Universidad de las Ciencias Informáticas

Facultad 10

Título: “Colección de Entrenamiento DES: Clasificando en diferentes temáticas los contenidos de Internet”

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas.

Autores: Yaritza Díaz Oliva.

Yanet Torres Mendoza.

Tutor: Ing. Alain Guerrero Enamorado.

Ciudad de la Habana, Cuba, 2008.

“Año 50 de la Revolución”.

DECLARACIÓN DE AUTORÍA

Declaro que soy el único autor de este trabajo y autorizo a la Facultad 10 de la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los____ días del mes de_____ del año _____.

_____ _____

Yanet Torres Mendoza
Autora

Yaritza Díaz Oliva
Autora

_____

Alain Guerrero Enamorado
Tutor

"(...) El camino de la vida puede ser libre y bello; pero hemos perdido el camino. La avaricia ha envenenado las almas de los hombres, ha levantado en el mundo barricadas de odio, nos ha llevado al paso de la oca a la miseria y a la matanza. Hemos aumentado la velocidad. Pero nos hemos encerrado nosotros mismos dentro de ella. La maquinaria, que proporciona abundancia, nos ha dejado en la indigencia. Nuestra ciencia nos ha hecho cínicos; nuestra inteligencia, duros y faltos de sentimientos. Pensamos demasiado y sentimos demasiado poco. Más que maquinaria, necesitamos humanidad. Más que inteligencia, necesitamos amabilidad y cortesía. Sin estas cualidades, la vida será violenta y todo se perderá. (...) Luchemos por un mundo de la razón, un mundo en el que la ciencia y el progreso lleven la felicidad a todos nosotros (...)"

Charles Chaplin

Agradecimientos

Quisiera agradecerles a mis padres y a mis hermanas que siempre me hayan apoyado, han sido mi guía, mi fuerza, mis deseos de continuar adelante y la razón por la que he podido llegar a ser una profesional. A mi novio y a mis suegros, gracias por todo el apoyo y el amor que me han dado. Tinito y Rubén, han sido los hermanos varones que nunca tuve la dicha de tener, no se que hubiera sido de mi sin ustedes, a Lisandra, Raycel, Yurima, Dayana, Yubi, Daymara me siento dichosa por tener amigas como ustedes, gracias por estos 5 años tan lindos que pasamos juntas. A mis primas Yusmaris y Rosa María, gracias por todo su apoyo. Gisela a ti decirte que no tengo palabras para agradecerte el papel de madre que asumiste conmigo. A la profesora Vilma, le agradezco muchísimo todo lo que ha hecho por mi, pues quedan en el mundo muy pocas personas con tan buen corazón como usted, nunca sabré como pagarle. A todas aquellas personas que de una forma u otra me han ayudado para que yo llegara a ser una profesional. Y alguien que no puede faltar es mi compañera de tesis que sin su ayuda nunca hubiese logrado llegar hasta aquí, gracias Yari, y fue un honor hacer la tesis contigo. **Yanet Torres Mendoza**

En estos años de experiencia universitaria ha habido personas a quien le quiero dar las gracias. A mi familia gracias por la paciencia que tantas veces he necesitado, por velar por mí. Mama gracias por respetar mis decisiones, si te agradeciera todo lo que haces por mí no tendría fin. Papi gracias por haber sido un ejemplo de fortaleza y valor, por ser uno de los principales precursores de este logro. Mi hermana del alma gracias a que estás, siento la responsabilidad de ser mejor para darte el ejemplo, nunca dudes lo mucho que te quiero. Tata gracias por estar ahí para mi, muchas veces tenemos desacuerdos, pero ¿Quién no los tiene? Mis sobrinos Leo y Lore gracias por los momentos de dicha junto a ustedes, tan pequeños y no imaginan la alegría que me brindan. Mi negra (Isnayca) gracias por alentarme, por permitirme apoyarme en tu hombro, por abrir paso a mis decisiones, por estar ahí y permanecer, eres la hermana que escogí. Maikel, amigo mío, gracias por apoyarme, por aconsejarme, por consolarme, gracias nene por estar ahí y sobre todo porque todavía puedo contar contigo. A Daviel y su familia gracias, por su apoyo en estos meses, a ti papi gracias por los momentos de dicha y felicidad, que esta relación sirva para demostrarnos que hay cosas que pueden cambiar. A mi padre gracias, por darme la vida. A Yane gracias por ser mi compañera de tesis, la decisión de realizar este trabajo juntas fue perfecta, hagámoslo de nuevo. A todos los que una forma u otra cooperaron con la realización de este trabajo, gracias. **Yaritza Díaz Oliva**

A nuestro tutor darle las gracias por dedicarnos sus horas extras a pesar de la carga de trabajo que lleva.

Dedicatoria

Dedico este trabajo de diploma a mis padres y hermanas que han sido mi apoyo, mi fuerza, mi razón de ser y el motivo por el cual yo he llegado ha ser lo que soy hoy...

Yanet Torres Mendoza

Dedicada en memoria de quien me acogió como hija, y no escatimó en brindarme su amor de padre. No estas aquí hoy, para verme cumplir la más grande de mis metas, pero se que donde quiera que estés, estás orgulloso de mi. **A ti papi**. A quien me ha heredado el tesoro más valioso que se le puede dar a un hijo: el amor. **A ti madre querida**. A quienes sacrificaron parte de su vida para formarme y educarme, a quienes la ilusión de su vida fue verme convertida en ingeniera. **A Luisa y Leonardo**. Sientan que este triunfo también es de ustedes, su apoyo fue la fuerza que me ayudó a conseguirlo. Los amo.

Yaritza Díaz Oliva

RESUMEN

En la actualidad uno de los problemas más notables es el acceso a contenidos inapropiados por parte de los usuarios de diferentes edades en Internet. Una de las soluciones con mayor potencial para regular el acceso a información web, es la categorización automática de las páginas web, su objetivo es encontrar la categoría a la que pertenece un determinado documento dentro de un conjunto de categorías definidas, necesitando para la etapa de aprendizaje del clasificador, la colección de entrenamiento, con la cual el clasificador aprende las características que debe poseer el documento para pertenecer a una determinada categoría. En el presente trabajo se realiza un estudio de diferentes colecciones de entrenamiento de dominio público así como de colecciones creadas en diferentes trabajos para identificar las características que debe poseer una colección de entrenamiento, valorando la influencia de dichas características para desarrollar una colección que influya en la disminución del margen de error a la hora de clasificar texto partiendo de documentos HTML. Antes de ser aplicado cualquier algoritmo de clasificación los documentos deben ser transformados de su formato inicial a una forma más adecuada al algoritmo que se vaya aplicar, por ello en la investigación se realiza un estudio de los modelos de representación más usados en la categorización de documentos, para así identificar cual sería el más adecuado para ser aplicado a los documentos que conforman la colección.

Palabras claves: categorización de documentos, colección de entrenamiento, representación de documentos.

ÍNDICE

INTRODUCCIÓN

El hombre desde tiempos remotos ha transformado información , transmitido señales, para vencer la distancia y así contactar con sus semejantes, satisfaciendo una de sus mayores necesidades: la comunicación.

Puesto que el tesoro más valioso de la raza humana es el conocimiento, es decir, la información, hoy más que nunca el hombre necesita:

- Intercambiar ideas
- Compartir conocimientos
- Relacionarse
- Buscar información

Internet ha revolucionado el mundo de las comunicaciones y la informática, es una oportunidad de difusión mundial, un mecanismo de propagación de la información, un medio de colaboración e interacción entre las personas y sus ordenadores independientemente de su localización geográfica. La cantidad de información que se genera a cada segundo en la red es incalculable. Internet crece, su desarrollo está alcanzando un ritmo desmesurado, cada día se publican en la red miles de documentos nuevos, y se conectan por primera vez miles de personas, la información ha empezado a jugar un rol fundamental en el desarrollo de la sociedad moderna, se está viviendo en una era, donde la información rige al mundo, la necesidad de estar conectado a Internet se esta haciendo inevitable, y el contenido de las páginas web se hace más difícil de controlar.

Internet y su implantación en la sociedad (la llamada Sociedad de la Información) aporta valiosos beneficios para los usuarios, posibilitando nuevas formas de comunicación, trabajo y educación. No obstante, su difícil control conlleva riesgos importantes para su aprovechamiento. Uno de los problemas más notables en la actualidad es el acceso a contenidos inadecuados (dígase pornográficos, relacionados con la droga, entre otros) por parte de los usuarios. Por una parte, existe la posibilidad de que los usuarios accedan a contenidos de Internet que son incapaces de juzgar correctamente, incluyendo los que promueven la violencia, el racismo y la adhesión a sectas, por otra parte que bajo completo

conocimiento accedan a páginas con contenido inapropiado. En ambos casos, es adecuada la adopción de sistemas de filtrado para regular el acceso a contenidos inadecuados.

La Universidad de las Ciencias informáticas (UCI), surgida en Cuba en el año 2002 como parte de la batalla de ideas que librara el pueblo, está inmersa en el desarrollo informático, dicha institución consta con un gran número de usuarios que tienen acceso a Internet, por tanto como parte de la incalculable información que surge en la red, la universidad no queda exenta del creciente problema del acceso a contenidos inadecuados por parte de la comunidad de usuarios.

La World Wide Web es un sistema de información global que ofrece la capacidad de publicar contenidos libremente, estableciendo un nuevo modelo de interacción entre individuos, viéndose un cambio en la forma de acceder al conocimiento. Dado el enorme tamaño y crecimiento que experimenta la Web, la aplicación de técnicas de clasificación automática a documentos HTML resulta útil para facilitar y mejorar la regulación del acceso a contenidos de Internet.

La clasificación automática de páginas web es considerada una de las soluciones con mayor potencial para regular el acceso a información web, el objetivo de la misma es encontrar dentro de un conjunto de categorías definidas, cual es la clase (categoría) a la que pertenece un determinado documento, comparando para ello el documento a clasificar con un conjunto de documentos clasificados manualmente de antemano (*colección de entrenamiento*).

Debido al carácter dinámico que tiene la Web se hace difícil categorizar manualmente el contenido de los documentos HTML (páginas Web) visitados, por ello existe en la Universidad de las Ciencias Informáticas el proyecto de Filtrado de Paquetes (FILPACON) que pretende ser una solución informática que permita la regulación de los contenidos de Internet utilizando listas de URLs previamente categorizadas. Por tal motivo el producto de dicho proyecto depende, en gran medida, de una base de datos constantemente actualizada que le permita tomar decisiones correctas respecto a los contenidos a regular. Para ello se hace necesario un Motor de Categorización Automática (MCA) que al utilizar algoritmos inteligentes necesita de una **colección de entrenamiento** diseñada para entrenar y validar

el aprendizaje de los algoritmos de manera que se logre disminuir el margen de error a la hora de categorizar textos en documentos HTML.

El presente trabajo trata de perfeccionar la regulación de los contenidos de Internet del proyecto FILPACON por lo que se ha planteado el siguiente **problema científico**, ¿Cómo lograr un pequeño margen de error a la hora de clasificar texto partiendo de documentos HTML?

El **objeto de estudio** es la clasificación automática de documentos HTML. El **campo de acción** abarca las colecciones de entrenamiento para categorizar textos en documentos HTML.

El **objetivo general** de la investigación es

- Proponer una colección de entrenamiento que cubra las categorías: deporte, salud y educación, en el idioma español.

Como **objetivos específicos** se tienen:

1. Sistematizar y estudiar los antecedentes de los tipos de colecciones de entrenamiento que se usan en la categorización de textos en documentos HTML.
2. Determinar cuales resultan ser las categorías más representativas para el proyecto FILPACON.
3. Sistematizar y estudiar los antecedentes de los tipos de Modelos de Representación Textual de documentos que se usan en la categorización de textos para documentos HTML.
4. Identificar que características de las colecciones de entrenamiento influyen en la disminución del margen de error de un categorizador de texto.
5. Identificar que modelo de representación textual posee las características idóneas para ser aplicado sobre la colección creada.

De acuerdo al problema y con la intención de alcanzar el objetivo propuesto, se plantean las siguientes **Preguntas Científicas**:

1. ¿Cuáles resultan ser los antecedentes de las colecciones de entrenamiento que existen para categorizar textos?
2. ¿Cuáles resultan ser los antecedentes de los modelos de representación textual que se usan para categorizar texto?
3. ¿Cómo crear una colección de entrenamiento que cumpla con las características adecuadas para reducir el margen de error en la categorización de textos?

Para alcanzar los objetivos enunciados anteriormente se definen las siguientes **tareas:**

- Sistematizar y estudiar las colecciones de entrenamiento para categorizar textos en documentos HTML.
- Identificar las categorías más representativas del proyecto FILPACON.
- Sistematizar y estudiar los modelos de representación textual para categorizar textos en documentos HTML.
- Identificar las características de las colecciones de entrenamiento.
- Crear la colección de entrenamiento.
- Identificar el modelo de representación textual para ser aplicado a la colección creada.

Para el desarrollo de esta investigación se utilizaron los **métodos** planteados a continuación:

Analítico – sintético: para distinguir los elementos relacionados con las colecciones de entrenamiento y proceder a la amplia y profunda investigación para la obtención de una colección de entrenamiento que logre disminuir el margen de error a la hora de categorizar los textos en los documentos HTML, analizando las categorías: educación, salud y deporte.

Histórico – lógico: para conocer los antecedentes y tendencias más actuales de las colecciones de entrenamiento a nivel mundial.

Observación: para ver y constatar los problemas que existen en el proyecto FILPACON por la falta de una colección de entrenamiento previamente analizada y estudiada para la categorización de los textos en los documentos HTML, precisamente tras estos problemas es que surge la idea de realizar la investigación.

La variante que se propone en la investigación constituye una alternativa para mejorar la calidad en el proceso de clasificación de documentos y garantizar reducir el margen de error a la hora de categorizar textos en documentos HTML en el proyecto FILPACON, regulando así el acceso a contenidos inadecuados por parte de la comunidad de usuarios de la UCI.

El presente trabajo está estructurado de la siguiente manera:

Capítulo 1: Las colecciones de entrenamiento en este capítulo se analizan aspectos relacionados con las colecciones de entrenamiento, conceptos de las colecciones de entrenamiento, tipos de colecciones, tipos de categorías en las que es posible clasificar un documento. En un primer momento se analiza el marco teórico en el que se enmarcan las colecciones de entrenamiento.

Capitulo 2: Modelos de representación textual de documentos en este capítulo se analizan aspectos relacionados con los modelos de representación textual, (primer paso a la hora de categorizar un documento), se establecen las características fundamentales de los modelos de representación de textos más usados en tareas de categorización de textos. Se muestran funciones empleadas para calcular la importancia, o relevancia, de un rasgo en el contenido de un texto.

Capitulo 3: Creación de la Colección de entrenamiento en este capítulo se presentan las características fundamentales de las colecciones de entrenamiento así como el método de representación de documentos más eficaz para ser aplicado a la colección que se confeccione, ayudando esto a la elaboración de una eficiente colección a utilizar por el proyecto FILPACON que minimice el margen de error a la hora de categorizar textos en documentos HTML.

Capítulo 1 COLECCIONES DE ENTRENAMIENTO

INTRODUCCIÓN

La implantación progresiva de Internet esta marcando una nueva era: la llamada era de la información, donde cada día se hace más inevitable el acceso a la web. Regular los diferentes contenidos a los que puede acceder una comunidad de usuarios se esta haciendo una necesidad cada vez mayor, los volúmenes de información en forma electrónica están experimentando un enorme crecimiento, y son disímiles los autores que publican en la red de redes diversos temas ya sean educativos e instructivos y otros que más bien son lesivos para la sociedad como temas pornográficos, violentos, etcétera.

Para un mejor desarrollo de la regulación del acceso a contenidos inadecuados se plantea la clasificación de texto, haciéndose necesario para ello utilizar un algoritmo de categorización automática que regule el contenido de los documentos.

Se pretende en el proyecto FILPACON que radica en la Universidad de la Ciencias Informáticas regular el acceso a contenidos inapropiados por parte de la comunidad UCI, clasificando en diferentes temáticas las páginas visitadas para así conocer si un determinado contenido es inadecuado o no, mediante un motor de categorización automática que necesita de una *colección de entrenamiento* que valide el aprendizaje automático de los algoritmos inteligentes.

En este capítulo se fija en un primer momento el marco teórico sobre el que se enmarcan las colecciones de entrenamiento, se trata brevemente el aprendizaje automático así como los algoritmos de aprendizaje, seguido se aborda lo referente a la clasificación de documentos, para a continuación presentar los aspectos relacionados con los documentos de entrenamiento, definiciones, detalles de algunas colecciones de entrenamiento. Seguidamente se introducen las categorías a clasificar los documentos por el proyecto FILPACON, así como las categorías o subcategorías que serán analizadas en el presente trabajo de diploma. Por último se muestran algunos trabajos en los que han sido realizadas algunas colecciones.

1.1 APRENDIZAJE AUTOMÁTICO

El aprendizaje automático fue concebido hace aproximadamente cuatro décadas con el objetivo de desarrollar métodos computacionales que implementaran varias formas de aprendizaje. El mismo es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender, es el procedimiento que permite desarrollar sistemas capaces de realizar una tarea de un modo automático, que mejore automáticamente mediante la experiencia.

Un programa ha aprendido a desarrollar una tarea si después de proporcionarle la experiencia el sistema es capaz de desempeñarse razonablemente bien cuando se presentan nuevas situaciones de dicha tarea. A pesar de que se desconoce como lograr que las computadoras aprendan al nivel de las personas, ciertos algoritmos propuestos en el campo han resultado efectivos en varias tareas de aprendizaje. Por ejemplo en la *clasificación*.

El enfoque del aprendizaje automatizado es similar al proceso de aprendizaje humano. Como mismo el hombre desarrolla gran parte de sus conocimientos mediante la lectura, una máquina aprende sobre diversas áreas o categorías a través de un conjunto de documentos preseleccionados (*conjunto de ensayo o de entrenamiento*) por especialistas en la materia. El aprendizaje automático consiste en la programación de sistemas informáticos de forma que se optimice un determinado criterio de rendimiento, empleando datos de entrenamiento o con ayuda de una experiencia pasada. En este proceso se define un modelo a partir de un conjunto de parámetros y, a través de un aprendizaje, se ejecuta un programa que optimiza dichos parámetros, usando un conjunto de datos que representan experiencia pasada. Todo esto con la idea de buscar métodos capaces de aumentar las capacidades de las aplicaciones habituales de manera que sean más flexibles y eficaces.

1.1.1 ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

Dependiendo de si se dispone o no de datos etiquetados, dentro del aprendizaje automático, se pueden distinguir dos tipos de algoritmos: El ***aprendizaje no supervisado*** y el ***aprendizaje supervisado***. Un ejemplo de este último algoritmo es el problema de

c*lasificación*, donde el sistema de aprendizaje trata de clasificar una serie de vectores utilizando una entre varias categorías (clases).

En el aprendizaje supervisado el algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. El aprendizaje supervisado se construye sobre ejemplos de categorías anteriores. Para cada ejemplo que tenemos conocemos el nombre de la clase a la que pertenece. La entrada a este tipo de aprendizaje es un conjunto de ejemplos clasificados. El resultado de este proceso es una representación de las clases que describen los ejemplos.

1.2 CLASIFICACIÓN

La ciencia de la clasificación es la taxonomía (del griego taxis o "ordenamiento", y nomos, "regla"). La clasificación es el ordenamiento por clases o categorías, según las propiedades del objeto o concepto en cuestión es la tarea de aproximar una función objetivo desconocida que describe las instancias del problema que deben ser clasificadas de acuerdo a un experto en el dominio, por medio de otra función llamada el clasificador, con un conjunto de categorías predefinida y un conjunto de instancias del problema. Cada instancia es representada como una lista de valores característicos conocidos como atributos.

Para generar automáticamente el clasificador es necesario un proceso inductivo, llamado el aprendiz, el cual para observar los atributos de un conjunto de instancias preclasificadas adquiere los atributos que una instancia no vista debe tener para pertenecer a la categoría. Por tal motivo, en la construcción del clasificador se requiere la disponibilidad inicial de una colección de ejemplos de valor conocido. A la colección se le llama *conjunto de entrenamiento*.

1.2.1 CLASIFICACIÓN DE DOCUMENTOS

La clasificación de documentos (categorización de texto o ubicación del tema) es el proceso por el cual se asocian una o más categorías a textos escritos en un lenguaje natural basándose tan sólo en su contenido. En la clasificación de documentos se distinguen tres casos:

Clasificación binaria. El clasificador devuelve una de entre dos categorías, o bien SI/NO.

Clasificación multi-clase. El clasificador proporciona una categoría de entre varias propuestas.

Clasificación multi-etiquetado. El clasificador puede proporcionar varias categorías de entre las categorías disponibles.

Aunque es posible construir de manera manual un categorizador, las técnicas estadísticas y automáticas son actualmente las preferidas puesto que no sólo ofrecen un rendimiento adecuado sino que resulta mucho más sencillo seleccionar un conjunto de ejemplos para entrenar un algoritmo que elaborar reglas manualmente. Esta tarea entra dentro del campo del *aprendizaje automático* y es posible aplicarle una gran variedad de técnicas disponibles.

1.2.2 CLASIFICACIÓN AUTOMÁTICA DE DOCUMENTOS

La clasificación automática de documentos es la tarea de clasificar automáticamente un conjunto de documentos en categorías (o temas) dentro de un grupo o categoría predefinido (Sebastiani, 1999). La dificultad de la clasificación automática es parte de un problema mayor de análisis automático de contenidos y ya en la década de los sesenta, se daban los primeros pasos en el campo. En ese entonces se planteaba saber si una máquina podría ser programada para determinar el tema del contenido de un documento y la categoría, o categorías, en las que debería ser clasificado. La construcción de los clasificadores se llevaba a cabo mediante un proceso manual que extraía el conocimiento del experto y lo representaba mediante reglas. Ya para la década de los 90 [Fabricio, 1999], con la amplia disponibilidad de documentación online, la clasificación automática de textos experimenta un creciente interés, que llega hasta nuestros días.

