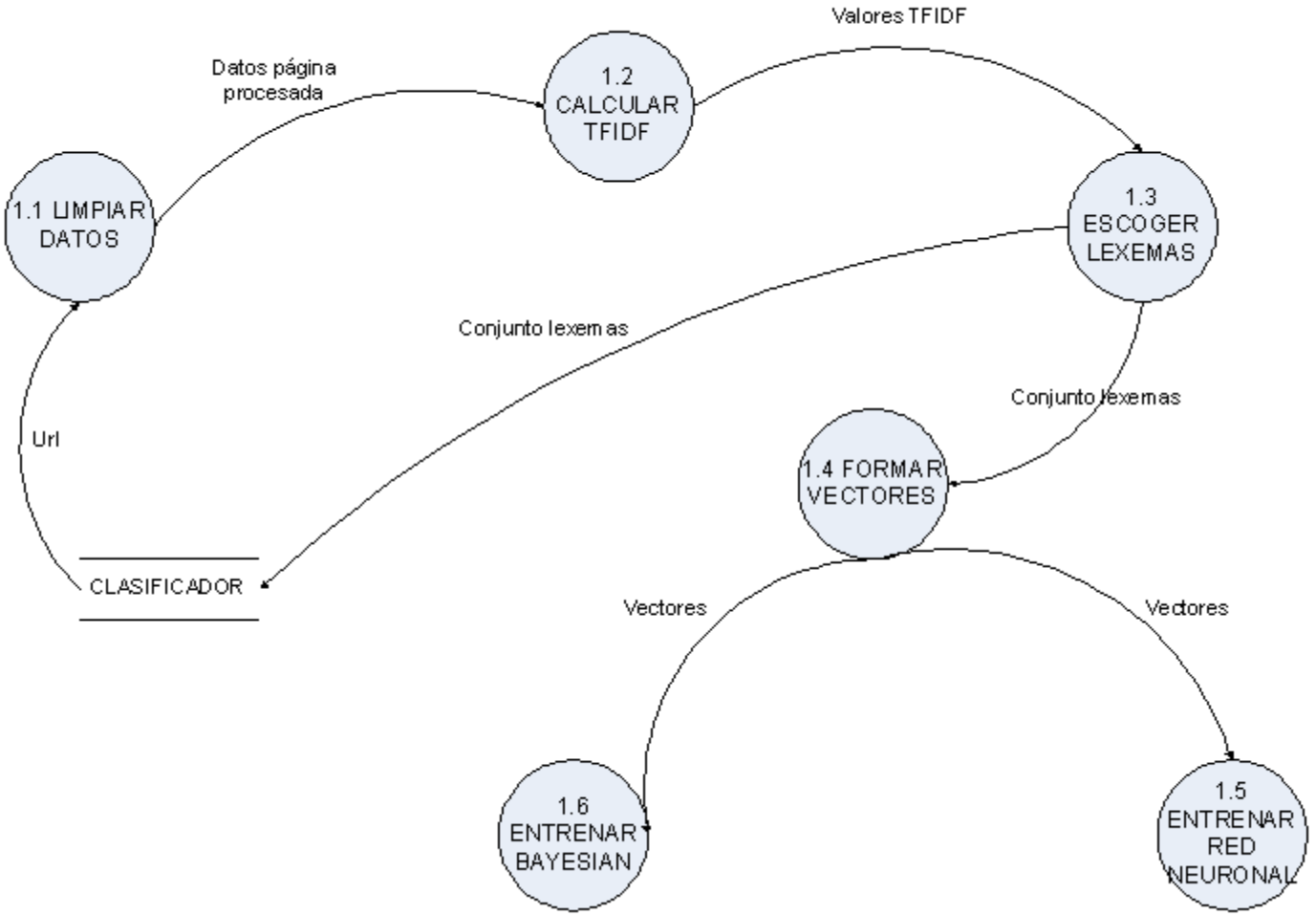

Valores TFIDF
Datos página procesada
1.2 CALCULAR TFIDF
1.1 LIMPIAR DATOS
1.3 ESCOGER LEXEMAS
Conjunto lexemas
Conjunto lexemas
Url
1.4 FORMAR VECTORES
CLASIFICADOR
Vectores
Vectores
1.6 ENTRENAR BAYESIAN
1.5 ENTRENAR RED NEURONAL

FIGURA 2

En la figura 2, se muestran todos los procesos necesarios para llevar a cabo la tarea de pre-procesamiento.

2.1.- Descargar Página: El proceso descargar página recibe como flujo de entrada un url del sistema PSearch y el flujo de salida lo constituyen los datos de la página descargada.

2.2.- Separar Página: El segundo proceso, recibe como flujo de entrada los datos de la página descargada. Retorna el título, meta datos y body de la página.

2.3.- Filtrar Página: El proceso filtrar página recibe el título, meta datos y body de la página para entregar el conjunto de datos filtrados, es decir solamente con los términos que nos interesa analizar.

2.4.- Calcular TFID: El proceso, recibe como flujo de entrada los datos de la página web procesada. Retorna los valores TFIDF de la página.

2.5.- Formar Vector: Este proceso necesita como entrada el conjunto de lexemas con sus valores TFIDF asociados para formar el vector de términos que define a la página.

3.- Clasificar: El proceso clasificar recibe el vector formado para determinar a qué categoría pertenece.

