Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 15 de Julio de 2006

Ingeniero

Carlos Azurdia

Coordinador de la Comisión de Tesis

Respetable Ingeniero Azurdia:

Por este medio hago de su conocimiento que he revisado el trabajo de graduación de la estudiante **LISBETH LINDSAY ARRIAZA RIVERA**, el cual se titula: **"REDES NEURONALES COMO MOTORES DE BUSQUEDA EN LA WEB"**, y a mi criterio el mismo cumple con los objetivos propuestos para su desarrollo, según el protocolo.

Sin otro particular, me suscribo de usted.

Atentamente,

Ing. Virginia Victoria Tala Ayerdi

Colegiado No. 4,628

Asesor de Tesis

Universidad San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 26 de Julio de 2006

Ingeniero
Jorge Armin Mazariegos Rabanales
Director de la Escuela de Ingeniería
En Ciencias y Sistemas

Respetable Ingeniero Mazariegos:

Por este medio hago de su conocimiento que he revisado el trabajo de graduación de la estudiante **LISBETH LINDSAY ARRIAZA RIVERA,** titulado: **"REDES NEURONALES COMO MOTORES DE BUSQUEDA EN LA WEB"**, y a mi criterio el mismo cumple con los objetivos propuestos para su desarrollo, según el protocolo.

Al agradecer su atención a la presente, aprovecho la oportunidad para suscribirme,

Atentamente,

Ing. Carlos Alfredo Azurdia
Coordinador de Privados
y Revisión de Trabajos de Graduación

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA - INGENIERIA EN CIENCIAS Y SISTEMAS - COORDINACION TESIS -

ESCUELA DE CIENCIAS Y SISTEMAS

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÌA
ESCUELA DE CIENCIAS Y SISTEMAS
TEL.: 24767644

"Todo por ti Carolinga mía"
Dr. Carlos Martínez Durán
2006: Centenario de su nacimiento

El Director de la Escuela de Ingeniería en Ciencias y Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor con el visto bueno del revisor y del Licenciado en Letras, de trabajo de graduación titulado ***"REDES NEURONALES COMO MOTORES DE BUSQUEDA EN LA WEB"****, presentado por la estudiante* ***LISBETH LINDSAY ARRIAZA RIVERA,*** *aprueba el presente trabajo y solicita la autorización del mismo*

"ID Y ENSEÑAD A TODOS"

Ing. Jorge Armin Mazariegos Rabanales
Director, Escuela de Ingeniería en Ciencias y Sistemas

29 de Agosto 2006

Universidad de San Carlos
de Guatemala

Facultad de Ingeniería
Decanato

Ref. DTG.302.06

El Decano de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería en Ciencias y Sistemas, al trabajo de graduación titulado: **REDES NEURONALES COMO MOTORES DE BUSQUEDA EN LA WEB,** presentado por la estudiante universitaria **Lisbeth Lindsay Arriaza Rivera,** procede a la autorización para la impresión del mismo.

IMPRÍMASE.

Ing. Murphy Olympo Paiz Recinos
DECANO

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANO
FACULTAD DE INGENIERIA

Guatemala, agosto de 2006

/cc

Todo por ti, Carolingia Mía
Dr. Carlos Martínez Durán
2006: Centenario de su Nacimiento

Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

REDES NEURONALES COMO MOTORES DE BÚSQUEDA EN LA WEB

Lisbeth Lindsay Arriaza Rivera

Asesorada por la Inga. Virginia Victoria Tala Ayerdi

Guatemala, agosto de 2006

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

REDES NEURONALES COMO MOTORES DE BÚSQUEDA EN LA WEB

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

LISBETH LINDSAY ARRIAZA RIVERA
ASESORADA POR LA INGA. VIRGINIA VICTORIA TALA AYERDI

AL CONFERÍRSELE EL TÍTULO DE
INGENIERA EN CIENCIAS Y SISTEMAS

GUATEMALA, AGOSTO DE 2006

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Murphy Olympo Paiz Recinos
VOCAL I	Inga. Glenda Patricia García Soria
VOCAL II	Lic. Amahán Sánchez Álvarez
VOCAL III	Ing. Julio David Galicia Celada
VOCAL IV	Br. Kenneth Issur Estrada Ruiz
VOCAL V	Br. Elisa Yazminda Vides Leiva
SECRETARIA	Inga. Marcia Ivonne Véliz Vargas

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Murphy Olympo Paiz Recinos
EXAMINADORA	Inga. Virginia Victoria Tala Ayerdi
EXAMINADORA	Inga. Floriza Felipa Avila Pesquera
EXAMINADOR	Ing. Edgar Estuardo Santos Sutuj
SECRETARIA	Inga. Marcia Ivonne Véliz Vargas

HONORABLE TRIBUNAL EXAMINADOR

Cumpliendo con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

REDES NEURONALES COMO MOTORES DE BÚSQUEDA EN LA WEB,

tema que me fuera asignado por la Dirección de la Escuela de Ingeniería en Ciencias y Sistemas, en julio de 2005.

Lisbeth Lindsay Arriaza Rivera
Guatemala, agosto de 2006

DEDICATORIA A:

Dios

Mi amigo Fiel, quien me da la vida cada día y me llena de bendiciones y fuerzas para seguir adelante.

Mis padres

Por haberme dado la oportunidad de estudiar y apoyarme a lo largo de la carrera.

Mis tíos

Por su cariño y palabras de ánimo.

Mis primos

Por la amistad y cariño que me han brindado y por ayudarme en momentos difíciles.

Mis amigos

Por su amistad y apoyo brindado de forma incondicional.

Mi amigo Dario

Por su cariño y ánimos en los momentos más duros de la carrera.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

GLOSARIO

Agente Inteligente	Un programa de software autónomo que ejecuta una tarea específica.
Aprendizaje Inductivo	Tipo de aprendizaje que generaliza a partir de la experiencia para mejorar su desempeño.
Axón	Es el camino de salida de la señal generada por la neurona.
Base de datos	Colección de datos relacionados diseñados para cumplir con las necesidades de información de determinada organización.
Buffer	Área de almacenamiento temporal en la memoria del sistema.
CGI	Pasarela de Interfaz Común -CGI, por sus siglas en inglés- es una tecnología de la Web que permite a un navegador solicitar datos de un programa ejecutado en un servidor Web. CGI especifica un estándar para transferir datos entre el cliente y el programa

Dendrita	Son las encargadas de recibir las informaciones de los axones de otras neuronas.
Explotación de Datos	Aplicaciones cuya finalidad es buscar datos escondidos en bases de datos para inferir a partir de ellos comportamientos en el futuro.
Freeware	El software freeware es software que se puede copiar, usar y redistribuir libremente, pero no incluye archivos fuentes.
Filesize	Tamaño de un archivo.
Grafo	Es un objeto matemático que se utiliza para representar circuitos, redes, etc.
Hash	Es una función utilizada para generar claves que representan a un documento, registro, etc.
Hipertexto	Texto disponible en la Web, el cual contiene enlaces a otros documentos.
Indexador	Es una herramienta que construye un índice, a partir de una serie de documentos.

Intranet	Es una estructura de red cuyo acceso está restringido a un grupo específico.
Licencia GNU	Esta licencia permite la modificación y redistribución de software únicamente bajo esta misma licencia.
Neurotransmisores	Son las principales sustancias de la sinapsis.
OCR	Reconocimiento óptico de Caracteres. El análisis de los datos digitalizados para reconocer caracteres de forma que puedan convertirse en texto que pueda ser editado.
Offline	Trabajar con documentos sin conexión a Internet.
Paradigma	Es un conjunto de supuestos teóricos que adoptan miembros de una determinada comunidad.
Peso	Representa la intensidad o fuerza de conexión de la sinapsis en una neurona.
Servidor	Máquina cuyo propósito es proveer datos de modo que otras máquinas puedan utilizar esos datos.
Shareware	Es un tipo de programas para poder evaluar gratuitamente por un tiempo limitado.

Sinapsis	Uniones que se producen entre el axón y las dendritas.
Spam	Son mensajes de tipo publicitario no solicitados enviados en cantidades masivas.
URL	Es la cadena de caracteres con la cual se asigna una dirección única a cada uno de los recursos de información disponibles en la Web.
WEB	Es un medio de comunicación de texto, imágenes, etc., a través de Internet.

RESUMEN

La búsqueda de información en Internet ha encontrado una serie de problemas, como por ejemplo, la excesiva cantidad de información que sigue creciendo día a día, por lo que es necesario incorporar nuevos algoritmos dentro de las herramientas de recuperación de información.

Dentro de las técnicas que se están incorporando, se encuentran las Redes Neuronales, hasta el momento no existen muchas herramientas que utilicen ésta tecnología, pero las Redes Neuronales pueden ser usadas de varias formas para optimizar el funcionamiento de los buscadores.