La clasificación automática es un proceso donde se construye automáticamente un clasificador por aprendizaje a partir de un conjunto de textos previamente clasificados en determinadas categorías la misma proporciona como principales ventajas (Sebastiani, 1999):

1. Una buena efectividad.

2. Una reducción del ser humano, ya que no se requiere un experto que defina las reglas de clasificación.
3. Independencia en el dominio de los textos a clasificar.

Actualmente la exactitud de los sistemas de clasificación automática de texto compite con la de profesionales humanos especializados. El estándar de la clasificación automática de documentos es construir y usar las llamadas máquinas de aprendizaje supervisado (clasificadores). Los clasificadores que no son más que los programas que ejecutan los algoritmos de clasificación, son entrenados con un grupo de documentos, previamente clasificados y etiquetados conformando una clase; para de esta manera, cumplir su objetivo que es decidir en qué categoría debe ir cada texto a clasificar.

La clasificación o categorización automática de documentos puede ser entendida como una tarea en la cual, en base a la identificación por medios matemático-estadísticos, un documento nuevo es asignado a una clase particular de documentos pre-existentes. El objetivo de la clasificación automática de textos es categorizar documentos dentro de un número fijo de categorías predefinidas en función de su contenido.

La construcción de un clasificador automático de texto comienza con la recopilación y clasificación manual de un conjunto de documentos (*documentos de entrenamiento*), después se llevan los documentos a una representación adecuada para finalmente aplicar distintos algoritmos de clasificación y obtener así el clasificador.

En resumen, en la realización de un proceso de categorización automático se consideran las siguientes tareas (Aas K, 1999), (Sebastiani, 1999):

Indexado. El objetivo es representar los documentos de texto en una forma adecuada de su contenido para su uso con el clasificador.

Reducción de dimensionalidad. Es común que el espacio de representación tenga una alta dimensionalidad. Así, con el propósito de evitar el sobreajuste en el proceso de aprendizaje e incrementar su eficiencia y efectividad, es necesario considerar sólo un subconjunto de los términos originales. La selección de los términos más

representativos es realizada por una función de características. Por ejemplo, la ganancia en la información y la estadística chi-cuadrada.

Aprendizaje. Durante el aprendizaje o entrenamiento de un clasificador se reúne y combina un número fijo de documentos de muestra para cada concepto, donde los documentos resultantes son preprocesados e indexados utilizando un algoritmo de clasificación.

Existen 3 componentes principales implicados en el proceso de categorización de texto (Kwan, 2006).

1. Los documentos textuales. Sea D= [d1, d2,…, dn] un conjunto de documentos.
2. Las categorías finales. Sea C= [c1, c2,..., cn] el conjunto de categorías a estudiar.
3. Un algoritmo de representación que actúa como clasificador. Un algoritmo de representación puede ser descrito como una función, tomando un documento como una entrada y produciendo una decisión binaria si el documento encaja en una categoría dada.

De manera general se entiende por clasificación automática de documentos: un conjunto de algoritmos, técnicas y sistemas capaces de asignar un documento a una o más clases o grupos de documentos, construidos según su afinidad temática, concibiéndose como un proceso de aprendizaje matemático-estadístico, durante el cual el algoritmo capta las características que deben poseer los documentos para pertenecer a una determinada categoría (**Anexo # 1**). Estas características no indican de forma absoluta la pertenencia a una categoría, más bien lo hacen en función de una escala o graduación.

1.2.2.1 TIPOS DE CLASIFICACIÓN AUTOMÁTICA

Dependiendo de si la clasificación es completamente automática o posee una parte manual se puede distinguir dos tipos de clasificación: la ***clasificación automática no supervisada*** o *clustering* (Chakrabarti, 2002) y la ***clasificación automática supervisada***. En la primera no hay categorías previas, los documentos se agrupan en función de ellos mismos, de su contenido. Este tipo de clasificación se conoce como no supervisada porque se efectúa de forma totalmente automática, sin supervisión o asistencia manual.

En el segundo caso se parte de una serie de clases o categorías prediseñadas, en la que la labor del clasificador (manual o automático) es asignar cada documento a la categoría que le corresponda. Esta clasificación no sólo se conoce como supervisada porque requiere la elaboración manual o intelectual del cuadro de categorías, sino también porque requiere un proceso de aprendizaje o *entrenamiento* por parte del clasificador, que debe ser supervisado manualmente en mayor o menor medida.

La categorización automática supervisada, requiere supervisión o intervención de personas, tanto para diseñar las clases o categorías como para enseñar o entrenar al sistema. Para esto último el sistema utiliza un conjunto de documentos de ejemplo para cada una de las categorías en las que se quiere clasificar (*colecciones de entrenamiento)*, de los cuales extrae el sistema las características de cada una de las clases. Después de una etapa de entrenamiento, el sistema queda ajustado de tal modo que ante nuevos ejemplos, el algoritmo es capaz de clasificarlos en alguna de las clases existentes. Cuanto mayor sea el conjunto de datos etiquetados mayor será la información potencial disponible y mejor resultara el aprendizaje.

1.2.2.2 ALGORITMOS DE CLASIFICACIÓN AUTOMÁTICA

El proceso inductivo necesario para generar automáticamente un clasificador es el algoritmo de clasificación, aunque la mayor parte de los algoritmos de clasificación han sido propuestos para clasificar todo tipo de cosas, algunos de ellos han sido utilizados para la clasificación de documentos.

Entre los más utilizados, se tienen:

- ***Algoritmos probabilísticos***
- ***Algoritmo de Rocchio***
- ***Algoritmo del vecino más próximo y variantes***
- ***Algoritmos basados en redes neuronales***
- ***Clasificador Máquinas de Vectores de Soporte (SVM)***

1.2.2.2.1 ALGORITMOS PROBABILÍSTICOS

Se basan en la teoría probabilística, en especial en el teorema de Bayes. Éste permite

estimar la probabilidad de un suceso a partir de la probabilidad de que ocurra otro suceso, del cual depende el primero. El algoritmo de este tipo más conocido, y también el más simple, es el denominado naive Bayes (Maron, 1961), Básicamente, se trata de estimar la probabilidad de que un documento pertenezca a una categoría. Dicha pertenencia depende de la posesión de una serie de características, de cada una de las cuales conocemos la probabilidad de que aparezcan en los documentos que pertenecen a la categoría en cuestión.

Las características son los términos que conforman los documentos, y tanto su probabilidad de aparición en general, como la probabilidad de que aparezcan en los documentos de una determinada categoría, pueden obtenerse a partir de los *documentos de entrenamiento*; para ello se utilizan las frecuencias de aparición en la *colección de entrenamiento*. Con las probabilidades obtenidas de la *colección de entrenamiento*, se puede estimar la probabilidad de que un nuevo documento, pertenezca a cada una de las categorías. La más probable, obviamente, es a la que será asignado. La implementación del naive Bayes es sencilla y rápida, y sus resultados son bastante buenos.

La idea principal en naive bayes es usar la teoría de probabilidad de Bayes para estimar las probabilidades de las palabras y categorías dado un documento, la parte ingenua de éste clasificador está en asumir la independencia de las palabras, es decir que la probabilidad condicional de una palabra dada una categoría se calcula asumiendo que dicha palabra es independiente de las probabilidades condicionales de otras palabras en el documento dada esa categoría.

1.2.2.2.2 ALGORITMO DE ROCCHIO

El llamado algoritmo de Rocchio (Rocchio, 1971) en el ámbito de la categorización, proporciona un sistema para construir los patrones de cada una de las clases del documentos. Partiendo de una *colección de entrenamiento*, categorizada manualmente de antemano, y aplicando el modelo vectorial, se pueden construir vectores patrón para cada una de las clases, considerando como ejemplos positivos los *documentos de entrenamiento* de esa categoría, y como ejemplos negativos los de las demás categorías.

El algoritmo de Rocchio es una forma de construir patrones de cada clase. Una vez que se

tienen los patrones de cada una de las clases, el proceso de entrenamiento o aprendizaje está concluido. Para categorizar nuevos documentos, simplemente se estima la similitud entre el nuevo documento y cada uno de los patrones. El que arroja un índice mayor indica la categoría a la que se debe asignar ese documento. El algoritmo de Rocchio ha sido utilizado en tareas de categorización con buenos resultados (Lewis, 1996).

1.2.2.2.3 ALGORITMO DEL VECINO MÁS PRÓXIMO

El algoritmo del vecino más próximo (Nearest Neighbour, NN) es uno de los más sencillos de implementar. La idea básica es: si se calcula la similitud entre el documento a clasificar y cada uno de los *documentos de entrenamiento*, el que de éstos, de más parecido, estará indicando a que categoría se debe asignar el documento que se desea clasificar. El algoritmo se basa en localizar el documento más similar o parecido al que se desea clasificar. Esto no es más que utilizar ese documento como si fuera una consulta sobre la *colección de entrenamiento*.

Una vez localizado el documento de entrenamiento más similar, dado que éstos han sido previamente categorizados manualmente, se sabe a qué categoría pertenece y, por ende, a qué categoría se debe asignar el documento que se está clasificando. Una de las variantes más conocidas de este algoritmo es la del *k-nearest neighbour* o *KNN*.

KNN parece especialmente eficaz cuando el número de categorías posibles es alto, y cuando los documentos son heterogéneos. La idea en este algoritmo es que dado un documento de prueba, el sistema encuentra los k-ésimos vecinos más cercanos entre los documentos de entrenamiento y usa las categorías a las que pertenecen los vecinos encontrados para proponer unas categorías candidatas al documento de prueba. El nivel de similaridad de cada documento vecino con respecto al documento de prueba es usado como el peso de las categorías de los documentos vecinos. Si varios de los k-ésimos vecinos comparten una categoría, entonces los pesos de cada uno de los vecinos es sumado y la suma ponderada resultante es usada como la probabilidad puntaje de esa categoría con respecto al documento de prueba. Finalmente una lista ordenada por puntajes de las categorías candidatas es generada para el documento de prueba.

1.2.2.2.4 ALGORITMOS BASADOS EN REDES NEURONALES

Las redes neuronales han sido propuestas en numerosas ocasiones como instrumentos útiles para la clasificación automática. Una de las principales aplicaciones de las redes neuronales es el reconocimiento de patrones. Básicamente, una red neuronal consta de varias capas de unidades de procesamiento o neuronas interconectadas; en el ámbito de la clasificación de documentos la capa de entrada recibe términos, mientras que las unidades o neuronas de la capa de salida mapea clases o categorías. Las interconexiones tienen pesos, es decir, un coeficiente que expresa la mayor o menor fuerza de la conexión. Es posible entrenar una red para que, dada una entrada determinada (los términos de un documento), produzca la salida deseada (la clase que corresponde a ese documento). El proceso de entrenamiento consta de un ajuste de los pesos de las interconexiones, a fin de que la salida sea la deseada.

Las técnicas basadas en redes neuronales usan conceptos de Inteligencia Artificial, un clasificador con redes neuronales, es una red de unidades en donde las unidades de entrada son los términos, las unidades de salida representan la categoría o categorías de interés y los pesos sobre los lados son las relaciones de dependencia. Dado un documento de prueba, sus términos peso son cargados dentro de las unidades de entrada, la activación de esas unidades es propagada a través de la red y el valor de las unidades de salida determina las decisiones de categorización.

1.2.2.2.5 MÁQUINAS DE VECTORES DE SOPORTE (SVM)

Los SVM son poderosos clasificadores, ampliamente utilizados en la clasificación de texto e imágenes. Dichos clasificadores construyen el Hiperplano de Separación Óptima, el cuál separa un conjunto de muestras positivas de un conjunto de muestras negativas maximizando el margen (la separación).

SVM permite construir clasificadores no lineales, pues representa datos de entrenamiento no lineales en un espacio de alta dimensionalidad (llamado espacio de características), y construye el hiperplano que tiene el margen máximo. Debido al uso de una función kernel para realizar el mapeo, es posible calcular el hiperplano sin representar explícitamente el espacio de características.

La idea básica de este algoritmo consiste en que los documentos son representados como puntos en un vector en el espacio, donde las dimensiones son las características seleccionadas, con base en ejemplos de entrenamiento, SVM encuentra el hiperplano único que minimiza el error de generalización esperado, lo que se logra maximizando la distancia más corta entre uno de los ejemplos de entrenamiento y el hiperplano. Los vectores de entrenamiento que definan la posición del hiperplano son llamados vectores soporte. Dicho hiperplano es el que mejor separa los datos en clases.

1.2.3 CLASIFICACIÓN AUTOMÁTICA DE PÁGINAS WEB

La clasificación automática de páginas web es una variedad de la clasificación automática de documentos, es una de las tareas más frecuentes de la minería de datos. Como toda tarea de clasificación requiere un proceso previo de aprendizaje que parte de la representación de los elementos a clasificar. Consiste en clasificar automáticamente el conjunto de documentos HTML en categorías dentro del conjunto de categorías predefinido.

Los primeros clasificadores aplicados a la categorización de páginas web fueron los Full-Text classifiers, clasificadores que se basaban en un análisis del contenido textual de los documentos HTML y no consideraban ninguna información extraída de la estructura del grafo de hipertexto. En estos casos, las funciones de ponderación empleadas no pasaban de ser funciones clásicas: binaria, TF y TF-IDF. Dentro de este tipo se encuentran los clasificadores bayesianos, así como el algoritmo de los k vecinos más cercanos.

Además de estos clasificadores basados en contenido se han propuesto muchos otros que exploran la estructura de hiperenlaces que forma la Web. Métodos que proponen el uso de los anchortexts para predecir la clase de los documentos apuntados por esos enlaces. Clasificadores que emplean métricas basadas directamente en la estructura de los hiperenlaces. En estos casos, las medidas se basan en un análisis de las correlaciones (co-citaciones) presentes en los documentos a clasificar.

1.2.4 CATEGORIZACIÓN AUTOMÁTICA BASADA EN APRENDIZAJE

El proceso consiste en la representación de un conjunto de documentos (en este caso, documentos en formato HTML) manualmente clasificados (llamado *colección de*

entrenamiento) como vectores de pesos, a los que se le aplica un algoritmo de aprendizaje de los abordado con anterioridad, que construye un modelo de clasificación o clasificador. Los documentos a clasificar se representan de igual forma, de modo que el clasificador es capaz de asignar una categoría a los mismos.

1.3 COLECCIÓN DE ENTRENAMIENTO

Existe gran variedad de algoritmos capaces de hacer clasificación supervisada, la idea básica es: conseguir construir un patrón representativo de cada una de las clases o categorías y aplicar alguna función que permita estimar el parecido o similitud entre el documento a clasificar y cada uno de los patrones de las categorías (He J, 2003). El patrón más parecido al documento es el que nos indica a qué clase debemos asignar ese documento. Para la construcción de los patrones de las categorías se utilizan documentos clasificados manualmente de antemano, que sirven como ejemplo. El proceso de formar esos patrones de cada clase a partir de esos documentos preclasificados se conoce como *entrenamiento* o *aprendizaje*; y la colección de documentos preclasificados utilizada se conoce como *colección de entrenamiento*.

Una colección de entrenamiento es un conjunto de documentos categorizados manualmente, que permite al sistema de clasificación asignar categorías a nuevos documentos, de acuerdo con su similitud a los documentos de la colección de entrenamiento. La misma es requerida para la construcción de los clasificadores de textos mediante métodos de aprendizaje automático.

1.3.1 TIPOS DE COLECCIONES DE ENTRENAMIENTO

Han sido desarrolladas con anterioridad varias colecciones de entrenamiento tanto de formato texto como de documentos HTML, algunas de ellas son de dominio público, otras tienen asociado un costo y restricciones para su uso, debido al creciente desarrollo que está alcanzando la clasificación de documentos. A continuación se detallan algunas colecciones ya existentes:

- **Semcor:**

La colección de textos Semcor (Miller, 1993) es un corpus etiquetado con conceptos de WordNet que se distribuye como suplemento de la propia WordNet (por ejemplo, para investigar o mostrar ejemplos del uso de conceptos). No existe otra colección etiquetada con información conceptual en tanto detalle. Semcor es un subconjunto del Brown Corpus y de la novela " The Red Badge of Courage", con alrededor de 250.000 palabras. En Semcor, cada palabra o expresión ha sido etiquetada con su concepto adecuado en WordNet.

Semcor no precisa ser adaptado para evaluar la categorización de texto. Contiene 352 fragmentos de textos obtenidos de diversas fuentes, cubriendo 15 géneros, tales como prensa, reportaje, religión, ficción, ciencia. SemCor es un lexicón, donde cada palabra en el texto hace referencia a su correcto significado en él. Es un corpus, en el que las palabras han sido etiquetadas sintáctica y semánticamente, en el que las frases de ejemplo pueden ser encontradas por varias definiciones.

- **Reuters-21578**:

La colección Reuters proporciona 5 categorías: intercambios (exchanges), orgs, personas (people), lugares (places) y tema (topics), aunque se suele emplear únicamente la categoría tema que contiene noticias en ingles de acuerdo con temáticas de carácter económico (indicadores económicos, monedas, bienes, etc.).

Esta colección se ha convertido en un estándar dentro del dominio de la categorización automática de documentos y es utilizada por numerosos autores de la materia. Los documentos en la colección son noticias reales que aparecieron en cables de la agencia Reuters durante 1987. Los documentos fueron recopilados y categorizados manualmente por personal de la agencia y de la compañía Carnegie Group, Inc., en 1987.

Actualmente es utilizada en los trabajos sobre categorización automática de documentos para asegurar una metodología de prueba uniforme. La colección se compone de 21.578 documentos (cantidad que le da nombre a la misma), distribuidos en 22 archivos. Cada documento tiene 5 campos de categorización distintos: Valor Bursátil, Organización, Persona, Lugar y Tema. En cada campo, el documento puede tener un sólo valor, varios, o ninguno.

- **BankSearch DataSet**

La colección BankSearch DataSet (Sinka, 2002) está formada por 11.000 páginas web escritas en ingles y preclasificadas manualmente en 11 categorías del mismo tamaño y organizadas en nivel jerárquicos. Fue recopilada y clasificada con el propósito de ser utilizada como una colección de referencia para evaluar clustering de páginas web.

Las categorías en esta colección se organizan del siguiente modo:

- Las tres primeras categorías pertenecen al tema general "Bancos & Finanzas" (Banking & Finance): subcolección ABC.
    - Bancos Comerciales (Commercial Banks, A)
    - Sociedades de Crédito Hipotecario (Building Societies, B)
    - Aseguradoras (Insurance Agencies, C)
- Las tres siguientes se corresponden con la temática "Lenguajes de programación" (Programming Language): subcolección DEF.
    - Java (D)
    - C/C++ (E)
    - Visual Basic (F)
- Las dos categorías siguientes están formadas por páginas relativas al tema "Ciencia" (Science): subcolección GH.
    - Astronomía (Astronomy, G)
    - Biología (Biology, H)
- Por ultimo, las dos últimas corresponden a "Deportes": subcolección IJ.
    - Fútbol (Soccer, I)
    - Deportes de Motor (Motor Sport, J)

Además de estas 10 categorías organizadas en una jerarquía de dos niveles, existe otra categoría (X) que constituye un superconjunto de la clase IJ y que supone añadir un nivel

superior en la jerarquía. Esta categoría X está formada por páginas relativas a diferentes deportes no incluidos ni en las categorías I ni J.

- **WebKB**

Esta colección fue creada para el estudio de la estructura de hiperenlaces que forman los documentos web. Su ámbito se reduce al entorno universitario. WebKB está formada por 8282 páginas clasificadas manualmente en 7 clases, descargadas de los departamentos de Computer Science de 4 universidades estadounidenses: Universidad de Cornell (Cornell University), Universidad de Texas en Austin (The University of Texas at Austin), Universidad de Washington (University of Washington), Universidad de Wisconsin-Madison (University of Wisconsin-Madison).

Categorías en las que se divide esta colección:

- Estudiante (Student), con 1641 páginas web.
- Facultad (Faculty), con 1124.
- Personal (Staff), con 137.
- Departamento (Department), con 182.
- Curso (Course), con 930.
- Proyecto (Project), con 504.
- Otros (Other), con 3764.

La clase otros supone un cajón desastre donde se encuentran aquellas páginas que no se sabia donde clasificar.

- **CCHMC**

Esta colección se trata de un corpus desarrollado por el centro de medicina computacional (The Computational Medicine Center), está formada por registros médicos anónimos recopilados en el departamento de radiología del Hospital infantil de Cincinnati (The Cincinnati Children's Hospital Medical Center's Department of Radiology – CCHMC).

La colección está formada por 978 documentos consistentes en informes radiológicos que están etiquetados con códigos del ICD-9-CM3 (Internacional Classification of Diseases 9th

Revision Clinical Modification). Se trata de un catálogo de enfermedades codificadas con un número de 3 a 5 dígitos con un punto decimal después del tercer dígito. Los códigos ICD-9-CM están organizados de manera jerárquica en los que se agrupan varios códigos consecutivos en los niveles superiores. El número de códigos asignados a cada documento varía de 1 a 7. El total de etiquetas distintas utilizadas en la colección es 142. La cantidad de información suministrada en cada documento es muy escasa pero relevante y bien estructurada.