FIGURA 2

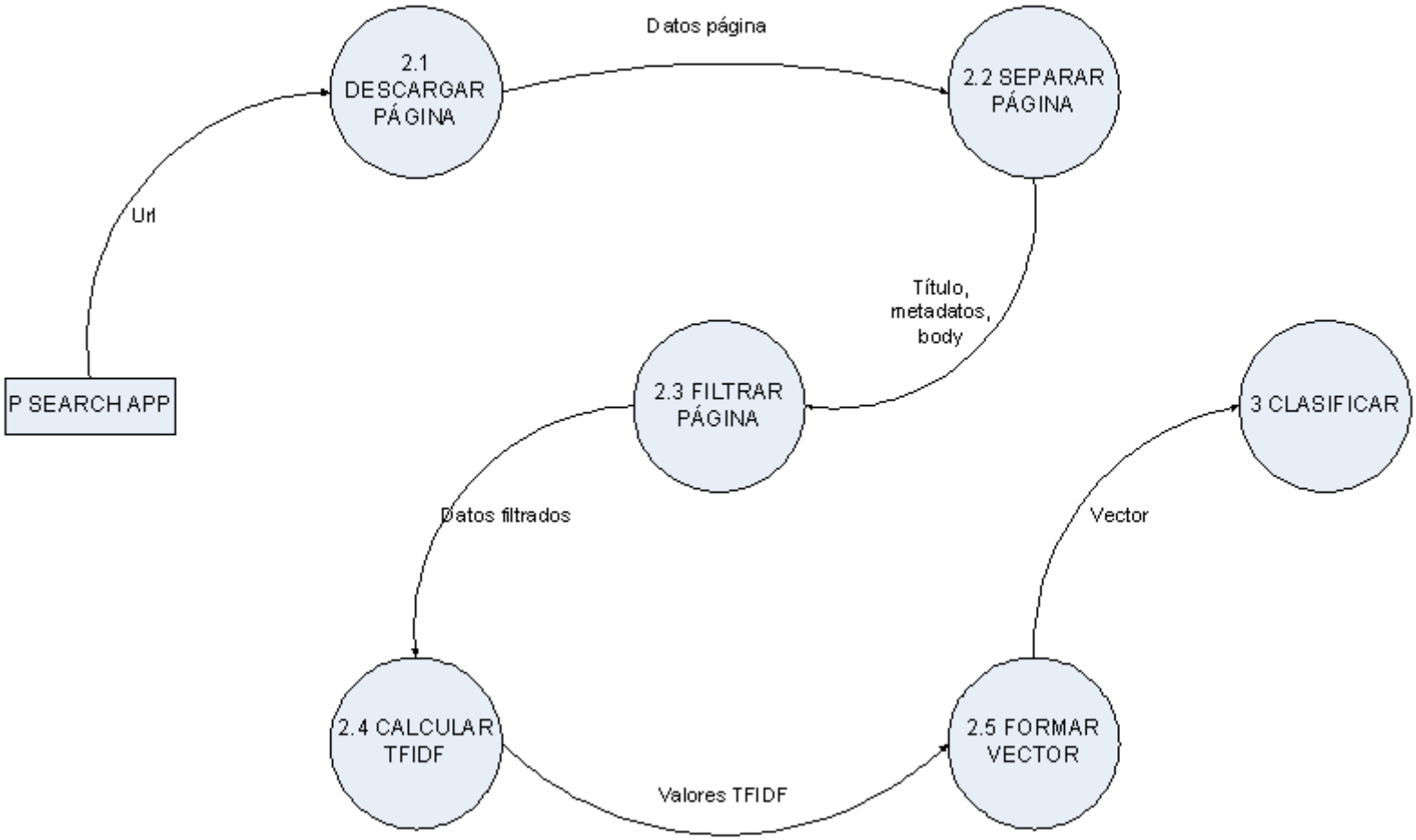

2.1
DESCARGAR
PÁGINA
Datos página
2.2 SEPARAR
PÁGINA
Url
Título,
metadatos,
body
P SEARCH APP
2.3 FILTRAR
PÁGINA
3 CLASIFICAR
Datos filtrados
Vector
2.4 CALCULAR
TFIDF
Valores TFIDF
2.5 FORMAR
VECTOR

FIGURA 3

En la figura 3, se muestran todos los procesos necesarios para llevar a cabo la tarea de clasificar.

2.- Pre – procesar: El proceso envía el vector formado para determinar a qué categoría pertenece la página web dada.

3.1.- Ejecutar Algoritmos: Este proceso recibe como flujo de entrada el vector que será utilizado por los algoritmos para finalmente enviar los resultados de clasificación.

3.2.- Comparar Resultados: El proceso, recibe como flujo de entrada los resultados de los clasificadores para compararlos y finalmente obtener el resultado final que es almacenado en la base de datos del módulo y enviada a su vez al sistema PSearch.