Existen muchas herramientas que no proveen relevancia en los resultados de las búsquedas, lo que representa una debilidad para las mismas, ya que los usuarios deciden cambiarlas por otras. Esta debilidad representa un gran problema, ya que es necesario que un buscador tenga bastante tráfico para poder obtener ganancias. Actualmente, hay algunos buscadores que han incorporado las Redes Neuronales en sus algoritmos de Ranking de documentos para proveer mayor relevancia a los usuarios.

También, se ha utilizado Redes Neuronales para mejorar el funcionamiento de los robots que recolectan páginas. Estas redes utilizan retropropagación, es decir, propagan el error a las capas de atrás para que la red pueda ajustar sus pesos y mejorar su funcionamiento.

En general, dentro de un buscador se encuentran varias fases que pueden ser optimizadas utilizando Redes Neuronales.

OBJETIVOS

1. Evaluar las propiedades y elementos de las redes neuronales para ser utilizados como motores de búsqueda.

2. Definir los beneficios que las redes neuronales pueden proporcionar en las búsquedas.

3. Comparar los actuales métodos de búsqueda con los que utilizan redes neuronales.

INTRODUCCIÓN

En la actualidad, es realmente necesario utilizar Internet para buscar o recuperar información, cada día crece más el número de personas que utilizan éste medio y las herramientas de búsqueda existentes.

La mayoría de personas solamente conocen una o dos herramientas de búsqueda y no están conscientes de que pueden realizar una búsqueda más profunda por medio de otras herramientas.

Los buscadores se han encontrado con una serie de problemas que no les permite entregar mejores resultados sobre las consultas realizadas, uno de estos problemas es el Internet Invisible que se detalla dentro de esta investigación.

Debido a que se han presentado problemas en el funcionamiento de los buscadores, surgió una nueva generación de herramientas, las cuales utilizan agentes inteligentes.

El objetivo principal de este trabajo es analizar un paradigma de la Inteligencia Artificial para ser usado como motor de búsqueda en Internet, presentando sus ventajas y desventajas, este paradigma es: Redes Neuronales.

1. SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN Y MOTORES DE BÚSQUEDA

1.1 Historia de los buscadores Web

1.1.1 Herramientas de primera generación

Las herramientas de búsqueda de primera generación, que también se conocen como sistemas basados en el acceso a computadoras remotas, se dividen en dos tipos:

- Buscadores: Los buscadores son aplicaciones que recopilan información que se encuentra en la red y luego la ingresan en una base de datos.
- Índices: También conocidos como directorios, son grandes bases de datos en donde la información se clasifica por temas.

Entre los primeros buscadores se encuentran: Jumpstation, el cual nació en 1993 y buscaba en una base de datos lineal haciendo parejas con las palabras clave, este fue un sistema muy lento; World Wide Web, éste nació en 1994 y usaba 4 bases de datos, y no presentaba los datos ordenados; Repository-Based Software Engineering (RBSE), fue el primer buscador que ofrecía los resultados basado en una clasificación, ordenando los resultados por relevancia de las palabras clave.

El primer directorio que se conoció fue Elnet Galaxy, que en sus búsquedas utilizaba: Telnet, Gopher y Web y entregaba los resultados clasificados en categorías. En 1994 nació Yahoo, el cual se basó en una colección de páginas Web, este buscador surgió en la preparación de la tesis doctoral de David Filo y Jerry Yang, los cuales al pasar tiempo en la web, juntaron una lista de sus sitios favoritos y la llamaron “La guía de Jerry para la World Wide Web”.

Algunos buscadores clasifican recursos y otros buscan por palabras clave entre los que se encuentran Alta Vista y Lycos.

Los buscadores más famosos en la actualidad son Google y Altavista.

Las herramientas de primera generación tienen una serie de limitaciones como por ejemplo: la enorme cantidad de información que existe en la WEB y que sigue creciendo a diario, los buscadores sólo son capaces de buscar en un 10 % de la red.

Cuando queremos localizar información por medio de estos buscadores, no estamos realizando la búsqueda en toda la red, sino que solamente en la base de datos del motor de búsqueda, y para que las páginas estén en esa base de datos deben ser estáticas y deben estar enlazadas por otras páginas, lo que resulta en un gran problema porque existe un gran porcentaje de páginas en la red que no están siendo visitadas.

Otra limitación es que los buscadores no pueden llegar a la información almacenada en servidores que no permiten el acceso público, ya que utilizan protocolos de limitación o prohibición de acceso a sus contenidos. Por ejemplo las revistas electrónicas requieren contraseñas o nombres de usuario para poder accederlas.

1.1.2 Herramientas de segunda generación

Para solucionar los problemas y limitaciones de los buscadores surgió una segunda generación de herramientas de búsqueda, a la cual se le denomina “Robots y multibuscadores”, las cuales utilizan robots y agentes inteligentes.

Estas herramientas tienen mayor capacidad para rastrear, localizar y recuperar información en la red, utilizando criterios de clasificación y jerarquía con mecanismos inteligentes, los cuales mejoran los resultados de las búsquedas.

Entre los problemas que estas herramientas pueden solucionar tenemos:

- El límite del vaciado de los webs, esto quiere decir que estas herramientas permiten navegar offline, de una forma mas rápida, sin costo y sin problemas de conexión
- Selección, actualización y ordenación por intereses
- Buscan sobre la Internet Invisible
- Análisis documental
- Falta de actualización de los resultados de los buscadores

También tienen una serie de problemas y desventajas, entre los cuales están:

- Problemas de incompatibilidad con algunos sistemas operativos y antivirus
- Mayores requerimientos de hardware
- Son productos comerciales, aunque existen versiones shareware, freeware y demo.

Cuando se realizan búsquedas por medio de varios motores de búsqueda se encuentra información repetida, éste es un problema que los multibuscadores deben resolver. Los multibuscadores son herramientas que tienen la capacidad de utilizar varios motores de búsqueda al mismo tiempo y eliminar resultados repetidos. Entre los multibuscadores se podrían mencionar: Copernic y WebFerret.

1.2 Buscadores Web

1.2.1 Ranking de documentos

Cuando se realiza una consulta empleando un conjunto de términos, el buscador recupera una lista de documentos ordenados según su relación entre el contenido de los documentos y la consulta realizada, a este proceso se le denomina ranking de documentos.

Los modelos para realizar ranking se dividen en dos tipos: los que comparan la consulta con documentos individuales y los que comparan con un conjunto de documentos. Dentro de los que comparan con documentos individuales se encuentran los más utilizados: modelo de espacio vectorial y modelo probabilístico.

Las consultas realizadas por los usuarios y los documentos que se encuentran en la base de datos se pueden representar por un vector $v_1, v_2, ..., v_n$, en donde v_n es igual a 1 cuando el término n está presente o igual a 0 si no lo está.

El modelo de espacio vectorial coloca a los vectores de los documentos y de la consulta en un vector de n dimensiones, n representa el número de términos únicos en la base de datos. El sistema calcula el coseno del ángulo que forma el vector de la consulta con los demás, el resultado determina la similitud entre la consulta y cada documento, y de esta manera permite ordenar los datos en función de esa similitud.

El modelo probabilístico dice que los términos que aparecen en los documentos recuperados anteriormente para una consulta dada tienen más peso que si no hubieran aparecido en esos documentos.

1.2.2 Proceso de búsqueda

Para realizar una búsqueda primero se debe necesitar algún tipo de información, luego el usuario debe colocar en el navegador la dirección del sitio que le ayudará en su búsqueda, este sitio puede ser un directorio para buscar por tema o un buscador Web para ingresar palabras clave.

El proceso de búsqueda en sí tiene los siguientes pasos:

- El buscador consulta un índice de páginas
- Luego obtiene una lista de documentos
- Con la lista que se obtuvo realiza un ranking
- Consolida los resultados (elimina repeticiones, reagrupa, etc.)
- Presenta los resultados al usuario

Luego que el usuario obtiene los resultados, se encuentra con una gran cantidad de direcciones o enlaces, desafortunadamente es muy común que no encuentre lo que buscaba.

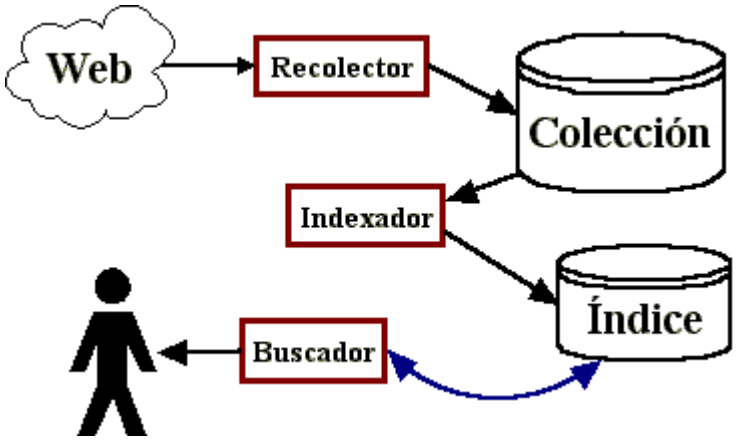
1.2.3 Arquitectura de un buscador Web

Un buscador Web contiene los siguientes elementos, como se muestra en la Figura 1:

- Una interfaz para los usuarios
- Un buscador: éste recupera algunas páginas del índice basándose en los requerimientos del usuario.
- Un índice: luego de que el indexador revisa las páginas con las palabras clave ingresadas por el usuario, las ingresa en un índice, para que luego sea usado por el buscador para presentar la información al usuario.
- Un indexador: normalmente utiliza un índice invertido, esto quiere decir que la colección de páginas se convierte en una lista de palabras, cada palabra apunta a una lista de documentos dentro de la colección. Este esquema se muestra en la figura 2.