- **OHSUMED**

OHSUMED (Hersh, 1994) es una colección que contiene 348,566 referencias compuestas por el título y/o resumen, derivadas del subconjunto de las 270 revistas incluidas en el KF MEDLINE (base de datos bibliográfica de estudios de literatura médica mantenida por la Biblioteca Nacional de Medicina (NLM)) de Atención Primaria de productos que abarca los años 1987 a 1991. La mayoría de las referencias que contiene la colección se refieren a artículos de revistas, aunque también posee un pequeño número de referencias a cartas al editor, actas de congresos y otros informes. El corpus contiene las 101 preguntas que generadas por los médicos en el curso de atención al paciente, así como una breve declaración de los pacientes.

- **Datos disponibles de forma pública facilitados por USC (Universidad de Santiago de Compostela), UIUC (Universidad de Illinois en Urbana Champaign) y TREC (Concurso internacional o Conferencias de evaluación y recuperación de texto).**

Este conjunto de datos está conformado por 5500 preguntas, el mismo fue etiquetado manualmente por el UIUC de acuerdo con las siguientes categorías.

- ABBR: abreviatura (abbreviation), expansión (expansion). Con 86 preguntas.
- DESC: definición (definition), descripción (description), forma (manner), razón (reason). 1162 preguntas.
- ENTY: animales (animal), cuerpo (body), color (color), creación (creation), moneda (currency), enfermedad médica (disease/medical), evento (event), alimento (food), instrumentos (instrument), lenguaje (language), carta (letter), otros (other), plantas (plant), productos (product), religión (religion), deporte (sport), sustancia (substance),

símbolo (symbol), técnica (technique), expresión (term), vehículo (vehicle), palabra (Word). 1251 preguntas,

- HUM: descripción (description), grupo (group), individual (individual), title. 1223 preguntas.
- LOC: ciudad (city), país (country), montaña (mountain), otro (other), estado (state). 835 preguntas.
- NUM: código (code), cuenta (count), fecha (date), distancia (distance), dinero (money), orden (order), otro (other), por ciento (percent), periodo (period), velocidad (speed), temperatura (temperature), tamaño (size), peso (weight). 896 preguntas.

- **Brown corpus**

La colección Brown Corpus (Francis W, 1979), está formada por 1.014.294 (se compone de 5000 textos de 2.000 palabras) de palabras de inglés americano contemporáneo, es el primer corpus en formato electrónico recopilado en la Universidad de Brown a principios de los sesenta.

Principales categorías con sus principales subdivisiones y el número de muestra en cada una de ellas, de la colección.

A. *PRENSA: Reportaje* (44 textos). Políticos, Deportes, Sociedad, Spot News, Financiero, Cultura.
B. *PRENSA: Editorial* (27 textos). Diario institucional, Personal, Cartas al Editor.
C. *PRENSA:* Comentarios (17 textos). Teatro, Libros, Música, Danza.
D. *RELIGIÓN* (17 textos). Libros, Periódicos, Traces.
E. *HABILIDADES Y AFICIONES* (36 textos). Libros, Periódicos.
F. *POPULAR LORE* (48 textos). Libros, Periódicos.
G. *BELLES LETTRES: Biografía, Memorias, etc* (75 textos). Libros, Periódicos.
H. *DIVERSOS* (30 textos). Documentos Gubernamentales, Informes de fundación, Informes de industria, Catálogo del colegio, Industria Cámara órgano.
J. *APRENDIDAS* (80 textos). Ciencias Naturales, Medicina, Matemáticas, Sociales y Ciencias de la Conducta, Ciencias Políticas, Derecho, Educación, Humanidades, Tecnología e Ingeniería.

K. *FICCIÓN: General* (29 textos). Novelas, Historias cortas.
L. *FICCIÓN: Misterio y Detective* (24 textos). Novelas, Historias cortas.
M. *FICCIÓN: Ciencia* (6 textos). Novelas, Historias cortas.
N. *FICCIÓN: Aventura y Occidental* (29 textos). Novelas, Historias cortas.
P. *ROMANCE Y LOVE STORY* (29 textos). Novelas, Historias cortas.
R. *HUMOR* (9 textos). Novelas, Ensayos.

1.4 CATEGORÍAS A CLASIFICAR LOS DOCUMENTOS

Disímiles son las categorías en las que se puede clasificar un documento. Para el desarrollo del proceso de clasificación de los documentos HTML visitados por los usuarios de la UCI en el proyecto FILPACON se tienen en cuenta las siguientes categorías, las cuales a su vez se dividen en subcategorías a la hora de realizar el trabajo de clasificación.

1. **Información/Comunicaciones**

   - ***Prensa digital/Revistas***

   Sitios relacionados con la prensa y revistas digitales con el afán de comunicar información.

   - ***Correo***

   Sitios que les permiten a los usuarios de Internet enviar y recibir correos.

   - ***Chat***

   Sitios que permiten a los usuarios compartir información con los demás de forma directa sitios y canales de chat.

   - ***Grupos de discusión/Boletines/Blogs/Foros***

   Sitios que permiten compartir opiniones e información sobre tópicos variados. También se incluyen las bitácoras personales.

   - ***SMS/Accesorios para teléfonos móviles***

Sitios que permiten a los usuarios enviar mensajes SMS vía Internet a otros usuarios. Se incluyen además proveedores de servicios relacionados con la telefonía móvil.

- ***Postales digitales***

Sitios que permiten a las personas enviar postales digitales a través de Internet.

- ***Buscadores/Catálogos/Portales***

Sitios que contienen buscadores, catálogos y portales Web.

2. TIC

- ***Hardware/software/Distribuidores***

Sitios relacionados con fabricantes de hardware, vendedores de software, y distribuidores de hardware y software.

- ***Proveedores/Alojamiento***

Sitios relacionados con proveedores de alojamiento web (web hosting, en inglés) son aquellos que permiten utilizar sus servicios para poner en línea nuestro sitio web, utilizando además su conexión a la red

- ***Seguridad***

Sitios que informan a los usuarios acerca de la seguridad y privacidad de los datos.

- ***Proxy anónimos***

Sitios que ofrecen a los usuarios la oportunidad de navegar en Internet de forma anónima.

3. Drogas

- ***Drogas ilegales***

Sitios que incluyen información acerca de drogas ilegales y las instrucciones para su uso. Sitios donde se promueve el uso de las drogas (LSD, heroína, cocaína, XTC, éxtasis, amphetamines).

- ***Alcohol/Tabaco***

Sitios que ofrecen información acerca del alcohol y el tabaco como actividades de placer. Se incluyen los sitios de los vendedores de estos productos. Sitios que promueven el consumo del alcohol.

- ***Consejería/Auto ayuda***

Sitios que brindan ayuda a problemas de adicción, consejos matrimoniales y consejería.

4. Vida diaria

- ***Relaciones personales***

Sitios que promueven y gestionan las relaciones interpersonales entre los usuarios.

- ***Bares/Restaurantes***

Sitios que brindar información acerca de bares, restaurantes, etc.

- ***Viajes***

Sitios relacionados con viajes, agencias de transporte, giras, etc.

5. Buscadores de trabajo

- ***Buscadores de trabajo***

Sitios relacionados con la búsqueda y propuestas de trabajo

6. Transporte

- ***Transporte***

Sitios que contienen información relacionada con los medios de transporte.

7. Armas

- ***Armas***

Sitios que brindan información acerca de armas. Compra y venta.

8. Medicina

- ***Salud***

Sitios que promueven la salud. Sitios sobre hospitales, farmacias, doctores, medicamentos, hábitos dietéticos, enfermedades, primeros auxilios, sexología, salud mental, psicología, universidades virtuales de salud. Sitios que tratan sobre trastornos afectivos: Depresión, manía, entre otros.

- ***Aborto***

Sitios que contienen información relacionada con el aborto, en la mayoría de los casos en contra del mismo.

9. Pornografía

- ***Pornografía***

Sitios que cuyos contenidos describen o muestran actividades sexuales explicitas y contenidos eróticos inadecuados para menores de edad. Sitios que tratan acerca de la prostitución, sitios con carácter obsceno de obras literarias o artísticas. Sitios de ofensas al pudor.

10. Negocios

- ***Negocios***

Cuando se habla de sitios de negocios en Internet, se hace referencia de sitios que te ofertan cualquier producto. Sitios de comercio electrónico. Sitios de compra, venta o cambio de mercadería o valores, donde los productos se ofertan a los clientes y los mismos pueden seleccionar productos y hacer órdenes. (Mayoristas y distribuidores, mercado inmobiliario, óptica, papel, productos de consumo, negocio electrónico, marketing, economía, publicidad de negocio).

11. Sociedad/Religión/Educación

- ***Organizaciones gubernamentales***

Sitios que brindan información acerca de organizaciones gubernamentales. Sitios de gobiernos o que se refieren a aspectos de suma importancia para el gobierno como son: cuestiones legales relacionadas con el comercio, como los derechos mercantiles .Además son sitios que tratan sobre el fenómeno migratorio, relaciones étnicas, demografía y población.

- ***Organizaciones no gubernamentales***

Sitios que brindan información acerca de organizaciones no gubernamentales. Cuando se hace referencia a un sitio de organizaciones no gubernamentales hablamos de sitios que tratan diferentes problemas en la sociedad, pues se abordan temas comerciales, medio ambientales, tratan las prácticas en materia de derechos humanos de gobiernos de cualquier tendencia política. De manera general son sitios que dedicados a servir a Organizaciones No Gubernamentales y a ciudadanos y ciudadanas que trabajan por la justicia social, la sustentabilidad y temas relacionados.

- ***Ciudades/Regiones/Países***

Sitios que brindan información acerca de diferentes regiones del planeta, países, ciudades y mapas. (Costumbres, principales zonas turísticas, mapas, revistas, portales).

- ***Educación***

Sitios de universidades, escuelas públicas y privadas, colegios, educación para adultos,

educación sexual, ofertas de cursos, diccionarios, enciclopedias, traductores. Sitios que promueven la acción y efecto de educar. Sitios que desarrollan facultades morales e intelectuales.

- ***Religión***

Sitios que contienen información acerca de las religiones y comunidades religiosas. Sitios de creencias o dogmas acerca de la divinidad y de las prácticas rituales. Sitios que brindan información sobre religiones antiguas, cristianismo, esoterismos.

- ***Sectas***

Sitios Web cuyos contenidos están relacionados con sectas, cultos, grupos síquicos, ocultismo, satanismo. Sitios que profesan una creencia común, que hacen proselitismo y buscan adeptos.

12. Actividades criminales

- ***Actividad ilegal***

Sitios que contienen actividades contrarias a la ley como son: construcción de bombas, pornografía infantil, piratería de música y software, difamación y calumnias, escritura de código malicioso, abuso de tarjetas de crédito, juegos online fraudulentos, entre otros.

- ***Políticas extremas/Odio/Discriminación***

Sitios cuyos contenidos están relacionados con corrientes políticas extremas, machismo, racismo, represión de minorías (sitios neo nazis). Sitios de aversión, de aborrecimiento hacia algo determinado.

- ***Warez/Hackeo/Software ilegal***

Sitios que contienen información relacionada con ataques informáticos, piratería y crack para software.

13. Extremos

- ***Violencia***

Sitios que promueven aplicar medios violentos a cosas o personas, sitios que promueven el abuso físico. Sitios que tratan sobre la discriminación, la violencia sexual y trata de mujeres, entre otros.

14. Juegos

- ***Apuestas***

Sitios de loterías, casino y agencias de apuestas.

- ***Juegos de computadora***

Sitios relacionados con los juegos para computadoras. Sitios de entretenimiento.

- ***Juguetes***

Sitios que contienen información relacionada con juguetes, juegos de mesa, juegos de sala. Sitios relacionados con objetos destinados a la diversión.

15. Deporte

- ***Deporte***

Sitios cuya información está relacionada con el deporte, tipos de deporte, técnicas deportivas, clubes, equipos, federaciones deportivas, resultados deportivos, juegos olímpicos, campeonatos mundiales.

16. Cultura/Entretenimiento

- ***Televisión/Cine***

Sitios cuyos contenidos están relacionados con el cine, la televisión, información de programas y vídeo.

- ***Recreación/Diversión***

Sitios que contienen información sobre actividades recreativas (zoológicos, piscinas públicas, parques de diversiones). Sitios de entretenimiento y distracción.

- ***Arte/Monumentos/Memoriales***

Sitios relacionados con las artes (artes escénicas, artes plásticas, artes gráficas, arquitectura, artesanía,), teatros, museos, monumentos históricos y memoriales. Sitios relacionados con la pintura, la arquitectura, la literatura.

- ***Música***

Sitios de estaciones de radios, radios en línea, mp3, bandas musicales, vendedores de música, y descarga de música.

- ***Literatura/Libros***

Sitios que contienen literatura y libros digitales. Bibliotecas online que prestan servicios literarios y libros.

- ***Humor/Cómico***

Sitios que ofrecen contenidos humorísticos, graciosos, noticias insólitas, disparatadas.

1.4.1 CATEGORÍAS A TENER EN CUENTA PARA LA CREACIÓN DE LA COLECCIÓN

Las categorías a tener en cuenta en el presente trabajo a la hora de realizar la o las colecciones de entrenamiento para el correcto aprendizaje y validación de los algoritmos inteligentes utilizados por el motor de categorización automática del proyecto FILPACON son las siguientes:

1. **Medicina**

- ***Salud***

Sitios que promueven la salud. Sitios sobre hospitales, farmacias, doctores, medicamentos, hábitos dietéticos, enfermedades, primeros auxilios, sexología, salud

mental, psicología, Universidades virtuales de salud. Sitios que tratan sobre trastornos afectivos: Depresión, manía, entre otros.

2. Deporte

- ***Deporte***

Sitios cuya información está relacionada con el deporte, tipos de deporte, técnicas deportivas, clubes, equipos, federaciones deportivas, resultados deportivos, juegos olímpicos, campeonatos mundiales.

3. Sociedad/Religión/Educación

- ***Educación***

Sitios de universidades, escuelas públicas y privadas, colegios, educación para adultos, educación sexual, ofertas de cursos, diccionarios, enciclopedias, traductores. Sitios que promueven la acción y efecto de educar. Sitios que desarrollan facultades morales e intelectuales.

Las subcategorías son la temática a tratar a la hora de realizar la colección de entrenamiento.

1.5 DESCRIPCIÓN DE LA TAREA

Dado un documento HTML visitado por un usuario el objetivo del motor de categorización automática es decidir a que tipo de información está accediendo el usuario, clasificando la página web de acuerdo a las diferentes categorías en las que puede ser clasificado un documento. Para esto el sistema hace uso de la *colección de entrenamiento*, preclasificada de antemano, para comparar la similitud entre el documento que esta siendo visitado y la colección, clasificando la página web. Por tanto se presenta la problemática

¿Cómo elaborar una eficiente colección de entrenamiento?

1.5.1 TRABAJOS RELACIONADOS CON LA CREACIÓN DE COLECCIONES

Son disímiles los trabajos en los que han sido elaboradas diferentes colecciones de entrenamiento, para el desarrollo de clasificación de documentos. Muestra de ello:

- ***Filtrado de contenido web en español dentro del proyecto Poesía*** **(Puertas, 2004)**

Para entrenar y evaluar el sistema, se construyó una colección de documentos web con contenidos pornográficos y otra con contenidos seguros, todos en español. Mediante un robot partiendo del directorio público DMOZ, se recuperaron todas la páginas que colgaban de las categorías regionales en español, constando así la colección con más de 35.000 documentos para entrenar y evaluar el filtro. La colección de documentos recuperados fue postprocesada para eliminar páginas vacías y aquellas que a pesar de estar categorizadas dentro de una categoría en español, su contenido tenía mayor cantidad de términos en otro idioma (se usó un reconocedor de idioma para realizar la labor de filtrado).

- ***Clasificación automática de textos de desastres naturales en México*** **(Téllez, 2003)**

Para la confección de esta colección, si hizo uso del periódico Reforma (www.reforma.com) como fuente de información principal. De este sitio se recopilaron noticias relacionadas (tanto relevantes como irrelevantes) con los fenómenos naturales de huracán, inundación y sequía correspondientes a los últimos dos años. Las noticias relevantes incluyen información del fenómeno natural, mientras que las catalogadas como irrelevantes contienen palabras o frases usadas comúnmente en la descripción de un fenómeno natural pero que se usan en contextos muy diferentes. El conjunto de entrenamiento obtenido finalmente consistió de 375 documentos, de los cuales el 11.5 % son noticias relevantes y el 88.5 % restante son irrelevantes.

- ***Categorización de texto sensible al coste para el filtrado de contenidos inapropiados en Internet*** **(Gómez, 2003)**

Para el desarrollo de la colección se tomó como base el directorio Open Directory Project (ODP). Utilizando un procesador y un robot software programado al efecto, se construyó la colección. Fueron recolectadas todas las direcciones válidas para adultos en inglés y en español, y se seleccionó de manera aleatoria un subconjunto de ellas, obteniendo 5.335

direcciones válidas y 1.021 direcciones para adultos en español, y 5.091 direcciones válidas y 1.002 direcciones para adultos en inglés.

Fueron descargadas todas las páginas accesibles en las direcciones anteriores, excluyendo aquellas que tardaban más de 10 segundos en responder, obteniendo 4.956 documentos válidos y 966 para adultos en español, y 2.570 documentos válidos y 129 para adultos en inglés.

- ***Categorización automática de documentos en español: algunos resultados experimentales*** **(Figuerola, 2000)**

Para la construcción de esta colección de entrenamiento fueron utilizadas 2.741 noticias publicadas en enero de 1994 en el periódico español EL MUNDO. Se trataban de documentos de una extensión media de 3.603 caracteres, con notable uniformidad en cuanto a su tamaño. La selección fue exclusivamente con noticias correspondientes a diferentes secciones del periódico, habiendo desechado materiales como artículos de opinión, editoriales, etc. El número de noticias o documentos utilizados de cada una de las secciones fue aproximadamente similar, aunque no exactamente el mismo. Se tuvo en cuenta a la hora de buscar la información abarcar un rango temporal (referido a las fechas de las noticias) compacto, en la idea de que las características de cada sección pueden variar notablemente con el tiempo. Se partió de la base de que cada una de estas secciones constituye una clase o categoría.

- ***Clasificación de documentos escritos en euskara: impacto de la lematización*** **(Arregi O, 2002)**

Este corpus utilizado proviene del diario Euskaldunon Egunkaria, único diario escrito totalmente en euskara. Los textos utilizados corresponden a los artículos de prensa publicados durante los meses de enero y febrero de 1999. Todos los documentos que contiene esta colección son de la misma época y del mismo estilo. El corpus está compuesto de 5.809 documentos de ellos el 75% o sea 4357 documentos están destinados para el entrenamiento del clasificador, los mismos son clasificados en 7 categorías (economía con 330 documentos, europa con 367, sociedad con 694, deporte con 1294,

cultura con 568, mundo con 314 y política con 790), correspondientes a otros tantos artículos de prensa.

CONCLUSIONES

El objetivo fundamental de la investigación, el desarrollo de una colección de entrenamiento implica el estudio de la categorización automática de páginas web, así como de las categorías en las que se puede clasificar un documento. En el capítulo se han revisado algunos de los algoritmos más frecuentes en la clasificación supervisada de documentos; técnicas cuyos resultados son comparables a los conseguidos por las personas. Se ha establecido una definición formal de lo que es una colección de entrenamiento, necesaria para la validación de los algoritmos inteligentes en el proceso de aprendizaje automático supervisado, se han presentado algunos tipos de colecciones existentes, se han puntualizado las categorías a tratar en el presente trabajo: salud, educación y deporte, para así crear una adecuada colección que cumpla con la necesidad del proyecto FILPACON a la hora de regular de manera automática la información a la que acceden los usuarios de la UCI.

El aprendizaje automático es la rama de la inteligencia artificial que realiza estudios para saber como realizar programas que puedan aprender por si solos, conforme vayan adquiriendo experiencia, como el ser humano. Se pretende que las máquinas realicen actividades comunes para el hombre de manera rápida, eficaz, incrementando no sólo un proceso si no varios. Aunque esto no se ha logrado de manera total, se han obtenidos grandiosos avances, por ejemplo, en procesos de clasificación.

El desarrollo de un clasificador automático se hace necesario ya que el procedimiento mecánico de clasificación:

1. Puede ser mucho más rápido.
2. Puede evitar prejuicios adoptados por un clasificador humano.
3. Puede evitar acciones costosas.
4. Ayuda al supervisor humano a concentrarse en los casos particularmente difíciles de clasificar.

La clasificación de los documentos HTML se hace necesaria para una mejor regulación de los diferentes contenidos a los que acceden los usuarios en una comunidad. Es mucha la información que aparece en la red cada día, son numerosos los contenidos que podemos encontrar en la web, como mismo encontramos ofertas de teatro, de compra de música, se encuentran ofertas pornográficas, de adhesión a sectas, se encuentran páginas que promueven la violencia. El proyecto FILPACON pretende realizar la clasificación de las páginas web visitadas en diferentes temáticas, para así conocer a que tipo de contenido (inadecuado o no) esta accediendo un determinado usuario.

La clasificación automática es un proceso con el que se pretende asignar a un documento una categoría definida con anterioridad, partiendo de un entrenamiento previo del clasificador. Para que el programa analice las características que determinan a que categoría pertenece un documento necesita de una serie de documentos a los que se le haya asignado una categoría con anterioridad (las colecciones de entrenamiento), de esta forma cuando el clasificador procese un nuevo documento puede deducir a que categoría o clase pertenece el mismo. Esta deducción se basará en la cercanía o similitud que exista entre el documento a clasificar y la colección de entrenamiento.