FIGURA 3

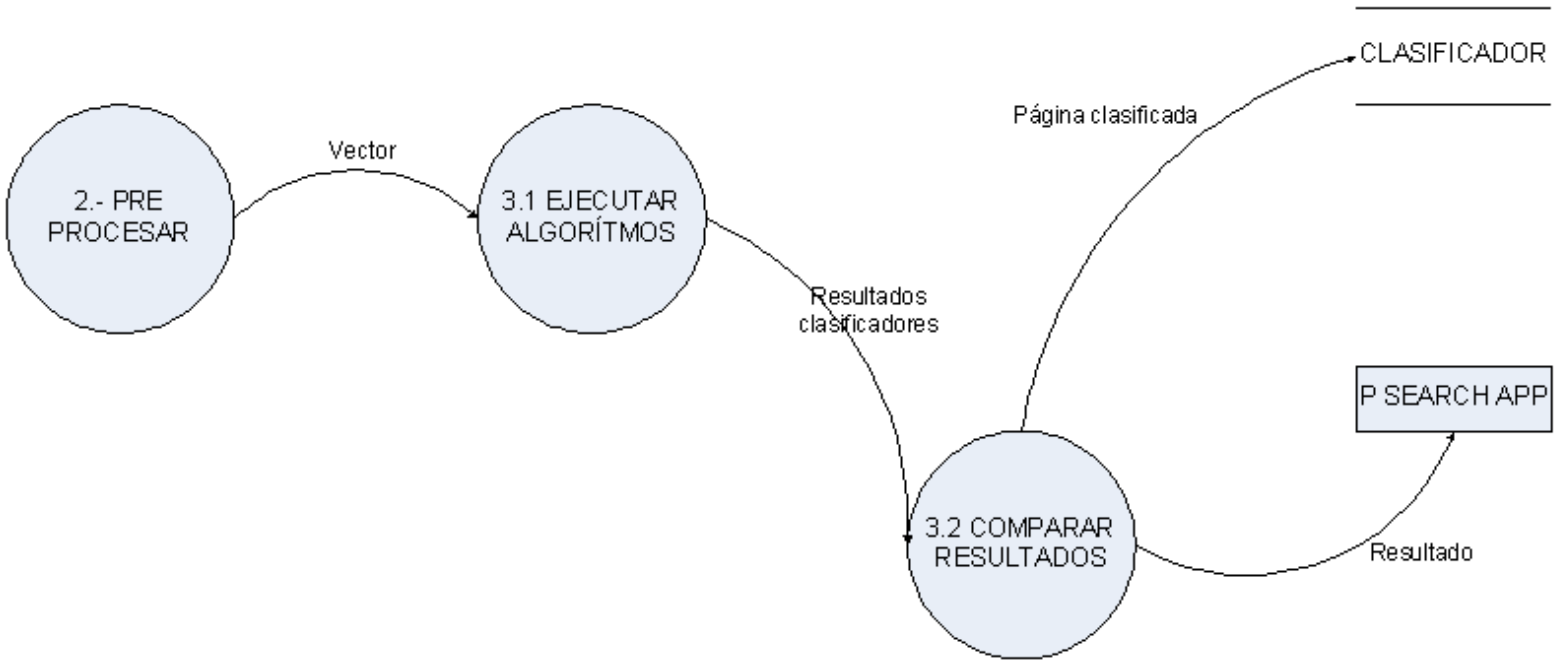
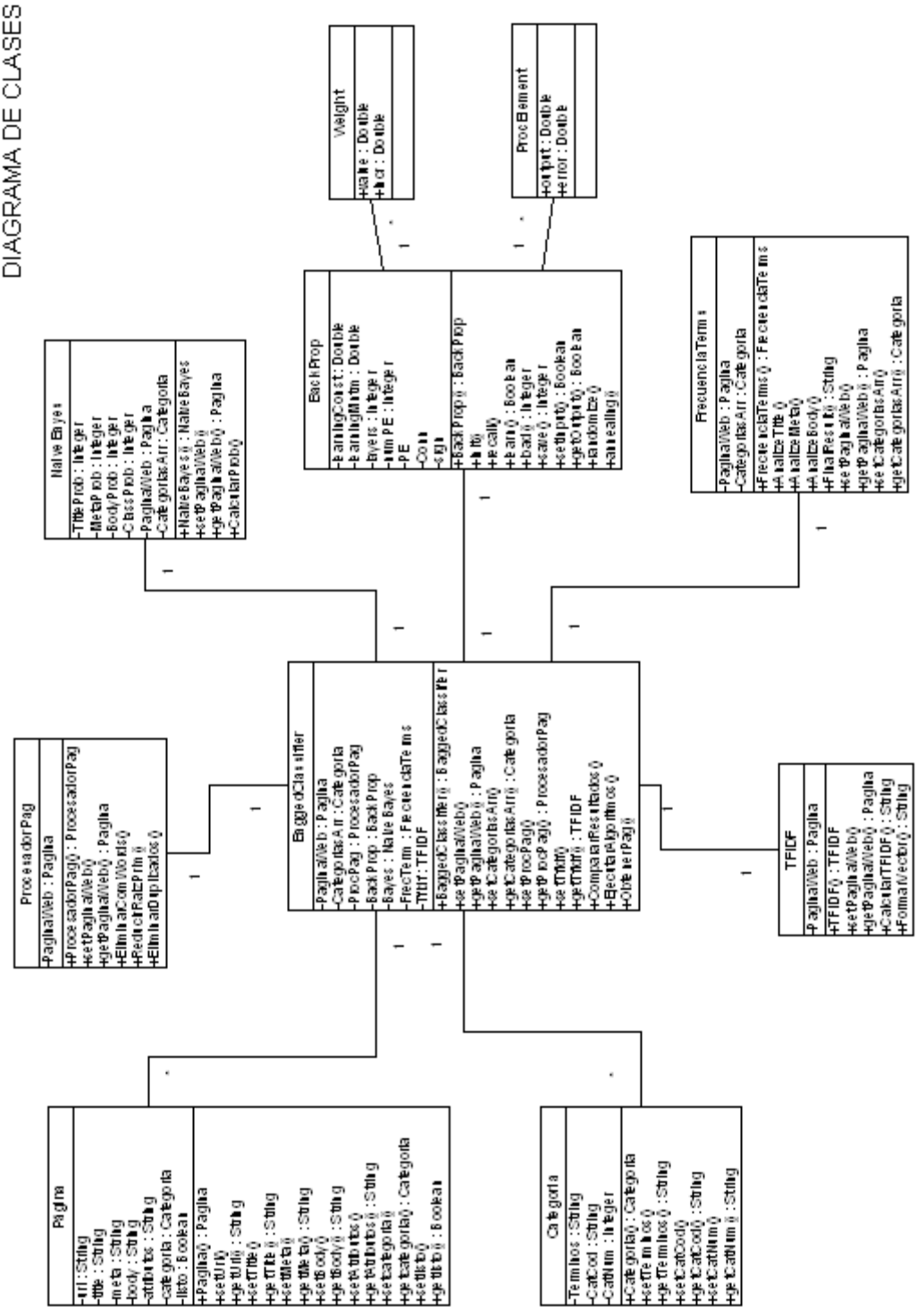