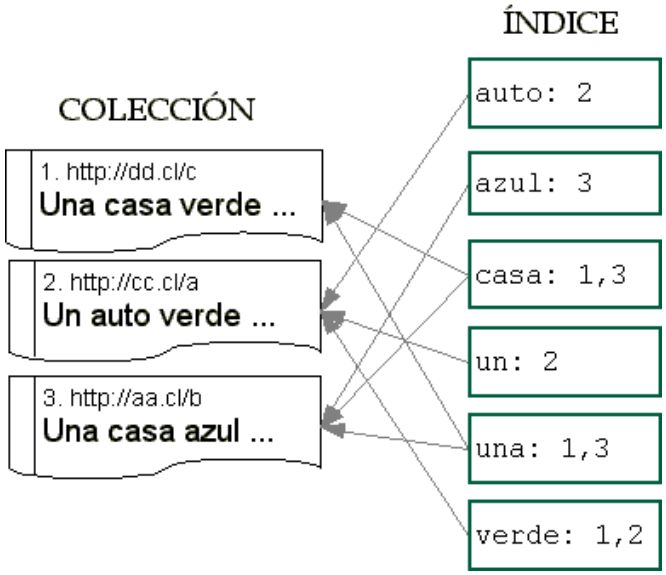
- Un robot o recolector: el objetivo de este robot es crear una colección de páginas. Para obtener esta colección, el robot visita varias páginas y las incluye en su colección, luego extrae sus enlaces y valida si ya existen en su colección, si no existen las visita.

Figura 1**. Arquitectura de un buscador Web**



Fuente: **Carlos Castillo. Búsqueda en la Web.**

Figura 2. **Indexador**



Fuente: **Carlos Castillo. Búsqueda en la Web.**

1.3 Los directorios o índices.

1.3.1 Introducción

Los directorios son un conjunto, en éste caso de páginas Web organizadas por temas, de forma jerárquica. Algunos motores utilizan robots para recuperar información, que luego asignan en diferentes categorías dependiendo de la frecuencia de palabras encontradas.

Un directorio no es un buscador, algunos le llaman “buscador temático”, una de las diferencias principales es el volumen y cantidad en la recuperación de información. Otra diferencia con los buscadores, es que el objetivo de un buscador es crear un índice de palabras mientras que el objetivo de un directorio es clasificar la información por temas.

Algunos de los directorios más conocidos son: Yahoo, Magellan, Yanoff, WWW Virtual Library.

1.3.2 Ventajas y desventajas de los directorios

- La principal desventaja, es la velocidad con que se realizan las búsquedas.
- No todos los documentos se pueden clasificar fácilmente, ya que algunos pertenecen a más de una clase.
- Una ventaja muy importante es que presenta resultados ya analizados para el usuario, y no sólo un montón de documentos con palabras en común.

1.3.3 Tipos de directorios

Directorios especializados: estos directorios ofrecen descripción de las páginas de una mejor manera que los directorios comunes. Los directorios especializados crean índices complejos de conceptos con el objetivo de orientar en la búsqueda de información.

Directorios orientativos: estos directorios ofrecen los documentos o recursos que se consideran de mayor interés para el usuario. Los directorios de este tipo sólo se preocupan en recoger los documentos más útiles e interesantes.

1.4 Problemas de la información en Internet

1.4.1 Infoxicación

La infoxicación es el exceso de información que se encuentra en esta era, la cual se ha denominado: la era de la información. Este problema ha impulsado la creación de sistemas que permitan encontrar información en Internet, dentro de esta información, también se encuentra una gran cantidad de información escondida, llamada Internet Invisible. La Internet Invisible es un conjunto de fuentes de información de gran valor, las cuales solo se pueden acceder por medio de una pasarela o formulario Web y que no pueden ser recuperados por los robots o buscadores, por lo que para ellos es invisible.

En el año 2000 Michael K. Bergman realizó un estudio que refleja el tamaño de la parte invisible en la red.[1]

[1] **Internet Visible e Invisible** , Biblioteca Nacional de Ciencias de la Salud, Instituto de Salud Carlos III

Según este estudio, los sitios invisibles reciben un 50% más de tráfico mensual que los visibles o superficiales. Una de las razones puede ser, según este estudio, que la calidad de los contenidos de los sitios invisibles es mucho mayor que los visibles, la siguiente tabla muestra estos datos:

Tabla I. **Internet Visible e Invisible, Internet visible vs. Internet Invisible**

	Internet Invisible			**Internet Visible**		
Consulta	**Total**	**"Calidad"**	**Porcentaje**	**Total**	**"Calidad"**	**Porcentaje**
Agricultura	400	20	5%	300	42	14.00%
Medicina	500	23	4.60%	400	50	12.50%
Finanzas	350	18	5.10%	600	75	12.50%
Ciencias	700	30	4.30%	700	80	11.40%
Derecho	260	12	4.60%	320	38	11.90%
TOTAL	**2,210**	**103**	**4.70%**	**2,320**	**285**	**12.30%**

Fuente: **Internet Visible e Invisible: Búsqueda y selección de recursos de información en ciencias de la salud. Pág. 9**

1.5 Los operadores de búsqueda

1.5.1 Introducción

Cuando se realizan búsquedas se pueden combinar dos o más términos utilizando operadores de búsqueda. Los operadores básicos son: or, and y not. Estos operadores indican cómo se debe realizar la búsqueda sobre los términos ingresados.

1.5.2 Operadores lógicos

- OR: Este operador indica unión de conjuntos. Si se realiza una búsqueda con este operador entre dos términos **x** e **y**, se obtendrá un conjunto de documentos indexados bajo el término **x**, ó **y**, ó ambos. Este operador es muy útil para aumentar el ámbito de una búsqueda.
- AND: Este operador indica intersección de conjuntos. Si se realiza una búsqueda con este operador entre dos términos, el resultado obtenido será un conjunto con elementos que son comunes a ambos términos.
- NOT: Este operador indica exclusión de conjuntos. Al realizar una búsqueda con este operador el resultado será un conjunto con elementos que no sean comunes al término ingresado.

1.5.3 Operadores de campo

Los operadores de campo y de proximidad son más restrictivos que el operador lógico AND, y normalmente se usan en búsquedas por medio de lenguaje natural.

Estos operadores son:

- (G): La búsqueda entre dos términos por ejemplo: x (G) y, recuperará todos los documentos que incluyan el término x como el término y, siempre que ambos términos se encuentren en el mismo campo. Por ejemplo si se tiene un documento titulado “El desafio de la gestión del riesgo y la disminución de la vulnerabilidad”, escrito por Pascal Girot, la búsqueda puede plantearse de la siguiente manera: riesgo (G) vulnerabilidad. Otra forma de hacer la búsqueda puede ser: Girot (G) Pascal.[2]

[2] **Proceso E: Búsqueda de información en las bases de datos**

- (F): La búsqueda: x (F) y, obtendrá los registros que incluyan x como y, siempre que los términos se encuentren en el mismo campo o en la misma ocurrencia de un campo repetible. Ejemplo: Girot (F) Pascal, podrá recuperar el registro por cuanto ambos términos se encuentran en la misma ocurrencia del campo de autor.
- **.** : Este operador es parecido a (F), con una restricción adicional, la cual es que los dos términos no se encuentren a más de n palabras de distancia, donde n es el número de puntos más uno. Ejemplo: x**..**y.
- **$**: Este operador también es parecido a (F), con una restricción adicional: los dos términos deben encontrarse exactamente a n palabras de distancia, donde n es el número de $ más uno. Ejemplo: x $ Y.

1.6 Algoritmos de búsqueda

1.6.1 Introducción

La búsqueda de información normalmente se clasifica como interna o externa. La búsqueda interna se refiere a información dentro de la memoria y la búsqueda externa es la información que se encuentra en dispositivos externos.

Los objetivos de los algoritmos de búsqueda son:

- Determinar si lo que se busca se encuentra dentro del conjunto en dónde se está realizando la búsqueda.
- Si lo que se busca está en el conjunto, encontrar la posición en la que se encuentra.

Los tipos de búsqueda interna principales son: búsqueda secuencial, binaria y búsqueda por tablas de Hash.

1.6.2 Búsqueda secuencial

Es el método en el cual se recorre y examina cada uno de los elementos que componen nuestro listado de valores hasta encontrar el o los elementos buscados, o hasta que se han examinado todos los elementos y no se han encontrado coincidencias.