Capítulo 2 MODELOS DE REPRESENTACIÓN TEXTUAL

INTRODUCCIÓN

El desarrollo y crecimiento de las redes de computadoras con el transcurso de los años, ha motivado la aparición de un creciente interés por los sistemas de clasificación automática de documentos. Estos sistemas realizan diferentes operaciones de clasificación basándose en el análisis del contenido del texto de los documentos que procesan. Resulta una tarea fundamental para el procesamiento automático realizar como primera acción a la hora de aplicar técnicas de clasificación automática, la representación de los documentos.

La representación de documentos supone el primer paso en la aplicación de cualquier técnica de clasificación automática, es necesario transformar un documento desde su formato inicial a una forma adecuada a la entrada del algoritmo de clasificación que se vaya a emplear posteriormente, es decir a un formato más adecuado a la tarea que se vaya a realizar.

Las técnicas de análisis y representación de documentos utilizadas en la actualidad en los sistemas de clasificación, se basan en criterios fundamentalmente estadísticos, centrados en frecuencias de aparición de términos en los documentos.

El presente capítulo se estructura como sigue: en primer lugar se establece una definición formal para un modelo general de representación de textos, seguido se presentan las características fundamentales de los modelos vectoriales más usados en tareas de categorización de textos. Se presenta una definición formal del modelo de espacio vectorial y del índice de latencia semántica, a continuación se muestran diferentes funciones de ponderación, seguido de algunas funciones de reducción de rasgos que podrían emplearse, en determinadas condiciones, como funciones de ponderación. Por último son presentadas las funciones de similitud que permiten estimar la cercanía entre dos documentos y así asignar correctamente una categoría al documento que está siendo procesado.

2.1 REPRESENTACIÓN DE DOCUMENTOS

Para llevar a cabo la clasificación automática es preciso contar con una forma consistente de representar el contenido de cada documento, para que la representación se genere de forma automática.

Una representación se puede considerar como el conjunto de cadenas (palabras), que representan el contenido del documento a representar, puesto que cualquier modelo de representación considera la palabra como elemento fundamental.

La representación se realiza con el fin de reducir la complejidad de los documentos y hacerlos más fáciles de manejar a la hora de categorizarlos, la versión completa del texto debe transformarse a un documento vectorial que describe el contenido del mismo. Un documento puede verse como un conjunto de términos que tienen diferentes patrones de ocurrencia.

Los términos se definen como palabras filtradas mediante una lista de parada (stoplist) y desprovistas de sus afijos (extracción de raíces o stemming). Los términos han de ser buenos desde el punto de vista semántico: deben capturar lo máximo posible el significado de los textos, y desde el punto de vista de aprendizaje: deben permitir un aprendizaje eficiente y efectivo.

Un modelo de representación de documentos se puede definir como una cuádrupla (X, B, μ, F), donde (X, B, μ) es un espacio de medida y F es una función sobre el conjunto de objetos X, de forma que $F : f(X) \rightarrow R$.

Todo modelo de representación se define a partir de los siguientes elementos:

1. Un *conjunto de objetos X* que formará el vocabulario empleado en la representación, denominados rasgos.
2. Un *álgebra B*, que generaliza las relaciones entre los objetos.
3. Una *función de medida μ*, con la que se establece la distancia o similitud entre objetos del espacio medible (X, B).

4. Una *función de proyección o de ponderación F* aplicable sobre el conjunto de objetos X, que establezca la proyección de cada objeto del espacio sobre la recta real, indicando la relevancia de dicho objeto en la representación.

Elegir una representación de la información contenida en lo documentos implica definir un espacio matemático para poder operar sobre él y utilizar aproximaciones de aprendizaje automático.

Para realizar el proceso de clasificación automática de textos se tiene que representar cada documento de los ejemplos de entrenamiento, de manera que esa representación se le pueda aplicar el algoritmo de clasificación.

Los principales objetivos de la representación de documentos son: mantener de manera fiel la información que aporta el contenido del documento y por otra parte ser adecuada a las especificaciones del algoritmo de aprendizaje a utilizar.

2.1.2 REPRESENTACIÓN DE PÁGINAS WEB

La información contenida en la Web se presenta principalmente por medio de documentos HTML, tanto estáticos como dinámicos. La World Wide Web puede verse como un conjunto de documentos HTML relacionados por medio de hiperenlaces.

Resulta fundamental para el procesamiento automático de las páginas web representar de un modo adecuado los documentos HTML. Esta representación debe ser fiel al contenido del documento, incluyendo la información necesaria para poder extraer un conocimiento útil y, a la vez, debe ser compatible con las entradas a los algoritmos que se empleen a continuación. Representar las páginas web varía en función de los elementos de la página que se tengan en cuenta: texto plano, texto enriquecido con etiquetas HTML, enlaces, textos asociados a enlaces.

Inicialmente para la representación de páginas web se aplicaron las mismas técnicas que se aplicaban a los textos incluyéndose estas representaciones dentro de la representación por contenido, que no tiene en cuenta ni la estructura, ni la topología del documento. Posteriormente se introdujeron propios de la web en la representación, como los hiperenlaces y las propias URLs.

Desde la creación del lenguaje HTML se han propuesto modelos de representación de las páginas web:

- ***Representaciones por contexto.***

Se analiza la topología del grafo de hipertexto, la estructura del sitio web o de una colección de páginas web a través de los datos relativos a su conectividad. Sirve para saber cómo está organizada una web, cómo está estructurada y cómo es la navegación a través de ella.

Mediante la minería de estructura o representación por contexto, se obtiene información acerca de si los usuarios encuentran la información que buscan, si la estructura del sitio es demasiado ancha(mucho contenido en una misma página) o demasiado profunda(muchos links que estructuran el mismo contenido), si los elementos están colocados en los lugares adecuados dentro de la página, si la navegación se entiende, cuáles son las secciones menos visitadas y su relación con el lugar que ocupan en la página central. En este tipo de representación se hace un análisis de la navegabilidad de los sitios, de cómo estos se interconectan con otros y relaciones de contenido de ligas con otros sitios.

- ***Representaciones por uso.***

Se analizan los registros de acceso almacenados en los servidores web. Trata de extraer patrones de uso de la web por parte de los usuarios. Para ello se utilizan los archivos Log de los servidores Web de forma que aplicando minería de textos sobre ellos se pueda extraer información útil. Este tipo de representación utiliza la información de la minería de estructura para determinar el tipo de comportamiento de navegación de los usuarios en los sitios de Internet.

La información que se puede obtener de un servidor, más los patrones encontrados en los hábitos de navegación, brindan una excelente información sobre el tipo de comportamiento que posee un determinado usuario sobre el sitio, el tipo de información que busca, como lo busca, como lo encuentra, porcentajes de éxito de búsqueda.

- ***Representaciones por contenido.***

La recogida de información es relativa a los contenidos de la página web. Trata de extraer información relevante sobre los mismos de manera que pueda ayudar clasificarlo, aumentando la organización del contenido, para así mejorar el acceso a dicha información. La minería de contenido o la representación por contenido describe la búsqueda automatizada de recursos e información que se encuentra en los sitios de Internet e involucra la minería de los datos de las páginas.

La minería de contenido se basa en el análisis no solo del texto, sino de los distintos elementos que se encuentran en los sitios de Internet, sin importar si su procedencia es estática ó dinámica. Las imágenes, videos, aplicaciones, meta leguajes, meta caracteres, cualquier dato es analizado por el motor de análisis de contenido. Actualmente, la mayoría de los sistemas de minería de contenido emplean representaciones de análisis de contenido de textos, por ejemplo, listas de conceptos o palabras clave y relaciones entre palabras, facilitando el análisis de los textos.

El objetivo de la presente investigación es obtener una colección de entrenamiento asentándose en el texto de las páginas web, las representaciones que se proponen tienen solo en cuenta el contenido textual de los documentos HTML adentrándose en la representación por contenido. Una vez analizada la página web se extrae el texto del documento HTML, reduciéndose el enfoque por contenido a un problema de representación automática de textos.

2.2 MODELOS VECTORIALES

Los modelos de representación vectoriales son técnicas sencillas, se basan en que el significado de un documento puede derivarse del conjunto de rasgos presentes en el mismo. Los documentos se modelan como conjuntos de rasgos que pueden ser individualmente tratados y pesados. Estos rasgos serían los vectores generadores de un espacio vectorial.

Mediante estos modelos los documentos pasan a ser representados como vectores dentro de un espacio euclídeo, de forma que midiendo la distancia entre dos vectores se trata de estimar su similitud como indicador de cercanía semántica.

Dos vectores cercanos en el espacio vectorial representan dos frases semánticamente próximas entre sí en el espacio textual. Una representación vectorial en el espacio de dimensiones definido por el léxico de un texto lleva implícita la representación semántica de dicho texto.

Los modelos vectoriales consideran cada rasgo como un objeto independiente, representan modelos más flexibles, al permitir realizar el pesado de cada rasgo individualmente de forma que este pueda considerarse más o menos importante dentro de un documento o de la colección. Aplicando diferentes métricas para el cálculo de la distancia entre documentos con una función de medida se puede intensificar o no la importancia de unos rasgos frente a otros.

Los primeros trabajos en representación vectorial surgieron en el ámbito de la clasificación de documentos. Esta se realizaba en un primer momento de manera manual, asumiendo que un documento pertenecía a una determinada clase si contenía determinados rasgos que previamente se habían etiquetado como pertenecientes a la clase. Posteriormente con el estudio realizado por parte de diversos conocedores del tema (H. P. Luhn, H. Borko, M. Bernick, M. Koll) se fueron perfeccionando los trabajos realizados con anterioridad, proponiendo diferentes modificaciones a las ideas iniciales, lo que ha hecho que estos modelos sean mas complejos y sofisticados (Luhn, 1953). (Luhn, 1957).(Koll, 1979).

Dentro de los modelos vectoriales se destacan como modelos fundamentales: el modelo de espacio vectorial y el índice de latencia semántica, aunque este último se puede incluir dentro del modelo de espacio vectorial.

2.2.1 MODELO DE ESPACIO VECTORIAL (Vector Space Model, VSM)

El modelo de espacio vectorial (VSM) es un modelo matemático que ha sido empleado para la representación de documentos en disímiles sistemas dentro del campo de la categorización de texto (Salton, 1983), así como para filtrado, recuperación, indexado y cálculo de relevancia de información. Este modelo propone un marco donde se asignan pesos no binarios, que pueden tomar un mayor rango de valores, haciendo posible la coincidencia parcial entre documentos.

Un documento puede expresarse como un vector (Figuerola, 2001):

D = (c1, c2, c3... cj),

Es decir, como un conjunto de características, hasta un total de j, en el cual c1 es un valor numérico que expresa en qué grado el documento D posee la característica 1, c2 lo mismo para la característica 2, y así sucesivamente.

En VSM cada documento puede ser representado mediante un vector de términos, y una colección de documentos forma una matriz de términos, en donde un término es la unidad mínima de información, ejemplo, una palabra. Cada término lleva asociado un coeficiente o *peso* que expresa la importancia o grado de representatividad del término en el documento. Este peso puede calcularse basándose en las frecuencias de los términos, tanto en toda la colección de documentos con que se trabaje, como dentro de cada documento en particular. La funcionalidad de este modelo estriba en la elección correcta de los pesos de cada término, en la función de asignación de pesos que se utilice.

Un mismo término puede ser más o menos significativo en un contexto que en otro, de manera que tendrá diferente peso en un documento que en otro. El tamaño o número de términos de cada documento también juega un papel importante. No es lo mismo que un mismo término aparezca dos veces en un documento largo, de muchas páginas; a que aparezca dos veces en un documento corto, de un par de párrafos. La relación existente entre documentos y términos clave se representan mediante vectores, cuyos componentes son los pesos de los términos clave en la entidad a la que corresponde el vector.

La idea de este modelo consiste en la construcción de una matriz (o tabla) de términos y documentos, donde las filas corresponden a los documentos y las columnas corresponden a los términos incluidos en ellos.

Las filas de la matriz (vectores) serían equivalentes a los documentos que se expresarían en función de la frecuencia de cada término. La longitud del vector de documentos sería igual al total de términos de la matriz (el número de columnas). De esta manera, un conjunto de *m* documentos se almacenaría en una matriz de *m* filas por *n* columnas, siendo *n* el total de términos almacenamos en el conjunto de documentos.

Todo documento dentro del VSM queda representado como una combinación lineal de vectores base, donde cada coeficiente de la combinación representa la relevancia de cada rasgo en el contenido del documento, calculada con una función F. Dentro del VSM, dos representaciones de un mismo documento serían diferentes siempre que el conjunto de valores que tomen las componentes de los vectores sean diferentes.

Para definir completamente el espacio vectorial de representación, es necesario determinar el valor que deben tomar los pesos. La estimación del peso de cada componente del vector en cada documento puede hacerse de diversas formas. El cálculo de los pesos se efectúa a partir de dos factores: la frecuencia de cada término en cada documento, y un elemento conocido como IDF (Frecuencia inversa del documento). Adicionalmente, suele aplicarse algún factor de normalización que permita soslayar las diferencias en tamaño de los documentos.

2.2.2 ÍNDICE DE LATENCIA SEMÁNTICA (Latent Semantic Indexing, LSI)

El índice de latencia semántica (LSI) se basa en el análisis de latencia semántica (LSA), que permite comparaciones de similitudes semánticas entre los textos (Landauer, 1998). La LSA es una técnica de extracción de términos que realiza una reducción de dimensiones del espacio de representación mediante la proyección de los documentos sobre un espacio ortogonal de baja dimensionalidad. Esta proyección se obtiene mediante la descomposición en valores singulares (SVD) de la matriz de términos por documento.

La principal característica de la representación LSI es: la dependencia semántica entre rasgos. Esta representación se plantea como una variante al modelo de espacio vectorial, en ella el texto se representa en un espacio de coordenadas donde los documentos y los rasgos se expresan como una combinación de factores semánticos subyacentes, formalizando esta similitud a partir del álgebra lineal.

La LSI se puede ver como un análisis de coaparición entre rasgos, donde se usa la función SVD, pretendiendo con ella medir la similitud entre diferentes rasgos de un vocabulario en base a coaparición entre rasgos.

La función SVD se considera también como un método de reducción de rasgos, ya que permite reducir la dimensión de las representaciones. Los conjuntos de rasgos co-ocurrentes son llevados a una misma dimensión en un espacio vectorial de dimensiones reducidas, pretendiendo incrementar la semejanza en la representación entre documentos cercanos semánticamente. SVD encuentra la proyección óptima a un espacio de dimensión reducida explotando patrones de coaparición entre rasgos. Aquellos rasgos que tienen similares patrones de coaparición son proyectados a una misma dirección. Estos patrones tratan de inferir similitud semántica entre rasgos, lo que en ocasiones no es acertado (Manning C, 1999).

El LSI es la aplicación de la función SVD a la matriz rasgo-documento. Una matriz rasgo-documentos es aquella matriz cuyos componentes almacenan información relativa a las frecuencias de aparición de los rasgos en los documentos. La matriz contiene en cada columna un vector que representa a un documento de la colección de modo que el componente representa la frecuencia de aparición del rasgo en el documento. Para el cálculo del componente además de la frecuencia de aparición pueden usarse otras funciones de ponderación.

El algoritmo de SVD es inutilizable para una colección grande, pues encuentra una proyección óptima a un espacio dimensional bajo, ésa es la característica dominante para los patrones de la co-ocurrencia de la palabra.

2.3 FUNCIONES DE PONDERACIÓN

Una función de ponderación es la función aplicada sobre el conjunto de rasgos de un documento, estableciendo la relevancia del rasgo dentro del contenido del documento. Una función de ponderación se corresponde con la función de proyección F dentro de la definición del modelo de representación de documentos.

Las funciones de ponderación se emplean para calcular la importancia, de un rasgo en el contenido de un texto, para el cálculo del peso de los términos en el contexto del documento. Son de carácter variado, dependiendo del uso que se vaya a dar a la representación. Estas funciones pueden emplear parámetros diferentes: desde la frecuencia

de aparición de un rasgo en el documento o en la colección, hasta probabilidades condicionadas de un rasgo a una clase en problemas de categorización de textos.

Las funciones de ponderación se basan en un conteo de frecuencias, ya sea dentro del documento a representar, o en el conjunto de documentos de la colección. Muchas de las funciones de ponderación permiten, ordenar un conjunto de rasgos para a continuación reducir el vocabulario con el que representar un documento. Basta con seleccionar el subconjunto de rasgos con mayores valores del total de rasgos de cada documento, o el subconjunto con mayores valores en el conjunto de la colección. Aunque no todas las funciones de ponderación pueden ser utilizadas como función de rasgos.

Dentro de las funciones de ponderación encontramos funciones de carácter local y global. En el caso de las funciones globales, las condiciones necesarias para asegurar vectores de representación diferentes, cuando se tienen documentos diferentes, son menos restrictivas que en el caso de las funciones locales.

A continuación son presentadas en el siguiente apartado varias funciones de ponderación, a ser empleadas en el VSM y otras utilizadas en LSI.

2.3.1 FUNCIONES LOCALES

Las funciones de ponderación local son aquellas funciones que toman exclusivamente información del propio documento para obtener una representación, sin necesidad de ninguna información externa. Asigna valores a las diferentes características de manera local al documento.

2.3.1.1 FUNCIÓN BINARIA (Binary, Bin)

La función binaria es el método de representación más sencillo, dentro de los modelos de representación vectorial es conocido como conjunto de palabras. En él, la función de ponderación F dentro de la definición del modelo de representación es una función binaria, que considera la presencia o ausencia de un rasgo en un documento para calcular su relevancia dentro del mismo.

La función de relevancia es un valor {0,1}, 1 si el rasgo aparece en el documento, 0 si no aparece.

$$F: \quad Bin(\vec{t_i}, \vec{d_j}) = \begin{cases} 1, & \text{si el rasgo } t_i \text{ aparece en } d_j \\ 0, & \text{si no aparece} \end{cases}$$

Con esta función, dos documentos tienen diferente representación si contienen diferentes rasgos.

2.3.1.2 FRECUENCIA DE APARICIÓN (Term Frequency, TF)

Dentro de los modelos no binarios la función TF, conocida como Bolsa de palabras (bag of words) (Dagan) es la representación más sencilla. La relevancia se representa por la frecuencia de aparición del rasgo en el documento.

$$F: \quad TF(\vec{t_i}, \vec{d_j}) = f_{ij}, \ frecuencia\ del\ rasgo\ t_i\ en\ d_j$$

Cada rasgo tiene una importancia proporcional al número de veces que aparece en el documento

La representación TF sobrevalora los rasgos muy frecuentes, los rasgos que suelen ser palabras de uso común y no resultan ser muy característicos dentro del contenido de un documento. En este caso, dos representaciones son diferentes si contienen rasgos diferentes o si las frecuencias de los rasgos no son todas iguales.

2.3.1.3 FRECUENCIA NORMALIZADA (Weighted Term Frequency, WTF)

La función frecuencia normalizada (WTF) supone una normalización de la frecuencia de un rasgo en un documento con la suma total de frecuencias del conjunto de rasgos presentes en el mismo.

En esta función la relevancia se calcula como la frecuencia de aparición normalizada del rasgo en el documento, generándose la representación conocida como bolsa de palabras normalizada.

$$F: \quad WTF(\vec{t_i}, \vec{d_j}) = \frac{f_{ij}}{\sum_{t_p \in d_j} f_{pj}}$$

fij es la frecuencia del rasgo ti en el documento dj

En este caso, dos representaciones son diferentes si contienen rasgos diferentes o si las frecuencias de los rasgos no son todas iguales, al igual que en la bolsa de palabras.

2.3.1.4 FRECUENCIA AUMENTADA Y NORMALIZADA (Augmented Normalizad Term Frequency, ANTF)

Esta función representa una frecuencia normalizada de un rasgo en un documento. La normalización se realiza con la mayor de las frecuencias presentes en el documento.

$$F: \quad ANTF(\vec{t_i}, \vec{d_j}) = 0,5 + 0,5\frac{f_{ij}}{max(\{f_{pj} \mid t_p \in d_j\})}$$

fij es la frecuencia del rasgo ti en el documento dj

Utilizando esta función, dos documentos tendrán diferente representación en las mismas condiciones de la función anterior, dos representaciones son diferentes si contienen rasgos diferentes o si las frecuencias de los rasgos no son todas iguales.

2.3.1.5 PASADO LOGARÍTMICO

La función pasado logarítmico es utilizada para el cálculo de la matriz rasgos-documentos dentro del LSI en conjunto con la función TF y la función binaria.

$$F: \quad L(\vec{t_i}, \vec{d_j}) = log_2(f_{ij} + 1)$$

fij es la frecuencia del rasgo ti en el documento dj

Este caso implica que dos representaciones son diferentes si los dos documentos son diferentes.

2.3.2 FUNCIONES GLOBALES

Las funciones de ponderación global generalmente toman información de la colección para generar las representaciones. Los coeficientes de los vectores de representación son calculados a partir de información externa al propio documento. En estas funciones se asigna valor a las características teniéndose en cuenta algún tipo de medida respecto al conjunto de documentos global.

Se pueden considerar asimismo como funciones de ponderación global aquellas que posean una parte local y otra global.

2.3.2.1 FRECUENCIA INVERSA DEL DOCUMENTO (Inverse Document Frequency, BinIDF)

La función BinIDF, supone que los rasgos que aparecen en muchos documentos de la colección no son tan descriptivos como aquellos que aparecen en unos pocos, enriqueciendo la representación binaria.

$$F: \quad BinIDF(\vec{t_i}, \vec{d_j}) = \begin{cases} 1 + \log(\frac{N}{df(\vec{t_i})}), & \text{si } f_{ij} \neq 0 \\ 0, & \text{si } f_{ij} = 0 \end{cases}$$

df (ti) es la frecuencia de documentos, el número de documentos en la colección en los que aparece el rasgo ti

fij es la frecuencia de ti en dj

N es la dimensión del corpus

Con esta representación un rasgo toma el mismo valor en cualquier documento de la colección.