2.- PRE PROCESAR
Vector
3.1 EJECUTAR ALGORÍTMOS
Resultados clasificadores
3.2 COMPARAR RESULTADOS
Página clasificada
CLASIFICADOR
P SEARCH APP
Resultado

DIAGRAMA DE CLASES



COEFICIENTES TFIDF PARA LOS TÉRMINOS SELECCIONADOS POR CATEGORÍA

Categoría	Término	Idioma	TFIDF
Artes y Humanidades (Arts & Humanities)	Animation	Inglés	3.26
	Architecture		3.26
	Graphic		3.26
	Photo		3.26
	Art		3.63
	Design		4.89
	Humanities		4.89
	Gallery		4.89
	Arte	Español	3.36
	Fotografía		3.15
	Diseño		3.25
	Arquitetura		3.36
	Gráfico		3.36
	Galleria		3.36
	animación		5.04
	humanidad		5.04
Ciencia (Science)	aeronautic	Inglés	3.19
	anthropology		3.23
	forensic		3.29
	Science		3.44
	aerospace		3.44
	geography		3.44
	astronomy		5.16
	Biology		5.16
	geografia	Español	3.20
	sistema		3.30
	Ciencia		3.41
	biologia		3.41
	quimica		3.41
	ecologia		3.41
	fisica		5.11
	meteorologia		5.11
Ciencias Sociales (Social Sciences)	language	Inglés	3.76
	urban		3.79
	social		4.08
	study		4.08
	theory		4.08
	women		4.08

	sociology		6.11
	research		6.11
	lenguaje	Español	3.42
	estudio		3.42
	teoria		3.42
	filosofia		3.42
	idioma		3.42
	social		3.42
	urbano		5.13
	mujer		5.13
Computadoras e Internet (Computers & Internet)	computer	Inglés	5.44
	hardware		5.44
	software		5.44
	program		5.44
	freeware		5.44
	shareware		5.44
	information		8.17
	technology		8.17
	computador	Español	3.55
	programa		3.55
	informacion		3.55
	tecnologia		3.55
	software		3.55
	hardware		3.55
	freeware		5.33
	shareware		5.33
Educación (Education)	education	Inglés	3.58
	school		3.58
	university		3.58
	academic		3.58
	college		3.58
	scholarship		3.58
	institute		5.37
	teach		5.37
	educacion	Español	3.29
	escuela		3.29
	universidad		3.29
	academia		3.29
	secundaria		3.29
	beca		3.29
	instituto		4.93
	profesor		4.93
Entretenimiento (Entertainment)	entertainment	Inglés	3.29

	magic		3.29
	game		3.29
	trivia		3.29
	movie		3.29
	music		3.29
	actor		4.93
	humor		4.93
	entretenimiento	Español	4.01
	magia		4.01
	juego		4.01
	pelicula		4.01
	cine		4.01
	musica		4.01
	actor		6.01
	parque		6.01
Negocios y Economía (Business & Economy)	business	Inglés	4.73
	economy		4.73
	retailer		4.73
	marketplace		4.73
	commerce		4.73
	finance		4.73
	sell		7.09
	offer		7.09
	economia	Español	5.44
	negocio		5.44
	vender		5.44
	comprar		5.44
	finanzas		5.44
	oferta		5.44
	servicio		8.17
	franquicia		8.17
Noticias y Medios (News & Media)	news	Inglés	5.87
	media		5.87
	magazine		5.87
	newspaper		5.87
	people		5.87
	radio		5.87
	television		8.80
	broadcast		8.80
	noticia	Español	4.53
	revista		4.53
	periodico		4.53
	periodismo		4.53