En la figura 3 se muestra el flujo de un algoritmo de búsqueda secuencial:

Figura 3. **Flujo de búsqueda secuencial**

1.6.3 Búsqueda binaria

Este método de búsqueda pide ciertos requerimientos necesarios para su utilización y uno de ellos es que el listado de valores debe estar ordenado. Esta búsqueda consiste en dividir nuestro listado por su elemento medio en dos sublistados más pequeños, y comparar el elemento con el del centro. Si coinciden, la búsqueda se termina. Si el elemento es menor, debe estar en el primer sublistado claro si se encontrara el valor, y si es mayor está en el segundo.

1.6.4 Búsqueda mediante transformación de claves (Hashing)

Este método consiste en aplicar una función que traduce un conjunto de posibles valores llave en un rango de direcciones relativas. Un problema potencial encontrado en este proceso, es que tal función no puede ser uno a uno; las direcciones calculadas pueden no ser todas únicas, cuando $R(k_1) = R(k_2)$ Pero : K_1 diferente de K_2 decimos que hay una colisión. A dos llaves diferentes que les corresponda la misma dirección relativa se les llama sinónimos**.**

A las técnicas de cálculo de direcciones también se les conoce como:

- Técnicas de almacenamiento disperso
- Técnicas aleatorias
- Métodos de transformación de llave - a- dirección
- Técnicas de direccionamiento directo
- Métodos de tabla Hash
- Métodos de Hashing

Pero el término mas usado es el de Hashing. Al cálculo que se realiza para obtener una dirección a partir de una llave se le conoce como función Hash.

La eficiencia de una función Hash depende de: la distribución de los valores de llave que realmente se usan, el numero de valores de llave que realmente están en uso con respecto al tamaño del espacio de direcciones, el numero de registros que pueden almacenarse en una dirección dada sin causar una colisión y por último la técnica usada para resolver el problema de las colisiones.

Los costos de esta búsqueda son: el tiempo de procesamiento requerido para la aplicación de la función Hash y los accesos E/S requeridos para solucionar las colisiones.

En la figura 4 podemos observar el flujo que se realiza en esta búsqueda:

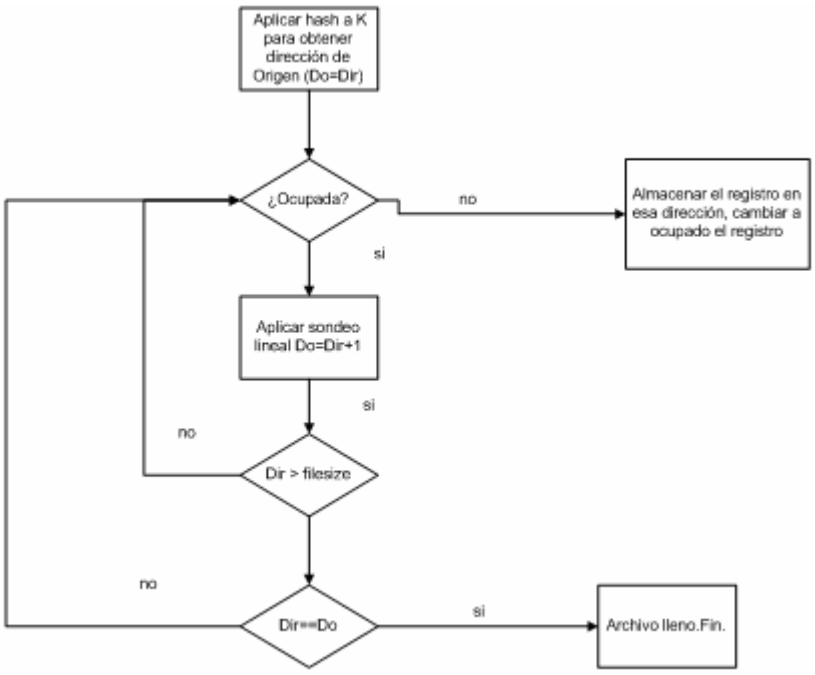


Figura 4**. Flujo de búsqueda por HASH**

2. REDES NEURONALES

2.1 Descripción de las Redes Neuronales

2.1.1 Historia de las Redes Neuronales

Desde hace muchos años se ha intentado imitar el funcionamiento del cerebro humano, por ejemplo a finales del siglo XIX se le comparó con la operación de una bomba hidráulica; desde 1920 hasta 1930 se intentó utilizar la teoría de conmutación telefónica como punto de partida de un sistema de conocimiento parecido al del cerebro. Entre 1940 y 1950 los científicos empezaron a pensar en las redes neuronales sugiriendo el siguiente concepto: las neuronas del cerebro funcionan como interruptores digitales (encendido – apagado) al igual que el computador digital. Así nació la idea de "revolución cibernética" que maneja la analogía entre el cerebro y el computador digital.

En 1943 Walter Pitts, Bertran Russell y Warren McCulloh intentaron explicar el funcionamiento del cerebro humano, por medio de una red de células conectadas entre sí ejecutando operaciones lógicas.

El ciclo "sentidos – cerebro – músculos", por medio de la retroalimentación producirían una reacción positiva si los músculos reducen la diferencia entre una condición percibida por los sentidos y un estado físico impuesto por el cerebro. También definieron la memoria como un conjunto de ondas que resplandecen en un circuito cerrado de neuronas. [3]

[3] **Informe sobre redes neuronales**, Gustavo Luis Pavía

Seis años después el fisiólogo Donald Hebb expuso que las redes neuronales podían aprender. Esta propuesta tenía que ver con las conexiones entre las neuronas. Hebb explicó que la activación repetida de una neurona por otra a través de una sinapsis o conexión, aumenta su conductividad y la hacía más propensa a ser activada sucesivamente, esto induce a la formación de un circuito de neuronas estrechamente conectadas entre sí.

En 1951, Marvin Minsky y Dean Edmonds realizaron la primera máquina de redes neuronales que estaba compuesta de 300 tubos de vacío y un piloto automático de un bombardero B-24. Esta máquina fue llamada “Sharc” y se trataba de una red de 40 neuronas artificiales que imitaban el cerebro de una rata. Cada neurona hacia el papel de una posición de un laberinto y si se activaba daba a entender que la rata sabía en que punto del laberinto estaba. Las neuronas que estaban conectadas alrededor de la activada, hacían la función de alternativas que seguir por el cerebro, la elección entre derecha o izquierda, que es la activación de la siguiente neurona, estaba dada por la fuerza de sus conexiones con la neurona activada.

Después de la creación de esta red neuronal, Minsky escribió su tesis doctoral acerca de la misma y exponía que si se realizaba este proyecto a gran escala, con millones de neuronas más y con diferentes sensores y tipos de retroalimentación, la máquina sería capaz de razonar.

En 1960 se creó la primer red neuronal para resolver un problema, este problema era eliminar ecos en las líneas telefónicas, esta red fue llamada Adaline (Adaptive Linear Neuron) y fue creada por Bernard Widrow y Marcial Hoff.

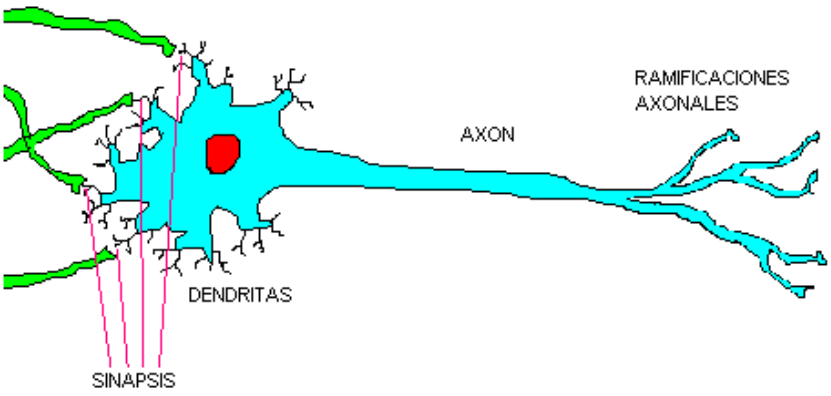
2.1.2 Definición de una Red Neuronal

2.1.2.1 Definición de una Red Neuronal Biológica

Para poder entender una red neuronal artificial, es necesario entender las redes neuronales biológicas, debemos recordar que el cerebro humano está compuesto por billones de neuronas interconectadas entre sí formando circuitos o redes que desarrollan funciones específicas.

Una neurona recoge señales que vienen de otras neuronas a través de unas estructuras llamadas dendritas. Esta neurona emite impulsos de actividad eléctrica a lo largo de una fibra llamada axón, que se separa en millares de ramificaciones.