El hecho de que dos documentos tengan conjuntos de rasgos diferentes hará que sus representaciones sean también diferentes.

En esta función cuantas más veces aparece un término dentro de un documento más representativo es, a la vez que cuanto mayor sea el número de documentos en los que aparezca menos discriminante será. La importancia de un rasgo si aparece en un documento es inversamente proporcional al número de documentos que lo contienen.

2.3.2.2 FRECUENCIA DEL TÉRMINO X FRECUENCIA INVERSA DEL DOCUMENTO (Term Frequency - Inverse Document Frequency, TF-IDF)

La combinación de la función TF con el factor IDF se propone para evitar que el valor del rasgo presente en el documento sea constante. Puesto que esta función asigna pesos mayores a aquellos términos que aparecen en un solo documento.

$$F: \quad TF - IDF(\vec{t_i}, \vec{d_j}) = f_{ij} \times \log(\frac{N}{df(\vec{t_i})})$$

df (ti) es la frecuencia de documentos, el número de documentos en la colección en los que aparece el rasgo ti

fij es la frecuencia de ti en dj

N es la dimensión del corpus

En este caso, la frecuencia del rasgo corrige el factor IDF de forma que el valor que toma un mismo rasgo en dos documentos es diferente siempre que la frecuencia de dicho rasgo en cada documento sea también diferente. Empleando esta función de ponderación, las

representaciones de dos documentos son diferentes si las frecuencias del conjunto de rasgos que contienen son diferentes.

En la función TF-IDF cada componente consta de dos partes, una local, que indica la importancia del término en cada documento (y que varía con la frecuencia en cada uno), y otra global, que indica la importancia del término en el total de la base (y que sería la misma para todas las coordenadas de un término).

2.3.2.3 FRECUENCIA INVERSA PONDERADA (Weighted Inverse Document Frequency, WIDF)

La función WIDF es una extensión de IDF, que incorpora la frecuencia del término sobre la colección de documentos, normaliza las frecuencias de un rasgo en un documento con la frecuencia de dicho rasgo en la colección.

$$F: \quad WIDF(\vec{t_i}, \vec{d_j}) = \frac{f_{ij}}{\sum_{d_k \in C} f_{ik}}$$

fij es la frecuencia de ti en dj

La sumatoria recorre los valores de 1 a N, *N es la dimensión del corpus*

Esta función supone una corrección a la sobreponderación que realiza la función TF con los rasgos frecuentes. Distingue, de un modo diferente a la función IDF, entre los rasgos que resultan frecuentes en un documento y los que son frecuentes en el conjunto de la colección, penalizando los rasgos que son frecuentes en la colección.

2.3.2.4 FRECUENCIA INVERSA PROBABILÍSTICA (Probabilistic Inverse Frequency, PIF)

La función PIF establece una corrección a la sobre ponderación de la frecuencia del documento penalizando los rasgos que aparecen en un mayor número de documentos.

$$F: \quad PIF(\vec{t_i}, \vec{d_j}) = log(\frac{N - df(\vec{t_i})}{df(\vec{t_i})})$$

fij es la frecuencia de ti en dj

N es la dimensión del corpus

Esta función se emplea en el ámbito del LSI como parte global dentro de la función de asignación de pesos a las componentes de la matriz rasgo-documento. Para asegurar representaciones diferentes en documentos diferentes en esta función, basta con cumplir que el conjunto de rasgos en dos documentos sea distinto.

2.3.2.5 FRECUENCIA GLOBAL X FRECUENCIA INVERSA DEL DOCUMENTO (Global Frequency - Inverse Document Frequency, GF-IDF)

Esta función GF-IDF calcula la relevancia de un rasgo mediante una relación entre la frecuencia global en la colección y la frecuencia del documento. GF-IDF suele emplearse en el cálculo de la proyección SVD del LSI.

$$F: \quad GF - IDF(\vec{t_i}, \vec{d_j}) = \frac{gf(\vec{t_i})}{df(\vec{t_i})}$$

df (ti) es la frecuencia de documentos, el número de documentos en la colección en los que aparece el rasgo ti

Empleando esta función como función de proyección F en la representación de un documento, dos documentos diferentes tendrán diferente vector de representación siempre que los documentos no tengan el mismo conjunto de palabras.

2.3.2.6 FUNCIÓN ENTROPÍA (H)

La función entropía se emplea en el cálculo de las componentes de la matriz rasgo-documento, dentro del modelo de representación LSI.

$$F: \quad H(\vec{t_i}, \vec{d_j}) = 1 - \sum \frac{p_{ij} log_2(p_{ij})}{log_2(N)} \text{ , con } p_{ij} = \frac{f_{ij}}{gf(\vec{t_i})}$$

$$f_{ij} = TF(\vec{t_i}, \vec{d_j}) \text{ y } gf(\vec{t_i}) = \sum_{j=1...N} f_{ij},$$

Como en el caso de la función GF-IDF, mientras que los documentos no tengan el mismo conjunto de palabras se asegura que dos documentos diferentes tengan representaciones diferentes.

En general, la función TF suele mejorar la representación binaria en problemas de categorización de textos; las representaciones con factor IDF (la función TF-IDF es la de uso más extendido) suelen ofrecer mejores resultados que la representación TF.

2.4 FUNCIONES DE REDUCCIÓN DE RASGOS EN LA CATEGORIZACIÓN DE TEXTOS

Dentro de modelos de representación vectoriales pueden emplearse como funciones de ponderación muchas de las funciones definidas para tareas de reducción de rasgos. Las funciones de reducción permiten realizar una ponderación en base a la cual se ordenan todos los rasgos contenidos en un vocabulario para, seguidamente, seleccionar un subconjunto de ellos. La selección se puede hacer estableciendo un umbral de ponderación mínima o prefijando una dimensión reducida, generando un vocabulario que resulte ser un subconjunto del vocabulario inicial, asignándose un valor a cada rasgo del vocabulario; por tal motivo la representación de un documento se podría realizar tomando estas funciones de reducción de rasgos como funciones de ponderación.

Las funciones de reducción de rasgos se pueden ubicar en dos enfoques: ***enfoque de filtrado*** donde las funciones resultan no orientadas a tarea puesto que son independientes del método de aprendizaje que se vaya a aplicar y el ***enfoque de envoltura*** donde la selección de rasgos se realiza con el mismo método de aprendizaje que será empleado posteriormente, y sobre una colección representada con el propio resultado del proceso de reducción. Estos métodos son orientados a tarea.

Una función de ponderación local y una función de reducción de rasgos se diferencian en que en las funciones de reducción el pesado se realiza en términos de una información extraída de la colección y no del documento. Lo mismo sucede con las funciones de ponderación global que no tienen componente local.

Las funciones de reducción de rasgos requieren un tratamiento previo, y en profundidad, de la colección de documentos que se esta considerando. No basta con realizar tareas de conteo de frecuencias, como en las funciones de ponderación global, sino que en muchos casos será necesario realizar cálculos complejos para determinar probabilidades condicionadas. A continuación, se muestran algunas de las funciones de reducción de rasgos más utilizadas, estas excreciones se pueden encontrar en (Sebastiani, 1999).

2.4.1 GANANCIA DE INFORMACIÓN (Information Gain, IG)

La medida IG se usa para establecer la calidad de un determinado rasgo. Consiste en medir el número de bits de información obtenida para predecir la categoría por medio de la presencia o ausencia de una palabra en el documento. A partir del cálculo de la ganancia de información de cada término es posible identificar aquellos términos con mayor poder discriminativo.

$$P(t_k, c_i) \cdot \log \frac{P(t_k, c_i)}{P(c_i) \cdot P(t_k)} + P(\bar{t}_k, c_i) \cdot \log \frac{P(\bar{t}_k, c_i)}{P(c_i) \cdot P(\bar{t}_k)}$$

El cálculo de esta función requiere la estimación de probabilidades condicionadas, siendo esta la probabilidad a posteriori de cada clase, dado un rasgo; así como el cómputo de la entropía. Dado un corpus de entrenamiento, para cada rasgo se computa su IG y se eliminan aquellos rasgos cuyo valor de IG no supere un umbral mínimo predeterminado.

La función IG es de carácter global, se debe cumplir la condición de que dos rasgos sean diferentes o que las frecuencias de los rasgos no sean todas iguales para que dos representaciones sean diferentes. Esta función mide qué tan bien un atributo dado separa el conjunto de entrenamiento conforme a las clases.

2.4.2 INFORMACIÓN MUTUA (Mutual Information, MI)

La función MI se emplea fundamentalmente para encontrar relaciones entre rasgos, en el contexto del modelado estadístico del lenguaje, puesto que toma un valor individual para cada clase.

$$\log \frac{P(t_k, c_i)}{P(t_k) \cdot P(c_i)}$$

En la función MI el valor que toma cada rasgo es el mismo independientemente del documento, se tiene que cumplir que dos documentos contengan diferentes rasgos para tener diferente representación. El pesado de esta función esta influenciado por la probabilidades marginales de los rasgos, siendo esto una debilidad de la función.

2.4.3 Chi-square (χ2)

La función χ2 mide la falta de independencia entre un rasgo y un documento. Esta medida toma un valor nulo en el caso de que haya total independencia entre el documento y la clase.

$$\frac{g \cdot [P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c_i}) - P(t_k, \overline{c_i}) \cdot P(\overline{t}_k, c_i)]^2}{P(t_k) \cdot P(\overline{t}_k) \cdot P(c_i) \cdot P(\overline{c_i})}$$

La medida χ2 es particular de cada clase, como en la medida la MI. χ2 no tiene dependencia del documento, se emplea como una función F dentro de un modelo de representación para obtener diferentes representaciones de dos documentos, dichos documentos deben contener diferentes rasgos.

2.4.4 Odds Ratio

En esta función, la probabilidad de que un rasgo sea característico de una determinada clase se calcula considerando la ocurrencia de los diferentes rasgos.

$$\frac{P(t_k|c_i) \cdot (1 - P(t_k|\overline{c_i}))}{(1 - P(t_k|c_i)) \cdot P(t_k|\overline{c_i})}$$

El valor Odds Ratio se mide tomando clases dos a dos, y no sobre todo el conjunto de clases. En esta función el valor que toma cada rasgo en el documento es independiente del mismo.

2.5 REDUCCIÓN DE DIMENCIONALIDAD

No todas las palabras que integran la colección de entrenamiento se pueden considerar, pues representados los textos mediante las palabras que lo componen, la dimensión de la colección suele ser alta. Por lo que se hace necesario aplicar alguna técnica para reducir la dimensión de la representación ya que de esta forma se reduce el tiempo de ejecución en la fase de aprendizaje o entrenamiento, y se elimina el ruido de los documentos, es decir, las palabras que no ofrecen información útil para la clasificación y pueden distorsionar los resultados.

Existen métodos de reducción para eliminar las palabras que se repiten en la mayor parte de los textos:

Reducción mediante selección de términos: se obvian aquellas palabras que aparecen en la mayor parte de los documentos de la colección, con el fin de mantener aquellas más relevantes, de mayor peso. Para ello se realiza una valoración de la relevancia de cada una de las palabras del vocabulario de la colección, se ordena el vocabulario de acuerdo a ese valor, y, finalmente, se descarta un porcentaje arbitrario de las palabras con menor relevancia.

Reducción mediante técnicas de extracción de términos: se puede combinar la información presente en varios atributos formando nuevos atributos sintéticos que conservan la semántica de los términos agrupados.

***Reducción mediante proyección aleatoria*:** asegura que cualquier conjunto de vectores de un espacio euclídeo puede ser proyectado sobre otro espacio de menor dimensión, de tal

forma que las distancias entre pares de vectores en el espacio original se mantienen de manera aproximada en el nuevo espacio.

Reducción mediante medidas de relevancia: asignan un valor numérico a cada uno de los términos, permitiendo realizar una reducción del vocabulario al quedarse sólo con aquellos términos más relevantes.

Para reducir la dimensión del espacio se suelen emplear las funciones de reducción de rasgos analizadas en el apartado anterior.

2.6 SELECCIÓN DEL VOCABULARIO. PRE-PROCESAMIENTO

Antes de representar un documento el contenido del mismo debe ser transformado en un conjunto de rasgos, esta fase de transformación se entiende como selección del vocabulario.

Esta selección del vocabulario se ve como un elemento de transformación de la información que inicialmente es de carácter cualitativo y que debe ser transformada a un conjunto de objetos X dentro de un espacio medible X, B, de carácter cuantitativo.

Las fases más comunes cuando la representación se basa en las palabras individuales son las siguientes.

2.6.1 ANÁLISIS LÉXICO

En la fase de análisis léxico se analiza un texto para distinguir las cadenas que formarán parte de la representación. Análisis de cada elemento del lenguaje, con el objetivo de determinar el tratamiento que se realizará sobre números, signos de puntuación, tratamiento de mayúsculas, minúsculas, nombres propios, etc.

En esta fase se extraen todas las palabras contenidas y se eliminan todas aquellas palabras que tengan una frecuencia relativa de aparición extremadamente baja.

2.6.2 IDENTIFICACIÓN DE LOS TÉRMINOS

Durante esta fase se identifican los términos que determinan el dominio concreto representado por los documentos de partida. Esta fase puede ser realizada mediante herramientas informáticas que reconozcan la información electrónica contenida en el corpus documental. La identificación de las diferentes palabras, se realiza a través de autómatas que mediante algoritmos, determinan las diferentes palabras y asignan una categoría gramatical, los términos serán guardados en una base de datos. A este proceso se le conoce como Normalización. En primer lugar se compara cada variante flexionada con un vocabulario del mismo idioma y después se aplican una serie de reglas para la normalización de los términos. De igual modo, se analiza cada una de las formas verbales que aparecen en el documento, se determina su flexión y conjugación para normalizarlo, a una sola forma como podría ser el infinitivo o estructuras verbales más complejas.

2.6.3 LEMATIZACIÓN Y TRUNCADO (stemming)

Son elevados los conjuntos de palabras que aun siendo diferentes tienen una raíz común y por tanto pueden tener un significado léxico equivalente. Son numerosas las palabras que teniendo una raíz común tienen un significado diferente y por ende no se catalogan como una misma palabra. También existen aquellas palabras que siendo diferentes tienen un mismo significado, como es el caso de los sinónimos.

Todo esto se analiza en el proceso de lematización que consiste en asignar a cada palabra su correspondiente lema. Así múltiples palabras con un lema común se agruparían. Formas diferentes como tendrán o tuvieran, compartirán el lema tener y serán considerados como un mismo rasgo dentro del vocabulario. Este proceso requiere recursos lingüísticos adecuados como pueden ser diccionarios electrónicos y un software específico.

El truncado (stemming), tiene como objetivo, reducir el número de rasgos del vocabulario. Este proceso consiste en el truncado de las palabras, eliminando los prefijos y sufijos, plurales, de manera que se pueda realizar una lematización más sencilla y relajada, sin necesidad de contar con un analizador morfológico ni recursos lingüísticos. Así, formas diferentes como comunicación, comunicante o comunicado, comparten el stem comunic y son considerados como un mismo rasgo dentro del vocabulario.

Una alternativa a la eliminación de sufijos es el aprendizaje de reglas de truncamiento a partir de grandes colecciones de documentos. Analizando un conjunto de palabras que forman parte de un idioma, se detectan los prefijos y sufijos que las forman, seleccionando como raíz de cada palabra el prefijo más probable.

La idea del proceso de truncamiento es la misma que la del proceso de lematización: diferentes palabras construidas sobre un mismo lexema suelen representar un mismo concepto.

2.6.3.1 ALGORITMOS DE STEMMING

Existen diferentes algoritmos de truncado para diferentes idiomas (ingles, francés, castellano, árabe, holandés, etc.) entre los que se encuentran:

- **Algoritmo Porter**

Porter fue presentado por primera vez en 1980 y fue desarrollado por Martin Porter en la universidad de Cambridge. Este lematizador lineal secuencial, es considerado uno de los mejores y más conocidos, remueve en cinco pasos controlados más de 60 terminaciones, removiendo terminaciones cortas sin excepciones. Cada paso resulta en la remoción de una terminación o la transformación de la raíz.

- **Algoritmo Lovins**

Lovins fue desarrollado por Jule Beth Lovins del MIT en 1968, este lematizador usa un algoritmo único y varias listas de excepciones para remover más de 260 diferentes terminaciones. Específicamente, este simple y secuencial algoritmo (remueve un máximo de 1 sufijo por vez y contiene 11 tablas conteniendo 260 sufijos, 29 casos para remoción de sufijos y 34 reglas para decodificar terminaciones

- **Algoritmo Paice**

Paice fue publicado 1990 y fue desarrollado por Chris Paice con la asistencia de Gareth Husk. Este lematizador simple e iterativo, remueve las letras finales de una palabra en un indefinido número de pasos. El algoritmo usa una lista separada de agregados finales, el

cual es un arreglo de una lista que esta dividida en series y secciones, cada una de las cuales es correspondiente a una letra del alfabeto.

2.6.4 ELIMINACIÓN DE PALABRAS VACÍAS (stop-words)

En este proceso se eliminan las palabras vacías o stop-words, conjunto de palabras más comunes en un lenguaje, que no aportan información semántica al texto, como las preposiciones, las conjunciones, artículos, adverbios, adjetivos de uso frecuente, etc., palabras que por si solas no tienen poder discriminante.

En esta fase se descartan los términos que se consideran poco discriminantes. Un término es poco discriminante si se encuentra en un documento y nos dice poco (o nada), acerca del posible contenido o tema del documento. También pueden considerarse como poco discriminantes aquellas palabras que están muy frecuentes en la colección. Aunque no todas las palabras muy frecuentes en un lenguaje pueden ser catalogadas como vacías pues en ciertos ámbitos pueden ser discriminantes a la hora de categorizar el texto, y por otro lado, palabras que no se consideran vacías en un ámbito concreto lo pueden ser, puesto que no aportan información alguna.

El listado de palabras vacías se ha de construir con una buena supervisión para no eliminar aquellas palabras que puedan tener cierto poder discriminante. En la comunidad científica se dispone de unas listas de stop-words para numerosos idiomas, entre las que se incluyen algunos verbos, adverbios o adjetivos de uso frecuente. Este proceso de eliminar las palabras vacías se realiza para reducir la densidad del espacio de los documentos.

2.6.5 IDENTIFICACIÓN DE LOS SEGMENTOS REPETIDOS

Los segmentos repetidos son secuencias de palabras que aparecen contiguas en el texto y que usadas de esta forma tienen un significado especial. Por ejemplo “marketing relacional”, “instrumentos de análisis”, etc. Dividir estos segmentos repetidos en los distintos términos que lo forman acarrearía una descontextualización y pérdida de significado.

Identificar los segmentos repetidos que aparecen en un texto podría hacerse fácilmente teniendo acceso a un diccionario que permita identificar la categoría gramatical de cada palabra (sustantivo, adjetivo, preposición, verbo, etc.). El problema consiste en identificar

qué segmentos repetidos tienen realmente una significación especial, y deberían tratarse como términos o conceptos.

En cualquier texto podemos identificar un gran número de segmentos de repetición con una misma estructura sintáctica (sustantivo->adjetivo, sustantivo->preposición->sustantivo), pero de todos ellos tan sólo una mínima parte tendrán un significado especial. Para solucionar este problema, cabe la posibilidad de aplicar técnicas estadísticas que seleccionen únicamente aquellos segmentos de repetición que ocurren con mayor frecuencia en los documentos, o reglas heurísticas que identifiquen únicamente los segmentos de repetición que aparecen en los títulos, títulos de sección, etc., de los documentos.

2.7 CÁLCULO DE SIMILITUD

Realizada la representación de los documentos, e independientemente del algoritmo de clasificación que se utilice después de la misma, es preciso estimar la cercanía, parecido o similitud entre los documentos a clasificar y la *colección de entrenamiento*, para así asignarle una categoría al documento o los documentos que están siendo clasificados.

La idea en este proceso consiste en establecer un vector para cada una de las clases o categorías posibles, que recoja las características de cada una de las clases. Midiendo la similaridad del vector de cada documento con cada uno de los vectores patrón que contienen las características de las clases o categorías. Aquel vector patrón de clase que ofrezca mayor similaridad con el vector del documento será el que más confianza indique, asignándose así la clase o categoría a la cual pertenece el documento en cuestión.

Los coeficientes de similitud y distancia son estimadores cuantitativos que describen el grado de asociación o semejanza entre los elementos comparados, expresado en valor numérico, entre 0 y 1, o en porcentaje (0 y 100%, respectivamente).

Algunas de las funciones de similitud disponibles para estimar la cercanía o parecido entre dos vectores se encuentran a continuación, las ecuaciones de las mismas pueden ser encontradas en (Salton, 1983):

2.7.1 FUNCIÓN DEL COSENO

Esta función de similitud referencia al coseno de ángulo que forman dos vectores que se representan en un espacio multivectorial. Dicha función se basa en el producto entre ambos vectores, aplicando un factor de normalización para obviar los efectos de la disparidad en los tamaños (número de términos) de los vectores o documentos.

La función se basa en calcular el producto escalar de dos vectores de documentos (A y B) y dividirlo por la raíz cuadrada del sumatorio de los componentes del vector A multiplicada por la raíz cuadrada del sumatorio de los componentes del vector B.