	medio		4.53
	radio		4.53
	television		6.80
	broadcast		6.80
Recreación y Deportes (Recreation & Sports)	sport	Inglés	3.58
	recreation		3.58
	auto		3.58
	travel		3.58
	outdoor		3.58
	hobbie		3.58
	pet		5.37
	map		5.37
	deporte	Español	3.58
	recreacion		3.58
	viaje		3.58
	puntaje		3.35
	campeonato		3.58
	torneo		3.58
	equipo		5.36
	liga		5.36
Referencia (Reference)	reference	Inglés	3.54
	phone		3.53
	number		3.54
	address		3.50
	quotation		3.54
	library		3.54
	dictionary		5.31
	encyclopedia		5.31
	postal	Español	4.63
	referencia		4.63
	direccion		4.49
	libreria		4.63
	biblioteca		4.63
	diccionario		4.63
	enciclopedia		6.94
	calendario		6.94
Regional (Regional)	country	Inglés	5.12
	govern		5.12
	region		5.12
	asia		5.12
	europe		5.12
	america		5.12
	state		7.68

	coast		7.68
	region	Español	4.92
	pais		4.92
	gobierno		4.92
	asia		4.92
	europa		4.92
	estado		4.92
	sierra		7.38
	america		7.38
Salud (Health)	health	Inglés	3.33
	disease		3.33
	drug		3.33
	fitness		3.33
	nutrition		3.33
	pharmacy		3.33
	condition		5.00
	medicine		5.00
	nutricion	Español	3.12
	medicamento		3.25
	salud		3.35
	enfermedad		3.35
	pastilla		3.35
	farmacia		3.35
	medicina		5.02
	doctor		5.02
Sociedad y Cultura (Society & Culture)	culture	Inglés	3.18
	myth		3.23
	society		3.36
	family		3.36
	folklore		3.36
	relationship		3.36
	love		4.22
	food		5.04
	folklore	Español	3.22
	mito		3.29
	sociedad		3.39
	cultura		3.39
	familia		3.39
	relacion		3.39
	pareja		4.99
	amor		5.09

RESULTADOS DE LAS PRUEBAS DE INTEGRACIÓN REALIZADAS

PRUEBA 1

					Sin Módulo de Clasificación				Con Módulo de Clasificación			
Número de Búsqueda	Búsqueda	Idioma	Perfil	Preferencias Seleccionadas	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)
1	Computers for Christmas	Inglés	Fijo	Computadoras e Internet (Computers & Internet)	3	7	43	0.05	7	7	100	0.05
2	Google Phone	Inglés	Fijo	Computadoras e Internet (Computers & Internet)	4	5	80	0.03	5	5	100	0.03
3	Construction Materials	Inglés	Fijo	Negocios y Economía (Business & Economy)	5	8	63	0.06	7	8	88	0.06
4	Android	Inglés	Fijo	Computadoras e Internet (Computers & Internet)	2	2	100	0.01	2	2	100	0.01
5	ipad	Inglés	Fijo	Computadoras e Internet (Computers & Internet)	7	8	88	0.06	7	8	88	0.06
6	justin bieber	Inglés	Fijo	Entretenimiento (Entertainment)	3	5	60	0.03	4	5	80	0.03
7	nicki minaj	Inglés	Fijo	Entretenimiento (Entertainment)	1	7	14	0.05	6	7	86	0.05
9	Mixer	Inglés	Fijo	Entretenimiento (Entertainment)	6	8	75	0.06	7	8	88	0.06
10	katy perry	Inglés	Fijo	Artes y Humanidades (Arts & Humanities)	5	7	71	0.05	6	7	86	0.05
11	Twitter	Inglés	Fijo	Noticias y Medios (News & Media)	4	4	100	0.02	4	4	100	0.02

Número de Búsqueda	Búsqueda	Idioma	Perfil	Preferencias Seleccionadas	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)
12	Gamezer	Inglés	Fijo	Recreación y Deportes (Recreation & Sports)	3	3	100	0.01	3	3	100	0.01
13	Facebook	Inglés	Fijo	Sociedad y Cultura (Society & Cultura)	2	4	50	0.02	4	4	100	0.02
14	swine flu	Inglés	Fijo	Salud (Health)	6	9	67	0.07	8	9	89	0.07
15	Wamu	Inglés	Fijo	Negocios y Economía (Business & Economy)	9	10	90	0.08	9	10	90	0.09
17	Mininova	Inglés	Fijo	Computadoras e Internet (Computers & Internet)	4	6	67	0.05	5	6	83	0.05
18	susan boyle	Inglés	Fijo	Entretenimiento (Entertainment)	7	8	88	0.06	7	8	88	0.06
19	slumdog millionaire	Inglés	Fijo	Sociedad y Cultura (Society & Cultura)	9	10	90	0.08	10	10	100	0.09
21	myspace layouts	Inglés	Fijo	Ciencia (Science)	6	7	91	0.05	6	6	100	0.05
23	national city bank	Inglés	Fijo	Negocios y Economía (Business & Economy)	3	5	60	0.03	4	5	80	0.04
24	funny image	Inglés	Fijo	Entretenimiento (Entertainment)	7	9	78	0.07	8	9	89	0.08
76	derechos humanos	Español	Fijo	Sociedad y Cultura (Society & Cultura)	5	7	71	0.05	7	7	100	0.05
77	Rafael Correa	Español	Adaptativo	Sociedad y Cultura (Society & Cultura)	8	9	89	0.08	8	9	89	0.09
78	catálogo en linea	Español	Adaptativo	Referencia (Reference)	3	5	60	0.03	4	5	80	0.03