Figura 5. **Neurona y sus conexiones**[4]



Fuente: **Introducción a las redes neuronales artificiales**

[4] **Introducción a las redes neuronales artificiales**

Las extremidades de estas ramificaciones llegan hasta las dendritas de otras neuronas y establecen unas conexiones llamadas sinapsis, en estas conexiones se produce la transformación de los impulsos eléctricos en un mensaje neuroquímico, mediante la liberación de unas sustancias llamadas neurotransmisores.

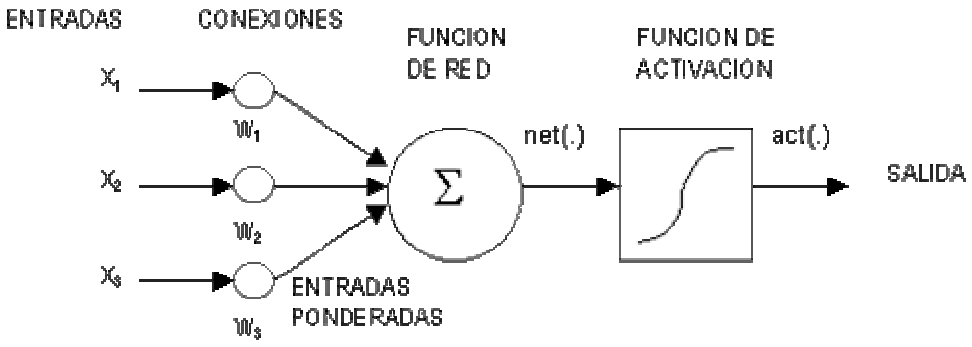
El efecto de los neurotransmisores sobre la neurona receptora puede ser excitación o inhibición, y es variable, de manera que se puede hablar de la fuerza de una sinapsis. Estas señales de excitación e inhibición recibidas por una neurona se combinan, y en función de la estimulación total recibida, la neurona toma un cierto nivel de activación, que se traduce en la generación de breves impulsos nerviosos con una determinada frecuencia o tasa de disparo, y su propagación a lo largo del axón hacia las neuronas con las cuales hace una sinapsis.

De esta manera la información se transmite entre las neuronas y va siendo procesada a través de las conexiones sinápticas y las propias neuronas. El aprendizaje de las redes neuronales se produce mediante la variación de la efectividad de la sinapsis, de esta manera cambia la influencia que unas neuronas ejercen sobre otras, de este concepto se deduce que la arquitectura, el tipo y la efectividad de las conexiones en un momento dado, representan en cierta manera la memoria o estado de conocimiento de la red.

2.1.2.2 Red Neuronal Artificial

Las neuronas se modelan mediante unidades de proceso, cada unidad de proceso se compone de una red de conexiones de entrada, una función de red (función de propagación), encargada de computar la entrada total combinada de todas las conexiones, un núcleo central de proceso, el cual se encarga de aplicar la función de activación y la salida, en donde se transmite el valor de activación a otras unidades.

Figura 6. **Unidad de proceso** [5]



Fuente: **Introducción a las redes neuronales artificiales**

La función de red calcula el valor de entrada total a la unidad, generalmente la suma ponderada de todas las entradas recibidas, esto quiere decir, de las entradas multiplicadas por el peso o valor de las conexiones. Equivale a la combinación de las señales de excitación e inhibitorias de las neuronas biológicas.

[5] Introducción a las redes neuronales artificiales

La función de activación es la que define el comportamiento de las neuronas. Se usan diferentes tipos de funciones, desde simples funciones de umbral a funciones no lineales. Esta función calcula el nivel de activación de la neurona en función de una entrada total.

Las conexiones ponderadas hacen el papel de las conexiones sinápticas, el peso de la conexión equivale a la fuerza de la sinapsis. La existencia de conexiones determina si es posible que una unidad influya sobre otra, el valor de los pesos y el signo de los mismos definen el tipo (excitación/inhibición) y la intensidad de la influencia.

La salida se calcula en función de la activación de la misma, se toma como salida el valor de activación. El valor de salida es similar a la función de la tasa de disparo de una neurona biológica.

2.2 Estructura de una red neuronal

Al diseñar una red se debe establecer como estarán conectas las neuronas y determinar los pesos de las conexiones. Lo usual es disponer las unidades en forma de capas, teniendo comunicación entre redes.

Inicialmente se desarrollaron redes de una sola capa, pero ahora lo usual es disponer tres o más capas, la primera capa actúa como un buffer de entrada, almacenando la información suministrada a la red, esta capa es llamada capa de entrada; otra capa actúa como interfaz o buffer de salida, esta almacena la respuesta de la red para que pueda ser leída, esta capa es llamada capa de salida; y las capas intermedias se encargan de extraer, procesar y memorizar la información, a estas se les llama capas ocultas.

También podemos hablar de redes recurrentes y no recurrentes o redes en cascada. En las redes en cascada la información fluye en una dirección de una capa a otra y no se admiten conexiones laterales. En las redes recurrentes la información puede regresar a los lugares por donde había pasado, formando ciclos y si se admiten conexiones laterales, incluso de una unidad con ella misma.

2.3 Características de las redes neuronales

- Aprendizaje inductivo: no se indican las reglas para dar una solución, sino que extrae sus propias reglas a partir de los ejemplos de aprendizaje, ellas mismas modifican su comportamiento en función de la experiencia.
- Generalización: después de entrenar a la red, se le pueden dar datos distintos a los usados en el aprendizaje. La respuesta que se obtiene dependerá del parecido de los datos con los ejemplos del entrenamiento.
- Abstracción: las redes neuronales artificiales son capaces de abstraer las características principales de las entradas aprendidas, de esta manera pueden procesar datos incompletos o distorsionados.
- Procesamiento en paralelo: las neuronas biológicas trabajan en paralelo; para que las redes neuronales artificiales trabajen en paralelo se deben usar varios procesadores, lo importante es que las redes neuronales tienen una estructura y modo de operación adecuados para el procesamiento en paralelo.
- Memoria distribuida: una red neuronal es capaz de seguir funcionando a pesar de sufrir lesiones en donde se destruyen sus neuronas o conexiones, ya que la información está distribuida en toda la red.

2.4 Tipos de Redes Neuronales

1. Asociadoras de patrones: son redes de dos o más capas cuyo objetivo es asociar a través de un proceso de aprendizaje supervisado, pares de estímulos, llamados patrón de entrada y patrón de salida. Este tipo de redes se puede comparar con los modelos de regresión estadística, los cuales tratan de hallar la relación entre una serie de variables.
2. Redes competitivas: son redes de una o varias capas que tienen en común la competencia entre unidades, con el fin de conseguir que una unidad quede activada y el resto no.
3. Modelos de satisfacción de demanda o de adaptación probabilística: en este tipo de redes cada unidad se actualiza durante un ciclo de procesamiento según cierta probabilidad. Su objetivo es alcanzar soluciones óptimas a problemas que tienen un gran número de demandas simultáneas.

2.5 Redes Neuronales Gráficas

Existen muchas situaciones en el mundo real que pueden ser descritas en un diagrama, que consiste en un conjunto de objetos junto con sus relaciones. Estas ideas producen el concepto de un grafo.

Las Redes Neuronales proveen un enfoque alternativo para resolver problemas de optimización. Pueden ser exitosas en resolver problemas con procesos de manufactura complejos y se han usado en áreas como procesamiento de imágenes, monitoreo y control, ruteo de redes de telecomunicación.

Una Red Neuronal se puede ver como un grafo directo que consiste en nodos e interconexiones sinápticas y enlaces de activación.

2.6 Funciones básicas de las Redes Neuronales

- Clasificación: una red neuronal puede utilizarse para decidir a qué clase se asigna un nuevo elemento a la red. Típicamente el número de clases es reducido en relación con el número de posibles datos de entrada, lo que hace complicado el proceso de clasificación.
- Asociación: la red funciona como un proceso de recuperación de un dato a partir de una entrada relacionada con el dato almacenado. Normalmente la entrada es un dato incompleto a recuperar y a partir de este dato se reestablece la información completa sobre el dato.
- Agrupamiento: esta función clasifica los datos de entrada, una manera de hacerlo es creando prototipos para cada una de las clases.
- Generalización: las redes neuronales tienen la capacidad de detectar relaciones entre los datos, con esto hacen posible implementar modelos funcionales que permiten extrapolar las relaciones numéricas presentes en los datos disponibles a otras situaciones similares.
- Optimización: en problemas de optimización, las redes neuronales proporcionan un procedimiento relativamente rápido para generar una solución que resulta suficientemente satisfactoria.
- Predicción y control: la facilidad de aprendizaje de la red permite seguir los cambios que se producen. Un ejemplo puede ser el análisis de series de tiempo. La red puede predecir las consecuencias de los cambios e indicar decisiones de control para compensar el futuro error de la trayectoria del sistema.