Si no existe coincidencia entre los componentes, la similitud de los vectores será cero ya que el producto escalar será cero. La similitud máxima en esta función sólo se da, cuando todos los componentes de los vectores son iguales, obteniendo en este caso la función del coseno su máximo valor, la unidad.

$$\frac{\sum_{i=1}^{m} X_i \cdot Y_i}{\sqrt{\sum_{i=1}^{m} X_i^2 \cdot \sum_{i=1}^{m} Y_i^2}}$$

Ecuación para el cálculo de la función coseno.

2.7.2 COEFICIENTE DE DICE

Esta función está basada en la asociación entre dos términos, calculando el coeficiente de intersección de los dos conjuntos y su unión. Esta aproximación es suficiente para estimar la correlación entre dos palabras, obteniendo el grado de similitud o asociación entre los términos usando una medida de similitud de conjuntos.

$$\frac{2 \cdot \sum_{i=1}^{m} X_i \cdot Y_i}{\sum_{i=1}^{m} X_i^2 + \sum_{i=1}^{m} Y_i^2}$$

Ecuación para el cálculo del coeficiente de Dice.

2.7.3 COEFICIENTE DE JACCARD

El coeficiente de Jaccard es uno de los más populares. Se define mediante la ecuación:

$$\frac{\sum_{i=1}^{m} X_i \cdot Y_i}{\sum_{i=1}^{m} X_i^2 + \sum_{i=1}^{m} Y_i^2 - \sum_{i=1}^{m} X_i \cdot Y_i}$$

Ecuación para el cálculo del coeficiente de Jaccard.

CONCLUSIONES

En este capítulo se estableció una definición formal de la Representación de Textos, fueron presentadas las características fundamentales de los modelos vectoriales más usados en tareas de categorización de textos, el Modelo de Espacio Vectorial (VSM) y el Índice de Latencia Semántica (LSI).

La diferencia entre VSM y la LSI es el enriquecimiento que la LSI incorpora por medio del análisis de coapariciones, requiriendo información de la colección. Este método ha venido aplicándose a colecciones controladas, realizando un análisis cruzado de los contenidos de todos los documentos para generar las representaciones. En el caso del modelo de espacio vectorial, se desarrollan funciones de asignación de peso completamente independientes de cualquier información de la colección.

Los modelos vectoriales representan los documentos dentro de un espacio vectorial en el que se establecen diferentes medidas de similitud (métricas del espacio μ) y de ponderación (funciones de proyección F) para cada una de las componentes X de los vectores de representación. En dichos modelos cada componente es tratado de un modo independiente.

La función de ponderación F es una pieza clave en la definición de los modelos, puesto que en conjunto con un espacio de medida X, B, μ, define el modelo de representación. Han sido revisadas las funciones F, las de carácter local, que no requieren de información previa sobre ningún vocabulario, así como las de carácter global. Fueron presentadas funciones de reducción de rasgos que pueden emplearse, como funciones de ponderación, en determinadas condiciones.

Con el fin de disminuir el tamaño de la colección de entrenamiento y hacerla más manejable, eliminando las partes no relevantes para continuar el proceso de clasificación automática se lleva a cabo el proceso de selección del vocabulario, para transformar el contenido de los documentos al conjunto de rasgos, pues no todos los elementos que aparecen en los documentos son útiles para su clasificación.

Hay elementos que por sí mismos no dicen nada del contenido del documento en el que se encuentran y por tanto pueden ser eliminados; entre ellos se incluyen: los signos de puntuación así como las etiquetas HTML; también aparecen palabras de uso muy frecuente, palabras que aparecen en una gran cantidad de documentos, lo que hace que su poder discriminatorio sea muy bajo; este tipo de palabras se les conoce como palabras vacías, ejemplos de ellas son las preposiciones, conjunciones, entre otras, las cuales pueden ser eliminadas. Otro paso en este proceso de transformación de los rasgos, es el proceso de lematización, mediante el cual se busca reducir las palabras a su raíz, ya que muchas palabras aunque diferentes, tienen la misma raíz léxica.

Realizadas las representaciones pertinentes se hace necesario para dar una categoría adecuada a los documentos que están en proceso de clasificación, estimar la cercanía o similitud entre los documentos y la *colección de entrenamiento*, todo independientemente del algoritmo de clasificación que se valla a utilizar para el proceso de clasificación.

Representar un documento que en la categorización de textos se basa generalmente en la utilización de raíces de palabras, es el primer paso para clasificar un documento, deben ser representadas en un primer momento las colecciones de entrenamiento a utilizar, para luego ser aplicado un algoritmo de clasificación y comenzar el proceso de asignación de categorías o clases a las mismas, así como a los documentos a clasificar por el motor de categorización automática.

Capítulo 3 CREACIÓN DE LA COLECCIÓN DE ENTRENAMIENTO

INTRODUCCIÓN

Internet se ha convertido desde hace años en la mayor fuente de información en formato digital disponible en el mundo. El crecimiento de esta red de comunicación global es imparable por lo que clasificar la información contenida en la web de un modo eficiente se ha convertido en una necesidad.

La idea de la clasificación es organizar los documentos en diferentes temáticas, desarrollándose en torno a ello diversas técnicas, buena parte de tales técnicas se basan en la utilización de medidas de semejanza entre dos documentos, basándose en la premisa de que documentos similares contienen términos similares.

Con la creciente disponibilidad de documentos en formato electrónico, surge la posibilidad de abordar la clasificación de documentos de manera automática. Partiendo de un esquema de clasificación previo, ya establecido, la idea es decidir en qué lugar de este esquema debe ir cada documento a clasificar, así como la selección de la categoría, establecida de antemano, en la cual se inserta cada documento de la colección.

En la clasificación de textos dada una colección de documentos y un conjunto de categorías, el algoritmo de clasificación asigna cada documento a la categoría o grupo de categorías que más se adecuen al contenido del documento.

El conjunto de documentos de entrenamiento debe estar previamente clasificado, para poder ser usado en el proceso de aprendizaje o entrenamiento del clasificador, y así dicho clasificador aprenda a clasificar los nuevos documentos en las diferentes temáticas o categorías. Una vez obtenido el conjunto de documentos de entrenamiento, el siguiente paso en la construcción de un clasificador de textos consiste en transformar los documentos, de su formato inicial, a una representación adecuada para el algoritmo de aprendizaje, y en general para la tarea de clasificación.

Una colección de entrenamiento es una herramienta indispensable para los entrenadores en el proceso de categorización de los documentos automáticamente, ya que permite el cumplimiento de la fase de entrenamiento del clasificador de texto.

Actualmente existen grandes colecciones de entrenamiento que son utilizadas por parte de investigadores para el entrenamiento de los clasificadores, ejemplo de ellas se encuentran en el epígrafe 1.4.1 del capítulo 1 de la presente tesis, algunas de ellas no son de dominio público sino que tienen asociado un costo y restricciones para su uso.

En el presente capítulo se plantean los lineamientos para la creación de una colección de entrenamiento en español a partir de las categorías educación, salud y deporte. La confección de esta colección destinada a la clasificación por parte del proyecto FILPACON, de los documentos HTML visitados por la comunidad UCI, en diferentes temáticas, para regular los contenidos a los que acceden los usuarios, persigue como objetivo disminuir el margen de error a la hora de clasificar los textos partiendo de documentos HTML.

El presente capítulo se distribuye de la siguiente manera: En un primer momento de plantean las características que debe poseer una colección de entrenamiento para su correcta elaboración, y posteriormente para la obtención de un buen resultado. Seguidamente se expresan cada uno de los pasos realizados manual y automáticamente para la confección de la colección de entrenamiento, una vez creada la colección se plantea una propuesta de un modelo de representación de documentos para ser aplicado por parte del proyecto FILPACON.

3.1 CARACTERÍSTICAS QUE DEBE POSEER UNA COLECCIÓN DE ENTRENAMIENTO

1. Se deben definir de antemano las categorías en las que van a ser clasificados los documentos, en las cuales se insertan cada nuevo documento que llega a la colección. En el desarrollo de la presente tesis se definen las categorías deporte, salud y educación para el desarrollo de la colección.
2. Se debe disponer de un número elevado de textos o documentos. Cuanto mayor es el conjunto de datos clasificados, mayor es la información disponible y mejor resultara el aprendizaje, pues cuando las colecciones de entrenamiento son pequeñas pueden producirse errores al estimar la probabilidad de que un determinado documento

pertenezca a una determinada clase, puede ser que un determinado término no aparezca en la colección, pero si aparezca en los documentos a categorizar.

3. Se deben transformar los documentos de la colección de su formato inicial, a una representación más adecuada para el algoritmo de aprendizaje que se vaya a utilizar posteriormente, y en general para la tarea de clasificación.
4. Reducir el tamaño del texto de los documentos de la colección, eliminando etiquetas html, signos de puntuación, palabras vacías entre las que se encuentran las preposiciones, conjunciones, así como eliminar sufijos y afijos.
5. Las colecciones son dependientes del idioma.
6. Las colecciones son sensibles a los errores de ortografía, términos diferentes constituyen componentes del vector diferentes.
7. El número de documentos por categorías o clases que conforman la colección, puede ser aproximadamente similar, aunque no exactamente el mismo.
8. En las colecciones de entrenamiento un documento puede estar presente en varias categorías a la vez.
9. Los términos (palabras) en la colección deben capturar lo máximo posible el significado de los textos, y permitir un aprendizaje eficiente y efectivo.

3.2 CONFECCIÓN DE LA COLECCIÓN

Para la confección de la colección los documentos fueron recopilados manualmente de diferentes sitios de Internet, en los cuales el contenido de las páginas se refería a términos propios de la salud, la educación y el deporte.

La cantidad de información que se puede encontrar en la web requiere un excelente servicio de búsqueda para que dicha información sea útil. Sin una poderosa herramienta de búsqueda, encontrar un sitio web específico puede ser muy difícil, es por ello que para el proceso de selección de las posibles páginas que conformarían la colección fueron utilizados en mayor o menor medida el Open Directory Project (ODP), Altavista y Google.

ODP o Dmoz como también es conocido, es el directorio más completo de la web, en esta base datos los contenidos están clasificados en diferentes categorías lo que permite una búsqueda rápida y fácil, además su acceso y uso es libre.

Altavista es rápido y fácil de usar, permite elegir el idioma en que se desee buscar, el número de búsquedas es mayor que en muchos otros buscadores, obtiene la fecha de cada archivo, además el motor aplica una lógica exhaustiva permitiendo que la búsqueda avanzada de resultados más exactos.

Google comprende más de 8.000 millones de direcciones URL, entra dentro de los mejores buscadores, constituye la colección más detallada de las páginas más útiles de Internet, produce resultados que contienen todos los términos de la búsqueda en el texto de la página o en los vínculos que le apuntan, extrae fragmentos de texto de los resultados que coinciden con su consulta. Esta característica ahorra el tiempo y la frustración de descargar una página irrelevante.

En un primer momento la colección contaba con un total de 1470 páginas. Después de un estudio exhaustivo de cada una de las páginas que conformaban la colección fueron filtradas manualmente todas aquellas que no se encontraban en el idioma español en toda su totalidad, fueron seleccionados aquellos documentos HTML que contenían textos con mayor contenido semántico, por tanto el número de documentos que conformaban la colección se redujo a 1203 páginas HTML.

Una vez escogidas las páginas que conformarían la colección estas fueron transformadas de formato html a un formato de simple almacenamiento del texto (txt). Procediendo al siguiente paso que consistió en el procesamiento de todo el conjunto de documentos que forman la colección.

3.3 PRE-PROCESAMIENTO DE LOS DOCUMENTOS DE LA COLECCIÓN

El propósito de esta etapa es reducir el tamaño de los documentos y aumentar la velocidad de proceso, eliminando las partes de los textos que no son relevantes, que no dicen nada sobre su contenido. Las tareas realizadas en cada uno de los documentos fueron las siguientes:

- *Eliminación de etiquetas HTML*. Los documentos son páginas Web y las etiquetas HTML no proporcionan información útil a la tarea de clasificación.

- *Eliminación de imágenes, videos, etc*. La colección de entrenamiento está creada para categorizar texto, todo lo que compone la página que no sea texto, no aporta información útil para la clasificación.
- *Eliminación de números y otros caracteres*. Son términos que no son significativos a la hora de asignarle una categoría al documento.
- *Revisión de la ortografía*. Palabras distintas constituyen términos distintos y por tanto componentes del vector diferentes, lo que influiría en los resultados obtenidos.

- *Eliminación de las palabras vacías*. Para reducir los términos que tienen poca capacidad semántica. (**Anexo # 2**).
- *Eliminación de las mil palabras más usadas en español*. Estas palabras luego de un estudio se estimó que tienen gran probabilidad de aparecer en la mayoría de los textos, por lo que poseen poco poder discriminativo, pues no diferencian un documento de otro. (**Anexo # 3**).
- *Eliminación de los acentos*.
- *Cambio de mayúsculas por minúsculas*. Para unificar las palabras y reducir así el número de términos de la colección.
- *Aplicación de un s-stemmers*. Para unificar términos terminados en plural.

3.4 INDEXADO

Obtenido el conjunto de entrenamiento y ya procesados los textos, el próximo paso a realizar consiste en transformar los documentos, de su formato inicial a una representación más adecuada para el algoritmo de aprendizaje.

Las páginas pueden ser representadas usando el modelo de espacio vectorial, representando el documento como vectores de pesos de términos. Este modelo es la representación más usada hoy en día, es muy efectivo en tareas de asignación de categorías a un documento en función de su contenido, es simple y rápido.

VSM está basado en el principio de independencia, considerando independientes entre si las palabras dentro de un mismo documento, esta suposición reduce la complejidad de cómputo brindando resultados aceptables. Esta representación permite un aprendizaje preciso, pues las palabras aisladas concentran una gran parte del significado del texto.

El VSM es aplicable a toda clase de idiomas siempre que puedan extraerse palabras de los documentos (sencillo en muchos idiomas pero más complejos en otros). El cálculo de los pesos de los términos se realiza a partir de datos extraíbles directamente de los textos sin ningún tipo de conocimiento lingüístico.

En este modelo cada documento se representa como una combinación lineal de vectores de términos, donde cada página se considera un vector, y cada término que aparece en al menos un documento será un componente del vector, asignándole un peso a cada término.

En el modelo vectorial se intenta recoger la relación de cada documento de la colección de documentos, con el conjunto de las características de la colección. Donde un documento puede considerarse como un vector que identifica en que grado el documento satisface cada una de las características. La representación conjunta de todos los documentos de la colección se realiza mediante una matriz denominada matriz término/documento.

Seleccionado el conjunto de términos caracterizadores de la colección de documentos, es necesario obtener el valor de cada elemento del vector del documento. El caso más simple es utilizar una aproximación binaria, de forma que si en el documento aparece el término su valor sería 1, en caso contrario sería 0. Como una palabra puede aparecer más de una vez en el mismo documento, y unas palabras pueden considerarse más significativas que otras, los componentes del vector obedecen cálculos más complejos que la asignación binaria.

Para el cálculo del peso de cada término en el vector se puede partir de dos ideas en cierto sentido contrapuestas: si un término aparece mucho en un documento, es importante para caracterizar ese documento. Pero si aparece en muchos documentos de la colección, no es beneficioso para distinguir un documento de los demás, dado su escaso poder discriminatorio.

Para determinar la capacidad de representación de un término para un documento dado, se computa el número de veces que aparece en dicho documento, obteniéndose la frecuencia del término en el documento, tf (Luhn, 1957). Los términos que más se repiten en un documento son, en un principio, más relevantes que los que se emplean menos. Por otra parte, si la frecuencia de un término en toda la colección de documentos es extremadamente alta, se opta por eliminarlo del conjunto de términos de la colección. Esto

es lo que se conoce como frecuencia inversa del documento (idf) (Spärck-Jones, 1972). Si un término aparece en muchos documentos no es útil para distinguir el documento de los otros de la colección. Los términos más frecuentes en la colección serán menos relevantes que los más raros.

Así, para calcular el peso de cada elemento del vector que representa al documento se tiene en cuenta la frecuencia inversa del término en la colección, combinándola con la frecuencia del término dentro de cada documento (tf x idf). De este modo se pondera con un valor más alto a los términos de baja frecuencia global, que aparecen poco en el conjunto de documento, pero de alta frecuencia local, que aparecen en un solo documento; se ponderan con valores bajos a los términos muy abundantes globalmente. En la ponderación tf x idf cada componente consta de dos partes, una local, que indica la importancia del término en cada documento (que variaría con la frecuencia en cada uno), y otra global, que indica la importancia del término en el total de la base (que sería la misma para todas las coordenadas de un término).

El proceso realizado para los documentos de la colección también puede aplicarse a los documentos a clasificar. Así pues, el mecanismo de obtención de pesos también se aplica a los documentos a clasificar, para de esta manera poder disponer de representaciones homogéneas de los documentos de la colección y los que están en proceso de clasificación, que posibiliten obtener el grado de similitud entre ambas representaciones.

VSM permite calcular una función de similaridad dada por el vector del documento a clasificar y los de cada uno de los que conforman la colección. El resultado de dicho cálculo mide la semejanza entre el documento y cada uno de los de la colección. La función de similitud más utilizada es el coseno del ángulo formado por ambos vectores. Dicha función aplica un factor de normalización para obviar los efectos de la disparidad en los tamaños de los vectores.

Representado un documento se debe de reducir el número de vectores originales que lo conforman, reduciendo la dimensionalidad del conjunto de entrenamiento, pues una alta dimensionalidad provoca que los resultados del clasificador lleguen a ser poco confiables debido a las deficiencia en los datos de entrenamiento al tomar en cuenta palabras con poca información del dominio.

Como función de reducción de dimensionalidad se puede considerar la *ganancia en la información* (GI), la cuál calcula la diferencia entre la entropía del sistema contra la entropía de cada palabra. Esta diferencia, medida en bits, indica que tan relevante y cuanta información aporta la palabra como factor determinante para llevar a cabo la clasificación. Evaluaciones presentadas han revelado que este método se encuentra entre los más efectivos (Yang, 1997). Esta función permite eliminar hasta un 99% de los términos originales, logrando un importante aumento de la eficiencia y la efectividad.

CONCLUSIONES

Fijadas las características de debe poseer una colección de entrenamiento, el siguiente paso consiste en la confección de la colección para las categorías deporte, educación y salud. Obtenidas las páginas que conformarían el conjunto de entrenamiento, el paso a seguir fue la creación de los vocabularios con los que representar los documentos HTML, empleando entre varias técnicas, funciones de reducción de dimensionalidad, o funciones de reducción de rasgos. En esta fase es transformada la información textual, de carácter cualitativo, ya que envuelve el tratamiento de significados de las palabras.

En este capítulo se ha propuesto un modelo de representación, para ser aplicado sobre el conjunto de documentos de entrenamiento. La obtención de una buena representación de las páginas web es la clave para obtener un buen resultado en el proceso de categorización de los diferentes contenidos de Internet. Pues la calidad de una representación se evalúa en función de los resultados que se obtengan tras su aplicación.

Una vez definida la colección el siguiente paso es transformar la versión del texto a un documento vectorial que describa el contenido del mismo, con el fin de reducir la complejidad de los documentos y hacerlos más manejables para la posterior tarea de clasificación. A raíz de la investigación realizada el modelo de representación a ser empleado sobre la colección de documentos es el Modelo de Espacio Vectorial. Donde cada documento es representado mediante un vector de términos, y la colección forma una matriz de términos, en la que un término es la unidad mínima de información y lleva asociado un peso que expresa la importancia de dicho termino en el documento. Este peso puede calcularse basándose en la frecuencia de los términos, tanto dentro de cada documento en particular, como en toda la colección de documentos.

Para el cálculo del coeficiente (peso) que lleva asociado cada término se propone la combinación de la frecuencia del término (tf) con la frecuencia inversa del documento (idf) el producto tf x idf. Indicando así la importancia del término en cada documento y la importancia del término en todo el conjunto de entrenamiento.

CONCLUSIONES GENERALES

El hecho de que el carácter de Internet sea universal y que el acceso a sus contenidos se realice desde diversos ámbitos sociales y culturales, hace que el proceso de lectura y escritura en la red de redes sea cada vez mayor. El comportamiento que pueda tener un usuario que busca información en la red puede estar dado en dos vertientes, que bajo conocimiento acceda a contenidos inadecuados o ilícitos en dependencia del gobierno del país donde radica, o que acceda a dichos contenidos sin saber juzgarlos correctamente, en ambos casos es adecuada la opción de regular el acceso a los diversos contenidos de la web. Clasificando para ello las páginas web visitadas en diferentes temáticas, objetivo fundamental del proyecto FILPACON.

La categorización automática de documentos tiene una etapa de aprendizaje o entrenamiento en la que se obtiene información de las categorías en las que puede ser clasificado un documento dado. A partir de un número de documentos que resulte representativo para cada categoría (la colección de entrenamiento).

El objetivo propuesto en la investigación era crear una colección de entrenamiento bajo las categorías deporte, educación y salud que minimizara el margen de error a la hora de clasificar texto en documentos HTML. Para ello se hizo un estudio del marco teórico en el que se enmarcan las colecciones de entrenamiento, se analizaron diferentes colecciones ya conocidas en el campo de la categorización (algunas de ellas a raíz del auge que está teniendo la categorización tienen un precio asociado) así como trabajos en los que fueron desarrollados algunos conjuntos de entrenamiento, para identificar las características idóneas que influyen en la correcta creación de una colección de entrenamiento, valorando la influencia de dichas características para obtener un buen resultado en el proceso de clasificación de los documentos HTML visitados, para regular con un mínimo de margen de error el acceso a los diferentes contenidos de Internet.