Número de Búsqueda	Búsqueda	Idioma	Perfil	Preferencias Seleccionadas	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)
79	buscar amigos	Español	Adaptativo	Entretenimiento (Entertainment)	7	8	88	0.06	7	8	88	0.06
80	universidad san francisco de quito	Español	Adaptativo	Educación (Education)	5	6	83	0.04	5	6	83	0.04
81	calentamiento global	Español	Adaptativo	Ciencia (Science)	3	5	60	0.03	4	5	80	0.03
82	via lacteal	Español	Adaptativo	Ciencia (Science)	2	7	29	0.05	6	7	86	0.06
83	clasificación	Español	Adaptativo	Educación (Education)	4	5	80	0.03	4	5	80	0.03
84	posicionamiento web	Español	Adaptativo	Computadoras e Internet (Computers & Internet)	6	8	75	0.06	7	8	88	0.06
86	desfile de modas	Español	Adaptativo	Entretenimiento (Entertainment)	3	5	60	0.03	4	5	80	0.03
87	celulares	Español	Adaptativo	Sociedad y Cultura (Society & Cultura)	6	7	86	0.05	6	7	86	0.05
88	ropa para perros	Español	Adaptativo	Recreación y Deportes (Recreation & Sports)	2	4	50	0.02	4	4	100	0.02
89	muñeco de nieve	Español	Adaptativo	Entretenimiento (Entertainment)	5	7	71	0.05	6	7	86	0.05
91	Geografia	Español	Adaptativo	Ciencias Sociales (Social Science)	6	7	86	0.05	6	7	86	0.05
93	deportes extremos	Español	Adaptativo	Recreación y Deportes (Recreation & Sports)	8	9	89	0.08	9	9	100	0.08
94	conjunto habitacional	Español	Adaptativo	Negocios y Economia (Business & Economy)	5	6	83	0.04	5	6	83	0.04

Número de Búsqueda	Búsqueda	Idioma	Perfil	Preferencias Seleccionadas	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)
96	Vitaminas	Español	Adaptativo	Salud (Health)	9	10	90	0.08	9	10	90	0.09
97	computadoras portatiles	Español	Adaptativo	Computadoras e Internet (Computers & Internet)	2	3	67	0.02	3	3	100	0.02
98	cédula de identidad	Español	Adaptativo	Regional (Regional)	2	2	100	0.01	2	2	100	0.01
99	cursos vacacionales	Español	Adaptativo	Sociedad y Cultura (Society & Cultura)	4	5	80	0.03	4	5	80	0.03
100	Mascotas	Español	Adaptativo	Recreación y Deportes (Recreation & Sports)	9	10	90	0.08	10	10	100	0.09
Promedio							84	0.0461905			90	0.0480952

PRUEBA 2

					Sin Módulo de Clasificación				Con Módulo de Clasificación			
Número de Búsqueda	**Búsqueda**	**Idioma**	**Perfil**	**Preferencias Seleccionadas**	**Resultados Relevantes**	**Resultados Totales**	**Exactitud (%)**	**Velocidad (s)**	**Resultados Relevantes**	**Resultados Totales**	**Exactitud (%)**	**Velocidad (s)**
8	Friv	Inglés	Fijo	Entretenimiento (Entertainment), Recreación y Deportes (Recreation & Sports)	1	4	25	0.02	4	4	100	0.02
20	circuit city	Inglés	Fijo	Negocios y Economía (Business & Economy), Computadoras e Internet (Computers & Internet)	8	8	97	0.07	8	9	90	0.07
22	michael jackson	Inglés	Fijo	Entretenimiento (Entertainment), Noticias y Medios (News & Media)	8	10	80	0.08	9	10	90	0.09
85	cumpleañ os	Español	Adaptativo	Recreación y Deportes (Recreation & Sports), Entretenimiento (Entertainment)	5	6	83	0.04	6	6	100	0.04
90	Islas Galapagos	Español	Adaptativo	Regional (Regional), Noticias y Medios (News & Media)	7	8	88	0.05	7	8	88	0.06
92	Patagonia Argentina	Español	Adaptativo	Regional (Regional), Noticias y Medios (News & Media)	6	8	75	0.05	7	8	88	0.06
Promedio							74	0.05			92	0.0542857