2.7 Aplicaciones de las Redes Neuronales

Explotación de datos: las redes neuronales se pueden usar como una herramienta realmente poderosa para la explotación eficiente de datos comerciales. Algunos usos que ha tenido son: clasificar clientes por hábitos de consumo con los datos obtenidos en los puntos de venta. Con esto se puede determinar los segmentos de la población que responderán con mayor probabilidad a la publicidad personalizada, reduciendo los costos de correo. De igual manera se pueden utilizar los resultados obtenidos para mejorar los métodos de mercadeo directo, por ejemplo por medio del teléfono. Otro uso que se le ha dado es la estimación de ventas con la previsión de inventarios. Esto quiere decir que la empresa no va a fabricar menos cantidad de producto que demandarán los clientes, y de esta manera no los van a enviar a la competencia.

Aplicaciones financieras y bancarias: ya se han afrontado varios problemas en el sector bancario por medio de la computación neuronal. Entre estos podemos mencionar: evaluación de los préstamos hipotecaros, verificación de firmas de cheques, detección de fraudes en tarjetas de crédito.

Aplicaciones en Medicina: una de las importantes áreas de aplicación de las redes neuronales, es la extracción de conocimiento a partir de un gran volumen de información. En el área de la medicina se ha demostrado que se pueden inferir criterios para realizar diagnósticos sobre los síntomas de una enfermedad a partir de los registros médicos sobre pacientes que han sufrido esa enfermedad.

Reconocimiento óptico de caracteres: una de las aplicaciones que han sido más difundidas es la identificación de la escritura impresa o manuscrita. Los sistemas que realizan este proceso se denominan Optical Character Recognition (OCR). Algunas aplicaciones que se han realizado se relacionan con la lectura de placas de vehículos, y el registro automático de los números de los cheques bancarios. Para reconocer letra escrita a mano se han desarrollado aplicaciones con redes neuronales para leer formularios llenados a mano o la lectura de códigos postales.

Telecomunicaciones: este campo fue uno de los primeros donde se emplearon las redes neuronales con éxito, como se mencionó en una sección anterior se utilizaron como filtros para la ecualización de líneas telefónicas y cancelación de ecos.

Aplicaciones en sistemas eléctricos de potencia: una aplicación que puede mencionarse es la predicción de la demanda de consumo de un sistema eléctrico.

Monitorización de procesos industriales: se han aplicado redes neuronales en procesos de polimerización, el control de hornos de arco eléctrico y se ha conseguido reducir el consumo de energía eléctrica y el desgaste de electrodos.

Motores de búsqueda: hasta el momento no se conocen muchas aplicaciones en ésta área, pero más adelante en esta investigación se evalúa este campo de aplicación, junto con sus ventajas y desventajas.

3. REDES NEURONALES COMO MOTORES DE BÚSQUEDA EN LA WEB

3.1 Introducción

El proceso de búsqueda o recuperación de información en Internet tiene una serie de pasos, como se mencionó en el primer capítulo, estos pasos son los siguientes:

- El recolector o robot se encarga de crear una colección de páginas.
- El indexador hace una comparación entre las palabras clave ingresadas por el usuario y la colección de páginas, y crea un índice con estas palabras.
- El buscador recupera ciertas páginas del índice basado en la consulta realizada, luego realiza un ranking de éstas y consolida los resultados.

Los robots que se utilizan para rastrear la red, pertenecen a un tipo de programas que se denominan agentes, estos agentes trabajan de forma independiente y sin supervisión de alguna persona, por lo que se les atribuye cierto grado de inteligencia. Los agentes utilizan la retroalimentación para mejorar su desempeño a través del tiempo.

Los agentes se basan en la relevancia que determinan los usuarios en la primera búsqueda para ponderar las palabras clave. Algunas herramientas como Direct Hit y Metabusca, utilizan la interacción con el usuario como medio para mejorar la relevancia.

Durante los últimos años se han adoptado varios paradigmas de aprendizaje automático para la recuperación de información y el análisis de textos, como las redes neuronales, el aprendizaje simbólico y los algoritmos genéticos. Lo que se necesita es una forma de aprendizaje automático que no requiera retroalimentación del usuario, esto se puede obtener por el método aplicado a la recuperación llamado "vida artificial".[6]

Las redes neuronales pueden utilizarse dentro de los buscadores para mejorar los procesos de los agentes que recolectan las páginas, ya que éstas pueden aprender y mejorar su desempeño sin necesidad de supervisión. También pueden ser usadas en el proceso de ranking de documentos, para proveer más relevancia en los resultados. En esta sección se evaluarán algunas herramientas que ya tienen incorporadas las redes neuronales dentro de sus procesos.

3.2 Herramientas que utilizan Redes Neuronales

DataparkSearch engine:[7] este sistema es un motor de búsqueda lanzado bajo la licencia GNU y fue diseñado para organizar búsquedas dentro de un sitio Web, grupos de sitios Web, intranet o sistemas locales. DataparkSearch se divide en dos partes. La primera parte es el mecanismo indexador. El indexador camina sobre las referencias del hipertexto html y almacena las palabras encontradas y nuevas referencias dentro de la base de datos. La segunda parte es un CGI que provee búsquedas utilizando los datos recogidos por el indexador.

---

[6] **Métodos y Técnicas para la indización y la recuperación de los recursos de la World Wide Web,** Mª Dolores Olvera Lobo

[7] **DataParkSearch Engine**

Dentro de las características principales de este sistema se puede mencionar que utiliza redes neuronales para realizar el ranking de los documentos.

MSN:[8] El motor de búsqueda MSN empezó a utilizar redes neuronales como parte de su algoritmo de ranking de documentos desde junio 2005. Ellos llaman a esta parte de su motor de búsqueda RankNet.

Hopfield Net Spider: Hopfield Net Spider es un robot que se encarga de recolectar las páginas. Este sistema empieza con un conjunto de URLs que son representados como nodos, luego activa los URLs vecinos, combina los pesos de los enlaces y determina los pesos de los nuevos nodos que descubrió. Este proceso es recursivo y termina hasta que un número requerido de URLs han sido visitados.

3.3 Aplicación de Redes Neuronales en el Ranking de Documentos

3.3.1 Introducción

El propósito principal de un motor de búsqueda es hacer dinero. La forma en que los motores de búsqueda hacen dinero es cuando tienen tráfico y tienen tráfico cuando proveen relevancia en las búsquedas.

[8] **Msn's new ranking system**

Para proveer relevancia en las búsquedas se utilizan complicados algoritmos. Un motor de búsqueda utiliza varios factores para realizar el ranking, por lo que se necesita una fórmula que combine dichos factores. Utilizando redes neuronales, se puede alimentar al motor de búsqueda con ciertas consultas y recalcular los pesos para que los motores coloquen ciertas páginas al inicio. De esta manera se dice al motor, que páginas deben ir al inicio para cierto tipo de consultas y la red neuronal ajustará la red para que el usuario obtenga los usuarios deseados para estas consultas entrenadas.

La fase de entrenamiento es muy sensible a cambios en el número de factores de entrada y la forma en que son calculados. Ejemplo: si se entrena la red neuronal en un índice que tiene muchas páginas con técnicas de spam, puede producir demasiado peso para esas páginas, lo cual no es favorable. Si luego se cambia la fase de indización para no incluir a las páginas con spam, se obtendrá un índice con mejor calidad, pero las búsquedas pueden no ser tan óptimas porque el motor de búsqueda probablemente castigará sitios con buena información que usen técnicas spam. En este caso, se tendrá que entrenar nuevamente a la red neuronal. [9]

[9] **Search-engine-algorithms**

3.3.2 Algoritmos para realizar el ranking

3.3.2.1 Redes Neuronales Gráficas[10]

La información hipervinculada en Internet es a menudo explotada para propósitos de ranking. Los algoritmos para ranking son usados por los motores de búsqueda para ordenar los URLs de acuerdo a su relevancia para las consultas del usuario. El enfoque más común es PageRank (ordenador de páginas). Éste considera un enlace de la página p a la página q como la aprobación de la calidad de la página q. Formalmente el PageRank (PR_n) de una página n está dado por:

$$PR_n = d * \sum_{u \varepsilon pa[n]} PR_U / h_u + (1-d),$$

en donde $pa[n]$ es un conjunto de padres de la página n , h_u es el grado de la página u , y d, que se usa como valor de atenuación, puede tomar valores entre 0 y 1.

Este enfoque es global en el sentido de que todas las páginas en Internet son tratadas igual. Sin embargo, desde el punto de vista de un usuario, el ranking puede no ser muy óptimo. Por ejemplo cuando uno busca por "Amazon", uno puede esperar ver los resultados de la librería en línea "Amazon" ó "Amazon" el río en Suramérica. Un usuario puede estar más interesado en páginas del río que en la tienda en línea. Mucho trabajo fue realizado en cómo modificar el método PageRank para reflejar los intereses del usuario. Algunas veces se le refiere a este trabajo como "ranking de páginas web adaptativo".

[10] **Adaptive Page Ranking with Neural Networks** Scarselli, Franco y Liang Yong ,Sweah.

No existen muchos trabajos actualmente que empleen los modelos de las redes neuronales para realizar ranking de páginas adaptativos. Una de las razones principales es porque no se han establecido modelos de redes neuronales capaces de procesar gráficos.