Planteadas las características de las colecciones, fue elaborada la colección, y posteriormente fueron transformadas cada una de las páginas que conforman el conjunto de entrenamiento al formato txt, seguido cada texto fue sometido a la etapa de pre-procesamiento. Para que el conjunto de entrenamiento de un buen resultado dentro de varias características debe cumplir con que el número de documentos de la colección sea

elevado, para evitar posibles errores en el proceso de clasificación a la hora de comparar los términos de un documento a clasificar con los términos de los documentos de la colección y que algún término no se encuentre dentro del conjunto de entrenamiento. Antes de comenzar con la confección de la colección se deben definir el conjunto de categorías en los que van a ser clasificados los documentos, por parte de FLPACON los documentos se pueden clasificar en 16 categorías, las cuales a su vez se desglosan en subcategorías.

Creada la colección de entrenamiento el siguiente paso a seguir es la correcta representación de los documentos que conforman la misma, para ello se hizo un estudio y análisis de los diferentes modelos de representación textual, se formalizó la representación automática de textos, se describió el modelo de espacio vectorial y el índice de latencia semántica, modelos vectoriales más utilizados en la representación automática de documentos. Fueron mostradas diferentes funciones de ponderación (función F) tanto de carácter local como global, así como funciones de reducción de rasgos o dimensionalidad, que pueden en mayor o menor medida ser empleadas como funciones de ponderación. Además fueron analizadas las diferentes fases en la etapa de pre-procesamiento por las que debe pasar un documento antes de serle aplicado un modelo de representación.

A raíz de del estudio de los modelos de representación se plantea como propuesta, para la representación del contenido de cada documento de la colección de entrenamiento el modelo de espacio vectorial, modelo muy efectivo en tareas de asignación de categorías. Así mismo como función de ponderación se propone la función tf x idf.

En la actualidad los métodos que mejores resultados han dado son aquellos que se basan en la extracción de bolsas de palabras (bags of words en inglés), dichos métodos buscan determinar el contexto de un texto o un conjunto de ellos por medio de las palabras que se repiten con mayor frecuencia.

RECOMENDACIONES

Para trabajos futuros se plantea ampliar el número de clases de la colección, para lo cual deben ser recolectados más documentos HTML por categoría, el conjunto de estos documentos debe ser en mayor o menor medida de igual número de páginas por categoría, aunque no necesariamente el mismo. Se debe tratar que el texto sea lo mas extenso posible, para así poseer mayor información, deben ser determinadas con antelación las diferentes categorías que conformarán la colección.

Se plantea profundizar sobre el Modelo de Espacio Vectorial, modelo de representación propuesto para ser aplicado sobre el conjunto de documentos que conforman la colección. Así también se plantea seguir estudiando las colecciones para continuar el análisis de las características que debe poseer una colección de entrenamiento para que resulte más eficiente y así alcanzar mejores resultados a la hora de categorizar los textos en documentos HTML.

REFERENCIAS BIBLIOGRÁFICAS

Aas K, Eikvil L. 1999. *Text Categorization a Survey.* [En línea] 1999. [Citado el: 9 de diciembre de 2007.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:14GgdlytLxsJ:www.cs.ualberta.ca/~dunwei/TM%2520papers/text%2520mining/aas99text.ps+Text+Categorization+a+Survey

Arregi O, Fernández I. 2002. *Automatic Document Classification.* s.l. : Journal of the Association for Computing Machinery, 2002.

Chakrabarti, S. 2002. *Mining the web: discovering knowledge from hypertext data.* 2002.

Dagan, I. *Mistake-Driven Learning in Text Categorization.* [En línea] [Citado el: 6 de abril de 2008.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:wdVuaR04AlUJ:acl.ldc.upenn.edu/W/W97/W97-0306.pdf+Mistake-Driven+Learning+in+Text+Categorization.

Figuerola, C. 2000. *Categorización automática de documentos en español: algunos resultados experimentales.* [En línea] 2000. [Citado el: 3 de Diciembre de 2007.] http://209.85.215.104/search?q=cache:7Y0vuSB48LQJ:reina.usal.es/pub/figuerola2000categorizacion.pdf+Categorizaci%C3%B3n+autom%C3%A1tica+de+documentos+en+espa%C3%B1ol:+algunos+resultados+experimentales&hl=es&ct=clnk&cd=1&gl=cu.

Figuerola, C. 2001. *Automatic vs. Manual Categorisation of Documents in Spanish.* [En línea] 2001. [Citado el: 3 de Mayo de 2008.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:6owZNIO6Bj8J:reina.usal.es/pub/figuerola2001automatic.pdf+author:%22Figuerola%22+intitle:%22Automatic+vs+manual+categorisation+of+documents+in+Spanish%22+.

Francis W, Kucera H. 1979. *BROWN COPUS MANUAL. Textos de especialidad y comunidades discursivas técnico-profesionales: una aproximación basada en corpus computarizado .* [En línea] 1979. [Citado el: 23 de Abril de 2008.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:znnaapgbDqYJ:mingaonline.uach.cl/scielo.php%3Fscript%3Dsci_arttext%26pid%3DS0071-17132004000100001%26lng%3Des%26nrm%3Dis.+BROWN+CORPUS+MANUAL.

Gómez, J. 2003. *Categorización de texto sensible al coste para el filtrado de contenidos inapropiados en Internet.* [En línea] 2003. [Citado el: 18 de febrero de 2008.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:vTmVPnunejQJ:www.sepln.org/revistaSEPLN/revista/31/31-Pag13.pdf+Categorizaci%C3%B3n+de+texto+sensible+al+coste+para+el+filtrado+de+contenidos+inapropiados+en+Internet.

He J, Tan A, Tan C. 2003. *On Quantitative Evaluation of Clustering Systems.* [En línea] 2003. [Citado el: 23 de Diciembre de 2007.] http://books.google.com/books?hl=es&lr=&id=WJsd7Mz7zJEC&oi=fnd&pg=PA105&dq=On+Quantitative+Evaluation+of+Clustering+Systems&ots=XqRTUKi5-e&sig=hGibBbVV92MHl36acE2OcZJFbrM#PPA278,M1.

Hersh, W. 1994. *OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research.* [En línea] 1994. [Citado el: 25 de Noviembre de 2007.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:8x_cDFxGfBMJ:www.billhersh.info/sigir-94-ohsumed.pdf+.

Koll, M. 1979. *WEIRD: An approach to concepr-based information retrieval.* 1979.

Kwan, Y. 2006. *Desafíos en la clasificación automatizada utilizando sistemas de clasificación bibliotecarios.* [En línea] 2006. [Citado el: 21 de febrero de 2008.] http://64.233.169.104/search?q=cache:b0OX89Q8lYIJ:www.ifla.org/IV/ifla72/papers/097-Yi_trans-es.pdf+Desaf%C3%ADos+en+la+clasificaci%C3%B3n+automatizada+utilizando+sistemas+de+clasificaci%C3%B3n+bibliotecarios&hl=es&ct=clnk&cd=1&gl=cu.

Landauer, T K. 1998. *An introduction to latent semantic analysis.* [En línea] 1998. http://lsa.colorado.edu/papers/dp1.LSAintro.pdf.

Lewis, D David. 1996. *Training algorithms for linear text classifiers.* s.l. : In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996

Luhn, H. 1957. *A statistical approach to mechanized encoding and searching of literaty information.* [En línea] 1957. [Citado el: 30 de Marzo de 2008.]

http://64.233.169.104/search?q=cache:2IrBnWuMuZcJ:www.research.ibm.com/journal/rd/014/ibmrd0104D.pdf+A+statistical+approach+to+mechanized+encoding+and+searching+of+literaty+information&hl=es&ct=clnk&cd=1&gl=cu.

Luhn, H. 1953. *A new method of recording and searching information.* s.l. : American Documentation, 1953.

Manning C, Schutze H. 1999. *Foundations of Statistical Natural Language Processing.* [En línea] 1999. [Citado el: 4 de Febrero de 2008.] http://64.233.169.104/search?q=cache:vlftAU0Uti8J:nlp.stanford.edu/fsnlp/+Foundations+of+Statistical+Natural+Language+Processing&hl=es&ct=clnk&cd=1&gl=cu MIT Press.

Maron, M. 1961. Automatic indexing: an experimental inquery. [En línea] 1961. [Citado el: 21 de marzo de 2008.] http://64.233.169.104/search?q=cache:sRMeZUlJhxMJ:stinet.dtic.mil/oai/oai%3Fverb%3DgetRecord%26metadataPrefix%3Dhtml%26identifier%3DAD0607467+Automatic+indexing:+an+experimental+inquery&hl=es&ct=clnk&cd=3&gl=cu.

Miller, G. 1993. *A Semantic concordance.* [En línea] 1993. [Citado el: 24 de marzo de 2008.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:1LHHs8sRyGMJ:acl.ldc.upenn.edu/H/H93/H93-1061.pdf+author:%22Miller%22+intitle:%22A+semantic+concordance%22+.

Chakrabarti, S. 2002. *Mining the web: discovering knowledge from hypertext data.* 2002.

Peláez J.I, La Red D, Sánchez P. 2002. *Un Clasificador de Texto Por Aprendizaje.* [En línea] 2002. [Citado el: 13 de enero de 2008.] http://cabrillo.lsi.uned.es:8080/aepia/Uploads/15/103.pdf.

Rocchio, J. 1971. *Relevance feedback in information retrieval.* s.l. : The SMART Retrieval System. Experiments in Automatic Document Processing, 1971.

Salton, G. 1983. *Introduction to Modern information RetrievaL.* s.l. : McGraw Hill, 1983.

Sebastiani, S. 1999 . *Machine Learning in Automated Text .* [En línea] 1999 . [Citado el: 20 de enero de 2008.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:p8j3r7vDQHYJ:www.softlab.ntua.gr/facilities/public/.

Sinka, M. 2002. *A large benchmark dataset for web document clustering.* [En línea] 2002. [Citado el: 24 de Noviembre de 2007.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:U3MiX5-yE3sJ:www.cs.reading.ac.uk/common/publications/02067.pdf+A+large+benchmark+dataset+for+web+document+clustering.

Spärck-Jones, Karen. 1972. *From Wikipedia, the free encyclopedia.* [En línea] 1972. [Citado el: 10 de Mayo de 2008.] http://en.wikipedia.org/wiki/Karen_Sp%C3%A4rck_Jones.

Téllez A. 2003. *Clasificación Automática de Textos de Desastres Naturales en México.* [En línea] 2003. [Citado el: 28 de enero de 2008.] http://66.102.1.104/scholar?hl=es&lr=&q=cache:VMcbma3MCAkJ:ccc.inaoep.mx/~mmontesg/publicaciones/2003/ClasificacionDesastres-CIIC03.pdf+Clasificaci%C3%B3n+Autom%C3%A1tica+de+Textos+de+Desastres+Naturales+en+M%C3%A9xico.

Yang, Y. 1997. *Feature Selection in Statistical Learning of Text Categorization.* s.l. : Proceedings of the 14th International Conference on Machine Learning, 1997

BIBLIOGRAFÍA

Combarro E, Montañés E, Diaz I, Cortina R, Alonso P, Ranilla J. 2007. *Un Framework para la Gestión Documental en las Administraciones Públicas.* [En línea] 2007. [Citado el: 5 de abril de 2008.] http://www.tecnimap.com/documentos/Departamentos/Coordinacion/Tecnimap/Comunicaciones%20definitivas/TCO-265-2007DC/Comunicaci%F3n%20TCO-265-2007DC.pdf.

Dumais S, Platt J, Heckerman D, Sahami M. 1998. *Inductive learning algorithms and representations for text categorization.* [En línea] 1998. [Citado el: 6 de marzo de 2008.] http://research.microsoft.com/~jplatt/cikm98.pdf.

Eito R, Senso J. 2004. *Minería textual.* [En línea] 2004. [Citado el: 6 de abril de 2008.] http://eprints.rclis.org/archive/00013341.

Figuerola C, Alonso J, Zazo A, Rodríguez E. 2004. *Algunas Técnicas de Clasificación Automática de Documentos.* [En línea] 2004. [Citado el: 5 de frebrero de 2008.] http://multidoc.rediris.es/cdm/viewarticle.php?id=28&layout=html.

Figuerola C, Zazo A, Alonso J. 2002. *La interacción con el usuario en los sistemas de recuperación de información: realimentación por relevancia.* [En línea] 2002. [Citado el: 10 de diciembre de 2007.] http://wotan.liu.edu/dois/data/Articles/juliccrcgy:2002:v:8:i:1:p:87-94.html.

Fresno, V. 2006. *Representación Autocontenida de Documentos HTML: una propuesta basada en Combinaciones Heurísticas de Criterios.* [En línea] 2006. [Citado el: 16 de marzo de 2008.] http://www.escet.urjc.es/~vfresno/tesis_VFresno.pdf.

Gascón A, De la Puente M, Rodríguez M. 2003. *Clasificación Jerárquica de contenidos Web.* [En línea] 2003. [Citado el: 25 de febrero de 2008.] http://www.nosolousabilidad.com/articulos/clas_facetadas1.htm.

Gayo, D. 2005. *BlindLight. Una nueva técnica para procesamiento de texto no estructurado mediante vectores de n-gramas de longitud variable con aplicación a diversas tareas de tratamiento de lenguaje natural.* [En línea] 2005. [Citado el: 11 de mayo de 2008.]

http://www.di.uniovi.es/~dani/downloads/blindlight-dgayo-disertacion-capitulo8-conclusiones.pdf.

Gómez J, Cortizo J, Puertas E, Buenaga M. 2004. *Experimentos en indexación conceptual para la categorización de texto.* [En línea] 2004. [Citado el: 9 de enero de 2008.] http://www.ainetsolutions.com/jccp/papers/ciawi04a.pdf.

Ismael R, Ponce M, José A, Zárate M, Juan C, Olivares R. 2006. *Recuperación de Información de Páginas Web mediante una Ontología que es poblada usando Clasificación Automática de Textos.* [En línea] 2006. [Citado el: 22 de febrero de 2008.] http://www.computer.org/portal/cms_docs_ieeecs/ieeecs/Communities/students/looking/2006summer/articulo06.pdf.

Martín M, Montejo A, Díaz M, Ureña A. 2007. *Integración de conocimiento en un dominio específico para categorización multietiqueta.* [En línea] 2007. [Citado el: 8 de marzo de 2008.] http://www.sepln.org/revistaSEPLN/revista/39/08.pdf.

Peláez J.I, La Red D, Sánchez P. 2002. *Un Clasificador de Texto Por Aprendizaje.* [En línea] 2002. [Citado el: 13 de enero de 2008.] http://cabrillo.lsi.uned.es:8080/aepia/Uploads/15/103.pdf.

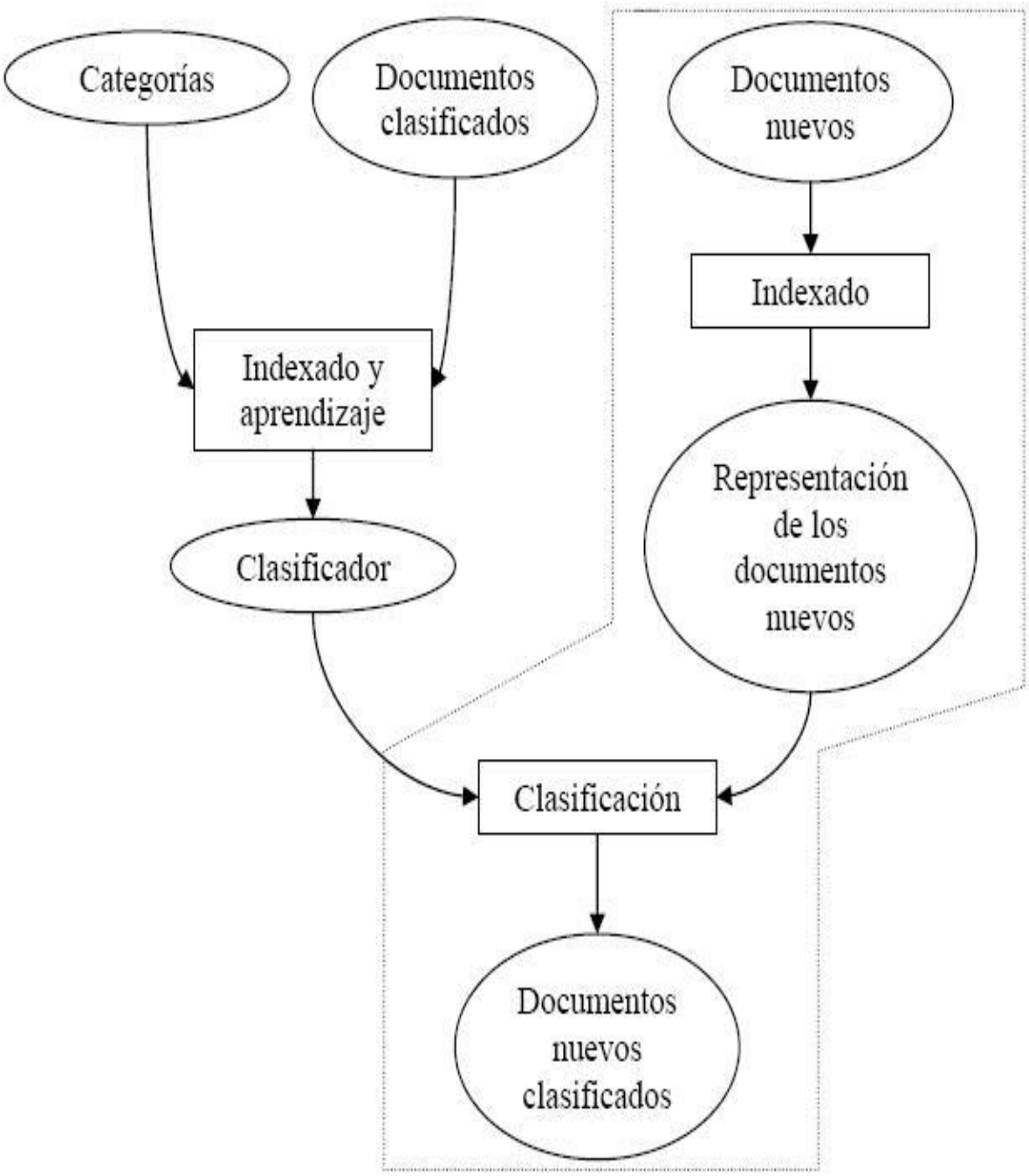
Téllez A, Montes M, Villaseñor L. 2004. *Aplicando la Clasificación de Texto en la Extracción de Información.* [En línea] 2004. [Citado el: 13 de diciembre de 2007.] http://ccc.inaoep.mx/~mmontesg/publicaciones/2004/IEconClasificacion-tallerENC04.pdf.

Téllez, A. 2005. *xtracción de Información con Algoritmos de Clasificación.* [En línea] 2005. [Citado el: 10 de diciembre de 2007.] http://ccc.inaoep.mx/~mmontesg/tesis%20estudiantes/TesisMaestria-AlbertoTellez.pdf.

Ureña A, Buenaga M. 1999. *Utilizando WordNet para Complementar la Información de Entrenamiento en la Identificación del Significado de las Palabras.* [En línea] 1999. [Citado el: 9 de diciembre de 2007.] http://cabrillo.lsi.uned.es:8080/aepia/Uploads/7/169.pdf.

Ureña A, García M, Buenaga M, Gómez J. 1998. *Resolución Automática de la Ambigüedad Léxica fundamentada en el Modelo del Espacio Vectorial usando Ventana Contextual Variable.* [En línea] 1998. [Citado el: 9 de enero de 2008.] http://www.esi.uem.es/~jmgomez/papers/aesla98.pdf.