PRUEBA 3

					Sin Módulo de Clasificación				Con Módulo de Clasificación			
Número de Búsqueda	Búsqueda	Idioma	Perfil	Preferencias Seleccionadas	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)
25	Keywords	Inglés	Fijo	Artes y Humanidades (Arts & Humanities)	8	9	89	0.07	9	9	100	0.08
26	jobs in computer science	Inglés	Fijo	Negocios y Economía (Business & Economy)	4	6	67	0.04	5	6	83	0.04
27	Amy Winehouse Dead	Inglés	Adaptativo	Computadoras e Internet (Computers & Internet)	7	10	70	0.08	9	10	90	0.09
28	blood bank	Inglés	Adaptativo	Educación (Education)	5	6	83	0.04	6	6	100	0.04
29	trade marks	Inglés	Adaptativo	Entretenimiento (Entertainment)	8	9	89	0.07	8	9	89	0.08
30	Paintball	Inglés	Adaptativo	Salud (Health)	5	7	71	0.05	6	7	86	0.05
31	ipod touch	Inglés	Adaptativo	Noticias y Medios (News & Media)	4	5	80	0.03	5	5	100	0.03
32	electric power	Inglés	Adaptativo	Recreación y Deportes (Recreation & Sports)	8	9	89	0.07	8	9	89	0.07
33	Crossfit	Inglés	Adaptativo	Referencia (Reference)	3	4	75	0.02	3	4	75	0.02
34	the view	Inglés	Adaptativo	Regional (Regional)	6	7	86	0.05	6	7	86	0.05
35	september 11	Inglés	Adaptativo	Ciencia (Science)	6	8	75	0.06	7	8	88	0.06
36	drop dead diva	Inglés	Adaptativo	Ciencias Sociales (Social Science)	5	7	71	0.05	6	7	86	0.05
37	droid bionic	Inglés	Adaptativo	Sociedad y Cultura (Society & Cultura)	7	8	88	0.06	8	8	100	0.06
38	Cspan	Inglés	Adaptativo	Artes y Humanidades (Arts & Humanities)	7	7	94	0.05	7	7	96	0.05
39	stock market today	Inglés	Adaptativo	Negocios y Economía (Business & Economy)	6	9	67	0.07	8	9	89	0.07
40	columbus dispatch	Inglés	Adaptativo	Computadoras e Internet	8	10	80	0.08	9	10	90	0.09

Número de Búsqueda	Búsqueda	Idioma	Perfil	Preferencias Seleccionadas	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)
42	weather boston	Inglés	Adaptativo	Entretenimiento (Entertainment)	1	4	25	0.02	3	4	75	0.02
43	immigration news	Inglés	Adaptativo	Salud (Health)	9	10	90	0.08	9	10	90	0.09
44	rescue me	Inglés	Adaptativo	Noticias y Medios (News & Media)	6	7	86	0.05	6	7	86	0.05
45	the town	Inglés	Adaptativo	Recreación y Deportes (Recreation & Sports)	2	3	67	0.01	2	3	67	0.02
46	dachshund	Inglés	Adaptativo	Referencia (Reference)	3	4	75	0.02	3	4	75	0.02
47	americas got talent	Inglés	Adaptativo	Regional (Regional)	4	6	67	0.04	5	6	83	0.04
48	dark knight rises	Inglés	Adaptativo	Ciencia (Science)	8	9	89	0.07	8	9	89	0.08
49	all star game	Inglés	Adaptativo	Ciencias Sociales (Social Science)	9	10	90	0.08	9	10	90	0.09
50	a dance with dragons	Inglés	Adaptativo	Sociedad y Cultura (Society & Cultura)	4	7	57	0.05	6	7	86	0.05
55	Gloria Trevi	Español	Fijo	Entretenimiento (Entertainment)	7	9	78	0.07	8	9	89	0.07
56	Finanzas	Español	Fijo	Salud (Health)	2	4	50	0.02	3	4	75	0.02
57	recetas con papa	Español	Fijo	Noticias y Medios (News & Media)	7	8	88	0.06	7	8	88	0.06
58	trabajo desde casa	Español	Fijo	Recreación y Deportes (Recreation & Sports)	9	10	90	0.08	9	10	90	0.09
59	pastel de chocolate	Español	Fijo	Referencia (Reference)	1	3	33	0.01	3	3	100	0.01
60	negocios mas exitosos	Español	Fijo	Regional (Regional)	3	5	60	0.03	4	5	80	0.03
61	adolescentes	Español	Fijo	Ciencia (Science)	2	3	67	0.01	3	3	100	0.01
62	Perros	Español	Fijo	Ciencias Sociales (Social Science)	7	8	88	0.06	7	8	88	0.06

Número de Búsqueda	Búsqueda	Idioma	Perfil	Preferencias Seleccionadas	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)
64	seguidor de línea	Español	Fijo	Artes y Humanidades (Arts & Humanities)	9	10	90	0.08	10	10	100	0.09
65	ingeniería mecánica	Español	Fijo	Negocios y Economía (Business & Economy)	6	7	86	0.05	6	7	86	0.05
66	Manzana	Español	Fijo	Computadoras e Internet (Computers & Internet)	9	10	90	0.08	9	10	90	0.09
67	Mercado	Español	Fijo	Educación (Education)	7	9	78	0.07	9	9	100	0.07
68	consejos de marketing	Español	Fijo	Entretenimiento (Entertainment)	7	8	88	0.06	7	8	88	0.06
69	Obesidad	Español	Fijo	Salud (Health)	4	5	80	0.03	5	5	100	0.03
70	conexión a Internet	Español	Fijo	Noticias y Medios (News & Media)	7	8	88	0.06	8	8	100	0.06
71	Ecuador	Español	Fijo	Recreación y Deportes (Recreation & Sports)	3	6	50	0.04	5	6	83	0.04
72	comida típica	Español	Fijo	Referencia (Reference)	5	8	63	0.06	7	8	88	0.06
73	Autos	Español	Fijo	Regional (Regional)	6	9	67	0.08	9	9	100	0.09
74	venta de autos	Español	Fijo	Ciencia (Science)	6	8	75	0.06	7	8	88	0.06
75	mayoristas	Español	Fijo	Ciencias Sociales (Social Science)	2	5	40	0.03	4	5	80	0.03
Promedio							74	0.0510638			89	0.0538298