Recientemente se introdujo una nueva clase de modelos de redes neuronales, llamado Redes Neuronales Gráficas (GNN), las cuales son capaces de procesar tipos diferentes de gráficas. Estas redes pueden ser usadas para la clasificación de páginas Web, ya que Internet puede representarse como un grafo en donde los nodos representan las páginas y las aristas denotan los vínculos. Los nodos pueden ser etiquetas numéricas que guardan el contenido de la página codificada.

Una GNN adjunta a cada página n un vector x_n , llamado estado, que recoge la representación de la página. El estado x_n está definido como la solución del sistema de ecuaciones:

$$x_n = \sum_{u \varepsilon pa[n]} A_{n,u} x_u + b_n ,$$

en donde $pa[n]$ es el conjunto de padres de n. Para cada página n, la categoría (rank) r_n está definido por:

$$r_n = (C_n^T) x_n ,$$

en donde T denota la transposición de un vector. Los vectores c_n , b_n y la matriz An, son los parámetros definidos por las salidas de las tres redes neuronales multicapa con retroalimentación. Estas redes neuronales procesan los parámetros usando las etiquetas, que pueden ser por ejemplo, el contenido de una página.

Las últimas dos ecuaciones asumen que la categoría (rank) de la página n depende de su contenido y en las páginas que tienen un vínculo direccionado a n. Esta suposición es parecida a la que tiene el PageRank de Google, con la diferencia que el PageRank no considera el contenido de las páginas sino que solamente los vínculos.

3.3.2.2 RankNet[11]

RankNet es el nuevo algoritmo que utiliza MSN en su motor de búsqueda para hacer ranking de páginas.

La siguiente ecuación corresponde a una red neuronal de dos capas. Para el entrenamiento i, la salida se denota como o_i.

$$o_i = g^3(\sum_j w_{i^3 j^2} g^2(\sum_k w_{j^2 k} x_k + b_j^2) + b_i^3) = g_i^3,$$

en donde para los pesos w y las compensaciones b, los índices superiores indexan la capa y los índices inferiores indexan los nodos dentro de cada capa correspondiente.

Para el problema de ranking se utiliza una red con una sola salida. La función de costo se convierte en una diferencia de la salida de dos entrenamientos consecutivos. Se asume que el primer patrón es conocido como el ranking mayor o igual que el segundo.

[11] **Learning to Rank using Gradient Descent.** Burges, Chris.

Las funciones utilizadas para estos algoritmos son modificaciones de una red de retropropagación. La retropropagación quiere decir que el error se propaga hacia atrás, de la capa de salida hacia la capa de entrada, ajustando los pesos de las conexiones con el fin de reducir el error.

3.4 Aplicación de Redes Neuronales en los Robots que recolectan páginas

3.4.1 Algoritmo del Hopfield Net Spider

Como se mencionó anteriormente este robot utiliza redes neuronales para construir la colección de documentos. A continuación se explica el algoritmo que utiliza.

1. Se inicializa la red con un conjunto de URLs. Cada URL se representa como un nodo de peso = 1. $\mu_i(t)$ se define como el peso del nodo i en la iteración t. Este robot recupera y analiza las páginas en la iteración 0. Los URLs nuevos que se encuentran en estas páginas se agregan a la red.
2. Para la siguiente iteración, el peso de cada nodo se calcula por medio de la siguiente fórmula:

$$\mu_i(t+1) = f_s(\sum w_{h,i}\,\mu_h(t)),$$

En donde $w_{h,i}$ es el peso del enlace entre dos nodos, f_s es la función que normaliza el peso a un valor entre 0 y 1. Luego de calcular los pesos, se debe decidir cuál nodo debe ser activado primero, es decir a que página va a visitar. Los pesos son los que deciden el orden en que las páginas deben ser visitadas.

El conjunto de nodos en la iteración actual son visitados y recuperados en orden descendiente en función del peso. Para poder excluir de la red los URLs de baja calidad, los nodos que tienen un peso menor a un valor θ no son visitados. Luego de que las páginas con peso mayor a θ han sido visitadas e incluidas en la colección, el peso de cada nodo en la nueva iteración se actualiza para reflejar la calidad y relevancia del contenido de la página de la siguiente manera:

$$\mu_i(t+1) = f_s[\mu_i(t+1) * p_i],$$

en donde p_i es el peso que representa la relevancia del contenido de la página i. Como se puede observar en la fórmula, una página con frases más relevantes tendrá un punteo más alto que las demás.

3. Este algoritmo se repite hasta que cierto número de paginas han sido recolectadas o hasta que el peso promedio de todos los nodos es menor a un error máximo permitido.

3.5 Ventajas del uso de Redes Neuronales

- Entrenar a una red no es complicado y provee un excelente desempeño para el problema del ranking de gran cantidad de documentos.
- Las Redes Neuronales han sido una gran ventaja para MSN, ya que pueden ahorrar tiempo en el cálculo de las fórmulas de ranking.
- Las Redes Neuronales Gráficas son capaces de generalizar con un pequeño número de ejemplos.

- Con las Redes Neuronales se obtienen los algoritmos adaptativos que tanto se requieren dentro de los motores de búsqueda, ya que ellas aprenden de sus mismos resultados.

4. COMPARACIÓN DE LAS TECNOLOGÍAS DE RECUPERACIÓN DE INFORMACIÓN MÁS USADAS EN LA ACTUALIDAD

4.1 Introducción

Como se ha mencionado a lo largo de esta investigación, las herramientas para la recuperación de información en Internet están evolucionando, es decir, están incorporando en sus algoritmos los conocimientos de Inteligencia Artificial.

Entre los problemas que han impulsado la evolución de los algoritmos, se puede mencionar: el gran crecimiento de Internet y la Internet Invisible. Los procesos actuales no tienen la capacidad de mejorar ante estos problemas, por lo que se hace necesario incorporar técnicas de Inteligencia Artificial, entre ellas, las Redes Neuronales.

Dentro de las características principales de las redes neuronales, se encuentra el aprendizaje inductivo, por medio de este aprendizaje no se le indica a la red las reglas para dar una solución, sino que extrae sus propias reglas a partir de los ejemplos de aprendizaje y la misma red modifica su comportamiento en función de su experiencia, por esta razón son ideales en los procesos de recuperación de información.

Actualmente no existen muchos motores de búsqueda que utilicen las redes neuronales dentro de sus algoritmos, recientemente el motor de búsqueda de Microsoft MSN, incorporó dentro de sus algoritmos, las redes neuronales para mejorar su proceso de ranking de documentos. El ranking de documentos es una de las partes más importantes y más complicadas de un motor de búsqueda, ya que es el que ordena las grandes cantidades de documentos encontrados en una búsqueda según la relevancia para el usuario.

En esta sección se realizará una comparación de las tecnologías utilizadas para el Ranking entre dos buscadores famosos en la actualidad: Google y MSN, también se realizará una comparación entre dos robots ó recolectores de páginas.

4.2 Comparación en técnicas de ranking

4.2.1 Google

El motor de búsqueda de Google utiliza para el Ranking de documentos PageRank. PageRank es un valor numérico que representa la importancia que una página Web tiene en Internet. Google asume que cuando una página tiene un enlace a otra es un voto para la página enlazada. Mientras más votos tenga una página, será considerada más importante. Este voto también es considerado como el peso.

PageRank es un dato muy importante, ya que por medio de él se ordenan los documentos (según su relevancia) que son el resultado de una búsqueda.

Por medio de éste método una página no puede controlar los enlaces que apuntan hacia ella, pero sí controla los enlaces que ésta tiene hacia otras páginas.

El valor PageRank de una página no se mantiene constante, Google una vez al mes lo recalcula en el Google Dance, que es el período que transcurre entre el comienzo y el fin de la actualización.

Desventajas:

- Una de las desventajas en éste método es que existen algunas páginas que no tienen enlaces de salida, y su peso se pierde en el sistema. Una gran cantidad de páginas en Internet son de este tipo.
- La actualización del PageRank es muy costosa y como se mencionó anteriormente, Google lo realiza una vez al mes.
- Es necesario utilizar otras técnicas para determinar realmente la relevancia de las páginas según la consulta realizada.

Ventajas:

- Se reduce el tiempo de la consulta, ya que solo se realiza una búsqueda de los documentos relevantes y se ordenan de acuerdo a su PageRank.

4.2.2 Microsoft MSN

El motor de búsqueda de MSN utiliza para el Ranking de documentos RankNet. RankNet es una Red Neuronal que utiliza aprendizaje supervisado, por ejemplo para una consulta sobre “noticias personalizadas”, uno de los resultados más relevantes es findory.com. Después de la fase de entrenamiento, se toma una cantidad de datos y se distribuye dentro de la red para tratar de enseñar al sistema a que haga lo correcto, es decir, que siempre coloque en los resultados más relevantes a findory.com cuando se realice una consulta sobre “noticias personalizadas”.