ANEXOS



Anexo # 1. Visión general de los procesos de clasificación de texto

a	añadió	aún	actualmente	ayer
adelante	además	afirmo	agrego	ahí
ahora	al	algún	algo	alguna
algunas	alguno	algunos	alrededor	ambos
ante	anterior	antes	apenas	aproximadamente
aquí	así	aseguro	aunque	bajo
bien	buena	buen	buenas	bueno
buenos	cómo	cada	casi	cerca
cierto	cinco	comentó	como	con
conocer	consideró	considerara	contra	cosas
creo	cual	cuales	cualquier	cuando
cuanto	cuatro	cuenta	da	dado
dan	dar	de	debe	deben
debido	decir	dejó	del	demás
dentro	desde	después	dicen	dice
dicho	dieron	diferente	diferentes	dijeron
dijo	dio	donde	dos	durante
él	ésta	éstas	éste	éstos
e	ejemplo	estos	el	ella
ellas	ello	ellos	embargo	en
estuvo	encuentra	entonces	entre	era
eran	es	esa	existe	esas
ese	esos	eso	está	están
esta	explicó	estaba	estaban	estamos
estar	estará	estas	este	expreso
fin	fue	fuera	fueron	gran
grandes	ha	había	habían	haber
habrá	hace	hacen	hacer	hacerlo
hacia	haciendo	han	hasta	hay
haya	he	hecho	hemos	hicieron
hizo	hoy	hubo	incluso	indicó
informó	igual	junto	la	lado

las	los	le	les	llegó
luego	lleva	llevar	lo	lugar
más	manera	manifestó	mayor	muy
me	mediante	mejor	mencionó	menos
mi	mientras	misma	mismas	mismo
mismos	momento	mucha	muchas	mucho
muchos	nada	nadie	ni	ningún
ninguna	ningunas	ninguno	ningunos	no
nos	nosotras	nosotros	nuestra	nuestras
nuestro	nuestros	nueva	nuevas	nuevo
nuevos	nunca	o	ocho	otra
otras	otro	otros	para	parece
parte	partir	pasada	primero	primeros
pueden	pasado	pero	pesar	poca
pocas	principalmente	propia	pues	poco
pocos	podemos	podrá	podrán	propias
propio	podría	podrían	poner	por
porque	propios	pudo	posible	próximo
próximos	primer	primera	pueda	puede
qué	que	quedó	queremos	quién
quien	quienes	quiere	realizó	realizado
realizar	respecto	sí	solo	se
señaló	sea	solamente	su	sean
según	segunda	segundo	seis	sola
sus	ser	será	serán	sería
si	solo	sido	siempre	siendo
siete	sigue	solos	siguiente	sin
sino	sobre	sola	son	tal
también	tampoco	tan	tanto	tres
tenía	tendrá	tendrán	tenemos	tener
tuvo	tenga	tengo	tenido	tercera
tiene	tienen	toda	todas	todavía

todo	todos	total	tras	trata
través	un	una	unas	uno
unos	usted	última	últimas	último
últimos	va	vamos	van	varias
varios	veces	ver	vez	y
ya	yo			

***Anexo # 2. Lista de palabras vacías. Stop-words*.**

de	la	el	que	en
y	a	los	del	por
las	se	un	con	para
una	efe	al	su	no
ha	es	hoy	como	más
lo	sus	entre	este	presidente
sobre	dos	esta	según	fue
gobierno	han	años	pero	millones
también	ciento	dijo	país	desde
ya	pasado	partido	contra	o
tras	España	Madrid	le	si
sin	año	hasta	general	está
durante	tres	parte	estado	ser
ante	ministro	cuando	son	sido
grupo	aunque	nacional	países	después
tiene	primera	donde	equipo	fuentes
uno	otros	dólares	acuerdo	quien
porque	sólo	muy	primer	estados
unos	e	próximo	EEUU	ex
mientras	José	hace	todo	personas
todos	mañana	gran	paz	internacional
fueron	política	unidos	cuatro	ciudad
Barcelona	días	español	además	antes

nuevo	nueva	mundo	hay	había
día	puede	será	ese	ahora
prensa	señaló	policía	vez	seguridad
capital	semana	mundial	caso	afirmó
mayor	elecciones	final	frente	consejo
cinco	comisión	situación	informó	pesetas
Juan	esa	reunión	así	Carlos
mar	están	mismo	ONU	medio
horas	estos	española	secretario	congreso
argentina	México	meses	otras	san
tiempo	otro	real	debe	zona
haber	guerra	momento	Luis	domingo
segunda	cada	Rusia	embargo	Organización
Bosnia	menos	defensa	ni	partidos
ayer	añadió	autoridades	fuerzas	trabajo
jefe	empresa	director	poder	europa
aseguró	tanto	lunes	hacer	miembros
mes	Brasil	hecho	ellos	europea
proceso	algunos	puntos	lugar	últimos
portavoz	forma	indicó	unión	explicó
militar	era	viernes	centro	seis
estadounidense	fin	norte	ministerio	Italia
miércoles	número	fútbol	otra	mejor
estas	ley	chile	varios	martes
m	Francia	Colombia	me	respecto
les	jueves	civil	problemas	total
banco	relaciones	club	vida	liga
desarrollo	actual	decisión	parlamento	sur
kilómetros	segundo	asuntos	Antonio	diario
bien	jugadores	militares	declaró	tienen
copa	agregó	hizo	derechos	negociaciones
plan	oficial	líder	mercado	jornada

ejército	hacia	encuentro	sino	empresas
proyecto	social	crisis	unas	puerto
bajo	Israel	mayoría	casi	González
posible	apoyo	Washington	económica	programa
exteriores	central	justicia	ambos	comercio
declaraciones	Cuba	aún	tarde	Clinton
último	cuenta	julio	jun	nov
anunció	resultados	poco	naciones	María
él	sábado	Manuel	jugador	importante
oct	diez	todas	sector	may
población	sea	junto	sistema	político
valencia	tuvo	república	selección	nos
puesto	visita	españoles	informaron	medidas
casa	jul	trabajadores	políticos	abr
enero	sociedad	interior	temporada	economía
público	muerte	debido	dentro	sep
comité	tribunal	feb	representantes	tener
siete	hora	García	grupos	Alemania
fiscal	juego	conferencia	ene	podría
toda	siempre	campaña	electoral	medios
falta	manifestó	destacó	informe	dic
tipo	grandes	cargo	posibilidad	juez
especial	comunidad	junio	federación	espera
problema	china	asociación	Sevilla	minutos
comunicado	futuro	ayuda	cambio	victoria
ago	alto	Miguel	misma	va
argentino	sede	control	pueden	mucho
mayo	partir	ello	estaba	región
última	información	ministros	televisión	ocho
costa	equipos	considera	condiciones	armas
presidencia	primeros	derecho	mi	propuesta
local	servicios	marzo	estar	humanos

diciembre	varias	principal	pesar	socialista
metros	cual	torneo	cerca	encuentra
popular	dirección	económico	llegó	septiembre
técnico	york	pues	fuerza	opinión
candidato	administración	octubre	oposición	investigación
presencia	participación	noche	serán	dar
servicio	subrayó	punto	juicio	fuera
campo	dinero	cooperación	serbios	premio
relación	oficiales	Venezuela	campeonato	ruso
pueblo	libertad	francisco	cualquier	Moscú
ningún	próxima	operación	nada	objetivo
vuelta	muchos	departamento	Santiago	sentido
localidad	pasada	diputados	América	abril
cumbre	llegar	base	Fernando	tan
todavía	primero	lucha	favor	Javier
principales	entrenador	evitar	internacionales	asamblea
buenos	casos	guardia	conjunto	haya
historia	habían	compañía	ejecutivo	pública
mostró	cabo	interés	cámara	mujeres
tropas	clasificación	conflicto	quienes	pese
francés	obras	previsto	agosto	pidió
unidas	violencia	deben	algo	mujer
tendrá	resultado	universidad	orden	OTAN
entonces	río	europeo	prueba	Pérez
datos	algunas	libre	reforma	nuevos
plazo	familia	noviembre	acciones	resto
Perú	ciudadanos	dirigentes	nivel	brasileño
París	cuyo	Uruguay	aumento	rey
declaración	hombres	división	serie	semanas
políticas	incluso	tenía	señala	Guatemala
cuya	Israelí	dado	votos	intervención
través	dice	paso	acuerdos	italiano

único	acción	febrero	tratado	campeón
junta	huelga	próximos	agencia	nombre
actualmente	posición	necesidad	Alberto	anterior
acto	instituto	federal	periodistas	comentó
movimiento	eso	prisión	doce	nuevas
partes	democracia	Sánchez	marcha	sería
rico	carrera	tercera	productos	corea
atentado	salvador	etapa	ecuador	fuerte
nunca	comercial	hombre	asimismo	creación
industria	plaza	causa	fuego	entrada
Portugal	rueda	fase	cuanto	producción
ninguna	niños	media	quiere	aires
Haití	vicepresidente	Sarajevo	salud	salida
finales	nueve	Bolivia	Felipe	jugar
deportivo	esos	asunto	tercer	recordó
Pedro	refugiados	jóvenes	responsable	soldados
propio	dirigente	da	mantener	exterior
izquierda	éste	triunfo	esp	personal
judicial	medida	hospital	afirma	trata
delegación	expresó	Panamá	crecimiento	firma
conversaciones	protección	alcalde	mexicano	roma
veces	precisó	conocer	máximo	solución
régimen	santa	aeropuerto	últimas	detenidos
diputado	buena	sigue	fondo	Fernández
título	conseguir	consideró	pueda	organismo
tal	ambiente	comenzó	reconoció	tenido
cultura	corrupción	vasco	representante	coalición
tráfico	heridos	territorio	detenido	debate
obra	ángel	accidente	palestinos	comunicación
ver	pide	liberación	Rodríguez	empleo
estuvo	agua	comicios	goles	parece
sí	zonas	ellas	sociales	estadio

existe	ataque	hechos	Zaragoza	responsables
alemán	ronda	extranjeros	dicho	van
largo	baloncesto	formación	locales	colombiano
audiencia	víctimas	principio	juegos	importantes
decir	buen	produjo	pudo	papel
organizaciones	radio	FAM	sectores	documento
Jorge	precios	fecha	superior	encuentran
muertos	funcionarios	estadounidenses	Japón	cuales
león	mantiene	fondos	agentes	mediante
recursos	hemos	expertos	luego	difícil
titular	provincia	isla	López	diálogo
especialmente	construcción	necesario	venta	calificó
ganar	yo	valor	unido	dio
oro	empresarios	manera	regional	joven
carta	asesinato	mas	lima	marco
entrevista	misión	esas	alta	operaciones
igual	lista	Banesto	generales	constitución
entidad	qué	seguir	alianza	esto
baja	petición	musulmanes	rica	supone
británico	logró	educación	cabeza	cruz
importancia	actividades	he	iniciativa	Bilbao
lograr	tierra	pruebas	Palestina	pretende
intención	mil	presentado	ciudades	estudio
calidad	Berlusconi	Yeltsin	ocasiones	gasa
viaje	miembro	embajador	dijeron	Martínez
actividad	intereses	frontera	once	primeras
confianza	participar	persona	anoche	nuestra
senado	estamos	Londres	sindicatos	resolución
éxito	poner	nuestro	Bruselas	cosas
manos	usa	socialistas	obstante	uso
ambas	Croacia	antigua	públicas	reino
94	1	2	000	10

3	11	12	5	6
4	15	20	16	18
17	14	13	7	0
19	21	30	8	22
23	09	25	9	07
08	03	06	24	01
05	27	1993	28	26
29	00	50	31	1992
40	1995	35	45	200
1991	100	32	60	33
34	70	1990	36	48
38	46	37	300	42
55	44	80	49	39
47	43	41	51	52

***Anexo # 3. Lista de palabras más usadas en el español*.**

GLOSARIO DE TÉRMINOS

A continuación se incluye una breve explicación sobre algunos conceptos utilizados en el presente trabajo, a fin de facilitar la comprensión por parte de lectores no demasiado especializados en estos temas:

Aprendizaje automático: El aprendizaje automático es la disciplina que estudia cómo construir sistemas computacionales que mejoren automáticamente mediante la experiencia. En otras palabras, se dice que un programa ha "aprendido" a desarrollar la tarea si después de proporcionarle la experiencia E el sistema es capaz de desempeñarse razonablemente bien cuando nuevas situaciones de la tarea se presentan.

Anchortexts: El "anchortexts" es el texto que aparece subrayado, el texto en el que hacemos clic para seguir un enlace, aunque a veces, según el diseño de las webs, es posible que no salga subrayado, pues el diseñador puede modificar el aspecto gráfico de los enlaces.

Archivos log: Los archivos log registran todos los accesos a su alojamiento, guardando entre otras, información acerca de la **IP** desde la que se ha realizado la conexión, fecha, hora, archivos y/o página a la que se accede, etc.

Bag of Words: Bolsa de palabras.

Base de Datos: Es un conjunto de informaciones que tratan sobre un dominio determinado, acumuladas, clasificadas y estructuradas de modo homogéneo en un sistema provisto de índices y facilidades para la búsqueda y explotación de dichos datos.

Categorización automática: Término que hace referencia a una amplia variedad de técnicas que tienen como objetivo asignar a un objeto dado una o más categorías (o etiquetas) de un conjunto predefinido. La categorización automática de documentos se realiza a partir del texto de los mismos y requiere una fase previa de entrenamiento durante la cual el categorizador es enfrentado con unos pocos ejemplos de las distintas categorías que debe reconocer.

Clustering: Un algoritmo de agrupamiento (en inglés, clustering) es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Un algoritmo de clustering permite extraer representantes de un conjunto de datos, que pueden ser posteriormente usados para transmisión, para eliminación de ruido o con una fase posterior de calibración, para clasificación de vectores en diferentes conjuntos.

Conocimiento a priori: Viene del término en latín "lo que precede".

Corpus: Es una colección, generalmente muy grande, de documentos (típicamente textos o habla transcrita) que muestran el uso real de una lengua natural que puede contener muestras de un único lenguaje (corpus monolingüe) o de varios lenguajes (corpus multilingüe).

Co-citaciones: Es un método que se emplea para establecer el grado de similitud entre dos documentos. Si el documento A y B son citados por el documento C, probablemente estén relacionados, a pesar de que no se citen mutuamente. Cuanto mayor sea el número de documentos que citen conjuntamente a los documentos A y B, mayor grado de similitud o relación habrá entre ellos.

Correlaciones: Es un grupo de técnicas estadísticas usadas para medir la fuerza de la relación entre dos variables.

Dirección IP: Representación numérica de la localización de un ordenador dentro de una red. Consiste en cuatro números de hasta 3 cifras separados por puntos pues todos los equipos informáticos conectados a Internet disponen de un número exclusivo, denominado dirección del Protocolo de Internet (IP).

Directorio: En informática, un directorio es una agrupación de archivos de datos, atendiendo a su contenido, a su propósito o a cualquier criterio que decida el usuario. Técnicamente el directorio almacena información acerca de los archivos que contiene: como los atributos de los archivos o dónde se encuentran físicamente en el dispositivo de

almacenamiento. En fin es un conjunto de ficheros agrupados bajo un mismo nombre, lo que facilita su utilización y administración.

Espacio Euclídeo: Un espacio euclídeo es un espacio vectorial normado de dimensión finita en que la norma es heredada de un producto escalar. El espacio euclídeo es el espacio matemático n-dimensional usual, una generalización de los espacios de 2 (plano euclídeo) y 3 dimensiones estudiados por Euclides.

Euskara: El euskera, vascuence, vasco, éuscaro o lengua vasca es un idioma que fue creado a partir del año 1968, impulsado por la necesidad de proporcionar a los hablantes una norma unificada para el registro culto, dada la inviabilidad de publicar en cada uno de los dialectos. Es una lengua de tipología aglutinante y genéticamente aislada, es decir, no muestra un origen común claro con otras lenguas, lo que ha llevado a diversas hipótesis científicas.

Función Kernel: Es una función de distancia que se usa para determinar el peso de cada ejemplo de entrenamiento.

Hiperplano: En un espacio de una única dimensión (como una recta), un hiperplano es un punto; divide una línea en dos líneas. En un espacio bidimensional (como el plano xy), un hiperplano es una recta; divide el plano en dos mitades. En un espacio tridimensional, un hiperplano es un plano corriente; divide el espacio en dos mitades.

Internet: Es un espacio virtual creado por dispositivos electrónicos de computación, redes de comunicación y redes de ordenadores. Es a la vez un lugar en el que se puede buscar información y recursos, un soporte para transportar dichos recursos, un medio para comunicarse con personas y un entorno en el que se pueden formalizar interacciones y hacer transacciones (negocios). Es, cada vez más, una buena fuente de información para actividades de vigilancia tecnológica y del entorno de una empresa. Sinónimos: WWW; World Wide Web; Ciberespacio; Web; W3; Telaraña o Malla Mundial.

Lematización: Se trata de la operación de hallar el lema del cual se deriva una palabra concreta. En este contexto, su finalidad es agrupar palabras que tienen un contenido semántico muy próximo; por ejemplo: *biblioteca, bibliotecas, o bibliotecario*. Esta agrupación

permite comparar o encontrar directamente todas las palabras que tienen el mismo lema, independientemente de que el usuario busque sólo por una de ellas; además, corrige los cálculos de frecuencias y demás, contabilizando lemas y no palabras particulares.

Lexicón: Es el diccionario en el que se registran las palabras que conoce un hablante. Este diccionario especifica los rasgos idiosincráticos de las piezas léxicas (palabras).

Lista de parada: Una stoplist es una lista que contiene un conjunto de palabras "vacías", que son eliminadas del documento en el primer paso de la indización.

Minería de datos: Es una actividad de extracción y análisis de datos, cuyo objetivo es descubrir patrones de comportamiento y relaciones ocultas entre dos o más variables contenidas en bases de datos. Utiliza diferentes combinaciones de técnicas y procesos, tales como aprendizaje automático, algoritmos genéticos, lógica difusa, redes neuronales, análisis estadísticos, técnicas de modelado, etc. para identificar patrones de comportamiento y relaciones sutiles y complejas entre los datos. A partir de estos análisis infiere o genera reglas de comportamiento que permiten hacer predicciones más fiables sobre la evolución de un sistema. Las aplicaciones más conocidas son: segmentación de mercados, definición del perfil de usuario, detección de fraudes, análisis de riesgo de crédito, Vigilancia tecnológica, etc. La minería de datos es de interés cuando el volumen de información a analizar es muy grande.

Motor de búsqueda: También llamado buscador o indexador de información. Es una herramienta usada para localizar los documentos en un repositorio por medio de palabras clave o una clasificación temática. Algunos requieren el uso de operadores booleanos (or, and, not), los más avanzados permiten la interrogación en lenguaje natural.

N-grama: Un n-grama es una secuencia de *n* elementos, palabras o caracteres, extraídos de un texto de forma no necesariamente correlativa. Sin embargo, se entiende habitualmente que un *n*-grama es una secuencia de *n* caracteres contiguos que puede contener espacios en blancos y, por tanto, estar formado por segmentos de varias palabras consecutivas

Palabras vacías: Se denominan *stop words* o palabras vacías aquellas palabras que, a pesar de un uso frecuente, aportan por sí solas poco significado a un texto. En la sentencia anterior se muestran subrayadas algunas palabras vacías del castellano.

Página Web: es la unidad de información, es la única unidad de información solicitada y recibida que después será mostrada por un navegador en la pantalla. Consiste en uno o más recursos Web, que serán vertidos simultáneamente como una unidad en la pantalla. Los recursos Web son el texto, imágenes, sonido, etc. Entonces para nuestro estudio, una página es un documento.

Sitio Web: Un sitio web se puede definir como un conjunto de páginas localizadas en la misma dirección IP que se relacionan por tener información común de relevancia.

Stemmers: es un programa (o su correspondiente algoritmo) que realiza la operación llamada *stemming*, una suerte de lematización, pero consistente en obtener la raíz a la cual se «pega» un determinado sufijo. Como puede adivinarse, se utiliza profusamente con documentos y consultas en inglés y, con esta lengua, parece dar buenos resultados.

Taxonomía: Es un sistema formal para la clasificación ordenada de conocimiento.

Término: Puede ser una unidad léxica simple, una combinación de varias unidades, un signo lingüístico, en forma de sustantivo, o un signo extralingüístico que pertenece a un lenguaje o un código artificial. Como signo extralingüístico, un término se puede expresar por números, letras u otros símbolos.

The Red Badge of Courage: Novela **"**La insignia roja del valor**"**.

Vector: Es un elemento de una estructura algebraica llamada espacio vectorial, que esencialmente es un conjunto de elementos con un conjunto de axiomas que debe satisfacer cada uno de ellos.

Web: La Web es un conjunto de documentos entrelazados en un sistema de hipertexto. El usuario entra en la Web a través de una página de inicio. La misma forma el conjunto total de documentos de hipertexto con enlaces entre ellos que residen en servidores HTTP al rededor de todo el mundo.

WordNet es una base de datos léxica que acumula información léxica sobre las palabras del idioma inglés.

LISTA DE ACRÓNIMOS

ANTF Función que representa una frecuencia normalizada de un rasgo en un documento, (Augmented Normalizad Term Frequency).

BinIDF Función de proyección basada en la frecuencia inversa del documento, (*Binary-Inverse Frequency Document*).

Bin Función de proyección binaria, (*Binary*).

DES Deporte, Salud, Educación.

FILPACON Proyecto de Filtrado de Paquetes.

GF-IDF Función de ponderación basada en la frecuencia global de un rasgo corregida con la frecuencia inversa del documento, (*Global Frequency - Inverse Document Frequency*).

H Función Entropía.

HTML Lenguaje de marcado de hipertexto, (*HyperText Language Markup*).

IA Inteligencia Artificial (*Artifial Intelligence*).

ICD-9-CM3 Clasificación Internacional de Enfermedades 9 ª revisión modificación clínica (*Internacional Classification of Diseases 9th Revision Clinical Modification*).

IG Ganancia de Información, (*Information Gain.*)

K-NN Algoritmo de los k vecinos más cercanos (*k-Nearest Neighbour*).

LSA Análisis de Semántica Latente, (*Latent Semantic Analysis*).

LSI Índice de Latencia Semántica, (*Latent Semantic Indexing*).

MI Información Mutua (*Mutual Information*).

NB Clasificador Naïve Bayes.

NLM Biblioteca nacional de medicina.

ODP /DMOZ Directorio de sitios web de contenido abierto (*Open Directory Proyect*).

PIF Frecuencia Inversa Probabilística (*Probabilistic Inverse Frequency*).

SVM Maquinas de Vectores de Soporte (*Support Vector Machine).*

TC Clasificación automática de textos, (*Text Classification*).

TF Función de ponderación basada en frecuencias de aparición o bolsa de palabras, (*Term Frequency*).

TF-IDF Función de ponderación basada en la frecuencia de un rasgo corregida con la frecuencia inversa del documento, (*Text Frequency - Inverse Document Frequency*).

TREC Concurso internacional o Conferencias de evaluación y recuperación de texto TREC, (*Text REtrieval Evaluation Conferences*).

UCI Universidad de las Ciencias Informáticas.

URL Localizador Uniforme de Recursos, (*Uniform Resource Locator*).

UIUC Universidad de Illinois en Urbana Champaign, (*University of Illinois at Urbana-Champaign*).

USC Universidad de Santiago de Compostela**.**

VSM Modelo de espacio vectorial, (*Vector Space Model*).

WIDF Función de ponderación basada en la frecuencia inversa ponderada, (*Weighted Inverse Document Frequency*).

WTF Función de ponderación basada en frecuencias normalizadas o Bolsa de palabras ponderada, (*Weighted Term Frequency*).