PRUEBA 4

					Sin Módulo de Clasificación				Con Módulo de Clasificación			
Número de Búsqueda	Búsqueda	Idioma	Perfil	Preferencias Seleccionadas	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)	Resultados Relevantes	Resultados Totales	Exactitud (%)	Velocidad (s)
51	claustrofobia	Español	Fijo	Artes y Humanidades (Arts & Humanities), Entretenimiento (Entertainment), Noticias y Medios (News & Media)	7	8	88	0.06	8	8	100	0.07
52	miss universo	Español	Fijo	Negocios y Economía (Business & Economy), Computadoras e Internet (Computers & Internet)	4	6	67	0.04	6	6	100	0.04
53	Lucero	Español	Fijo	Computadoras e Internet (Computers & Internet), Artes y Humanidades (Arts & Humanities)	4	7	57	0.05	6	7	86	0.05
54	explorando Marte	Español	Fijo	Educación (Education), Salud (Health)	9	10	90	0.08	9	10	90	0.09
Promedio							75	0.0575			94	0.0625

REFERENCIAS

[1] Grasso, M. 2006 Data Mining UTN-FRRo-SG2.

[2]Monografías.com 2009 Principios de Data Mining http://www.monografias.com/trabajos26/data-mining/data-mining.shtml.

[3] Sinnexus 2009 Datamining http://www.sinnexus.com/business_intelligence/datamining.aspx.

[4]Monografías.com 2009 Descubriendo Información Oculta http://www.monografias.com/trabajos/datamining/datamining.shtml.

[5] De Gyves, F. 2010 Web Mining: Fundamentos Básicos Universidad de Salamanca.

[6] Toolan, F. 2010 Web Mining Intelligent Information Retrieval Group University College Dublin.

[7] Arias, A. and Ovalle, D. 2010 Web Usage Mining: Revisión del Estado del Arte Escuela de Sistemas, Facultad de Minas Universidad Nacional de Colombia, Sede Medellín.

[8] Ajoudanian, S. and Davarpanah M. 2009 Deep Web Content Mining World Academy of Science, Engineering and Technology 49.

[9] Liu, B. 2010 Web Content Mining Department of Computer Science University of Illinois at Chicago http://www.cs.uic.edu/~liub.

[10] Esteban, M. 2007 Web mining y obtención de información para la generación de inteligencia Universidad de Zaragoza.

[11] Web-datamining.net 2010 Web Content Mining - Mining Text http://www.web-datamining.net/content/.

[12] Fivelstad, I. 2010 Web Structure Mining http://steinbit.agder-ikt.hia.no/~iisfel02/.

[13] Srivastava, J., Cooley, R., Deshpande, M. and Ning Tan, P. 2010 Web Usage Mining: Discovery and applications of usage patterns from web data Department of computer Science and Engineering University of Minnesota.

[14] Grcar, M., 2009 Web usage Mining Department of Knowledge technologies Jozef Stefan Institute Slovenia.

[15] Abraham, A. 2009 Business Intelligence from Web Usage Mining Department of Computer Science, Oklahoma State University.

[16] Zhou, X., Li, Y., Bruza, P., Tang Wu, S. and Xu, Y. 2009 Using Information Filtering in web data mining process Faculty of Information Technology Brisbane, Australia.

[17] Han, J. and Kamber, M. 2006. Data Mining – Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, 2nd edition.

[18] Unmexicanoenjapon.com 2009 http://tecnologia.unmexicanoenjapon.com/2010/11/17/tf-idf-pesando-las-palabras/#more-81

[19] Carrera, V. and García, M. 2008 Un algoritmo simple y eficiente para la Clasificación Automática de páginas web http://www.evcarrera.net/papers/andescon08.pdf

[20] Optiz, D. and Maclin, R. 1999 Popular ensemble methods: an empirical study. http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume11/optitz99a-html/node1.html

[21] Wikipedia 2009 http://es.wikipedia.org/wiki/Clasificadores_(matem%C3%A1tico)

[22] Buenastareas.com 2009 http://www.buenastareas.com/ensayos/Clasificador-Bayesiano/305157.html.

[23] Wikipedia 2009 http://es.wikipedia.org/wiki/Red_bayesiana.

[24] Statsoft 2010 Naïve Bayes Classifier http://www.statsoft.com/textbook/naive-bayes-classifier/

[25] Langley, P., Iba, W. and Thompson, K. 1992 An Analysis of Bayesian Classifiers AI Research Branch Moffet Field CA, USA.

[26] Viswanathan, K. 2010 Naïve Bayes classifier in 50 lines http://ebiquity.umbc.edu/blogger/2010/12/07/naive-bayes-classifier-in-50-lines/

[27] Larranaga, P., Inza, I. and Moujahid, A. 2010 Clasificadores Bayesianos Departamento de Ciencias de la Computación e Inteligencia Artificial Universidad del País Vasco.

[28] Rojas, R. 1996 Neural Networks Springer-Verlag Berlin.

[29] Red Neuronal Backpropagation – Presentation Transcript

[30] TREC 1999 Redes Neurales Artificiales http://electronica.com.mx/neural/informacion/backpropagation.html