Ventajas:

- RankNet es fácil de entrenar y tiene un gran desempeño en el problema de realizar el ranking en una gran cantidad de documentos.
- Utilizar Redes Neuronales es de gran utilidad, ya que para realizar una actualización del ranking de documentos, solamente se debe entrenar a la red con los nuevos resultados esperados.

4.2.3 Resumen Comparativo

	MICROSOFT MSN	**GOOGLE**
Velocidad	Msn también ofrece velocidad al mostrar los resultados, ya que recupera aquellos con mayor peso.	Google ofrece bastante velocidad ya que solamente realiza la búsqueda de las páginas con el mayor PageRank.
Desempeño en el Ranking	Tiene un gran desempeño al realizar ranking de grandes cantidades de documentos. Además la red es fácil de entrenar para que ella misma modifique los pesos ya que utiliza retropropagación.	La actualización es bastante costosa y toma bastante tiempo, por lo que se realiza una vez al mes.

Tabla II. **Resumen comparativo en técnicas de ranking**

4.3 Comparación en técnicas de recolección de páginas

4.3.1 Googlebot

Este robot fue adaptado para combinar el análisis basado en los enlaces. Los URLs que tienen un mayor punteo de PageRank son los que son visitadas primero. En cada paso o iteración, el robot obtiene el URL con el mayor PageRank, recupera el contenido y extrae los enlaces dentro de la página.

Por el gran crecimiento de la red Google tiene un sistema distribuido de robots. Un servidor de URLs contiene la lista de páginas que proporciona a un cierto número de robots, para que las visite. Usando 4 robots Google tiene la capacidad de obtener 100 páginas por segundo.

GoogleBot genera una estructura de datos basándose en los siguientes datos contenidos en una página:

- Título
- Descripción
- Palabras clave
- Encabezados
- Enlaces
- Texto

Además de basarse en la frecuencia de aparición de algún texto, Google también realiza un análisis de proximidad entre palabras para colocar una mayor ponderación.

Desventajas:

- Como se mencionó en esta sección, el análisis de PageRank se basa en su mayor parte en los enlaces de las páginas, lo que puede llevar a incluir páginas de baja calidad en su índice.
- Ya que este algoritmo se basa en los enlaces, lo hace más lento, porque visita los enlaces que encuentra dentro de las páginas que va descubriendo, aunque no tengan mayor relevancia, lo cual es una gran pérdida de tiempo.

4.3.2 Hopfield Net Spider

El robot Hopfield Net Spider es una Red Neuronal de una capa. Como se vió en la sección 3.4.1, durante primera iteración se tiene un conjunto de páginas con peso=1, para las siguientes iteraciones la red pondera según la relevancia entre los enlaces. Para saber cuál es la siguiente página a visitar se basa en la ponderación, ordenando en forma descendente.

Una de las cualidades de este robot es que la página debe tener un peso mínimo para poder ser visitada, de lo contrario no se toma en cuenta, lo que ayuda a resolver el problema de la relevancia de los documentos indizados. Después de cada iteración y que se han agregado nuevas páginas a la colección, se vuelven a calcular los pesos de cada nodo, para que se pueda reflejar la calidad del contenido de las mismas.

Ventajas:

- Este sistema es eficiente y veloz, ya que primero evalúa el contenido de las páginas y pondera según la relevancia, lo que permite solamente visitar aquellas páginas con mejor contenido, esto quiere decir que el número de visitas se reduce bastante.

4.3.3 Resumen Comparativo

	GOOGLEBOT	HOPFIELD NET SPIDER
Velocidad	Es relativamente lento, ya que visita todos los enlaces encontrados.	Este sistema es más veloz, ya que el conjunto de páginas se reduce, porque no sólo visita las páginas que encuentra sino que evalúa el contenido para decidir si deben ser visitadas.
Desempeño	Por basarse en los enlaces su indice puede contener páginas que no son útiles y que contengan spam.	El índice de este robot es mejor ya que revisa el contenido y en base a éste pondera las páginas. Además si se le entrena a la Red Neuronal a que no incluya páginas con spam, se obtendrá un mejor índice.

Tabla III. **Resumen comparativo en técnicas de recolección de páginas**

4.2 Tendencias

Google, el buscador más famoso y más utilizado en el mundo, utiliza poderosos algoritmos para la recuperación de información, dichos algoritmos hasta el momento no utilizan Redes Neuronales, son solamente algoritmos recursivos que no tienen una forma de aprendizaje como las Redes Neuronales, a las cuales solo se les entrena con una serie de datos y luego aprenden por sí mismas.

En la actualidad, como se ha mencionado anteriormente, se están incorporando técnicas de Inteligencia Artificial en los motores de búsqueda, un caso reciente es el motor de búsqueda de MSN.

MSN incorporó la tecnología de Redes Neuronales en su ranking de documentos, esto quiere decir que el motor de búsqueda aprende de una entrada dada y si es necesario, modifica su comportamiento para proveer mejores resultados, mientras que en este momento Google no tiene una forma de aprendizaje.

Es necesario que los buscadores empiecen a evaluar las diferentes tecnologías de Inteligencia Artificial para aplicarlas dentro de sus motores. Una buena alternativa son las Redes Neuronales, para poder entrenar a la red una sola vez y luego en base a su aprendizaje la misma red sea quien mejore su comportamiento.

Muchas ventajas se obtienen al incorporar Redes Neuronales, entre ellas están:[12]

- Escalabilidad, un algoritmo típico tiene un tamaño finito, mientras que una red neuronal provee la capacidad de crecer, esto se puede comparar con el funcionamiento del cerebro humano, cuando creamos nuevas ideas y pensamientos, estamos creando nuevos caminos para enlazar áreas dentro del cerebro que no habían sido enlazadas. La red realiza el mismo funcionamiento, crea relaciones entre elementos que no estaban enlazados.
- El sistema puede ser entrenado para entender que es lo que se considera como relevante, sin necesidad de modificar un algoritmo sino que solamente se le vuelve a entrenar cuantas veces sea necesario.

[12] MSN Search: The Only One to Implement Neural Networks

CONCLUSIONES

1. Las Redes Neuronales tienen dentro de sus propiedades y características principales: aprendizaje inductivo, procesamiento paralelo. Estas características pueden ser incorporadas dentro de las herramientas de recuperación de información para optimizar su funcionamiento.

2. Las Redes Neuronales pueden ser utilizadas dentro de las distintas fases que componen el proceso de búsqueda, por ejemplo, podrían ser utilizadas dentro de los procesos que realizan el ranking de documentos, para proveer mayor relevancia a los usuarios, ya que, las redes pueden aprender de los ejemplos que se le presenten y mejorar su funcionamiento a través del tiempo.

3. Al utilizar Redes Neuronales dentro del ranking, se reduce el tiempo, ya que, las Redes tienen capacidad de trabajar con gran cantidad de datos y devolver los resultados con más velocidad que los métodos actuales, pues poseen procesamiento en paralelo.

4. Aplicando las Redes Neuronales dentro de los agentes o robots que recolectan páginas, se obtiene un índice de mejor calidad, puesto que la Red evalúa el contenido de las páginas y las pondera según la relevancia, luego propaga el error hacia atrás para mejorar su funcionamiento, de esta manera, el índice sólo cuenta con páginas realmente importantes.

RECOMENDACIONES

1. Las Redes Neuronales tienen la capacidad de trabajar con gran cantidad de datos y con datos incompletos, ya que, sólo se le indica a la red una serie de ejemplos y ésta realiza una generalización. Se recomienda utilizar ésta tecnología para mejorar los procesos contenidos en las herramientas de búsqueda, pues uno de los problemas más grandes que tienen estas herramientas es la excesiva información que se encuentra en la Web.

2. Se recomienda evaluar los distintos tipos de Redes Neuronales y sus características para poder decidir la mejor estructura a utilizar en la construcción de motores de búsqueda en Internet.

BIBLIOGRAFÍA

Referencias electrónicas

1. Anatomía de un gran motor de búsqueda.
   http://www-db.stanford.edu/~backrub/google.html, Agosto 2005

2. Ranking adaptativo con Redes Neuronales.
   http://www2005.org/cdrom/docs/p936.pdf, Agosto 2005

3. PageRank de Google
   http://google.dirson.com/pagerank.php, Septiembre 2005

4. http://www.maestrosdelweb.com/editorial/googlehis/, Septiembre 2005

5. http://www.seroundtable.com/archives/002143.html, Septiembre 2005

6. http://www.alianzo.com/blogs/redessociales/2005/06/29/las_busquedas_sociales_de_yahoo, Septiembre 2005

7. http://www.baquia.com/noticias.php?id=1150, Octubre 2005

8. El Poder de Buscar
   http://mariachi.dsic.upv.es/uimp2004/Baeza/Baeza.pdf, Diciembre 2005

9. Web Spiders Personalizados
   http://faculty.ist.unomaha.edu/pdasgupta/courses/csci8980/papers/chau99.pdf, Diciembre 2005