DETECCIÓN DE IDIOMA DE SITIOS WEB MEDIANTE REDES NEURONALES

Piedad Garrido[1] **Francisco J. Martinez**[1] **Francisco J. Vela**[2] **Jesus Tramullas**[1] **Inmaculada Plaza**[1]

[1] Universidad de Zaragoza, {piedad, f.martinez, tramullas, iplaza}@unizar.es
[2] Universidad de Zaragoza, fjvela@gmail.com

Resumen

La detección del idioma de un documento puede tener especial importancia, sobretodo en entornos donde se trabaja con grandes volúmenes de documentos escritos en diferentes idiomas y que se desean clasificar. Normalmente esta detección se realizaba o de forma manual, o usando métodos estadísticos con computadores. En este artículo se presenta un proyecto que hemos realizado que permite identificar de forma automática el idioma de las páginas web, usando una nueva metodología basada en redes neuronales. Ha sido necesario desarrollar tres aplicaciones: (i) la primera ayuda a la creación, entrenamiento, proyección y visualización de redes neuronales, (ii) la segunda recoge y ajusta los datos, y (iii) la tercera sirve para comprobar si la red neuronal está bien entrenada, hasta alcanzar una tasa de fallos que pueda ser asumida. Los resultados demuestran que el uso de esta metodología da muy buenos resultados con páginas web de diferentes idiomas.

Palabras Clave: Internet, detección de idioma, Redes Neuronales.

1 INTRODUCCIÓN

La identificación del idioma es el proceso por el cual se determina cuál es el idioma en el que está escrito un texto dado. Tradicionalmente, la identificación del idioma se ha llevado a cabo en textos escritos, de forma manual y en entornos de bibliotecas. Se realizaba identificando palabras que aparecían con frecuencia, o identificando letras y signos de puntuación que son característicos en ciertas lenguas.

Desde hace un tiempo a esta parte, se han incorporado los llamados métodos computacionales (en los que se hace uso de los ordenadores) para resolver el problema, usando métodos de procesado de lenguaje natural basados en métodos estadísticos. De esta forma, se pueden clasificar multitud de documentos en función del idioma de una forma más rápida, barata y eficiente. Además, con la aparición de los documentos electrónicos, la detección automática del idioma ha pasado a tener una mayor importancia, al no estar sólo enmarcado en el ámbito bibliotecario, sino también a cualquier organización que tenga que trabajar con grandes volúmenes de documentos escritos en diferentes idiomas.

Desde hace tiempo, los científicos han estudiado el funcionamiento del cerebro, intentando desarrollar algunos modelos matemáticos que tratan de simular su comportamiento, ya que es capaz de procesar a gran velocidad grandes volúmenes de información, combinarla o compararla con información almacenada y dar respuestas adecuadas. Estos modelos se han basado en los estudios de las características esenciales de las neuronas y sus conexiones.

Las redes de neuronas artificiales (denominadas habitualmente como RNA o ANN, en inglés) [5] son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Este paradigma nos permite afrontar problemas de difícil solución con la programación tradicional. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Esta salida viene dada por tres funciones: (i) una función de propagación (también conocida como función de excitación), (ii) una función de activación, que modifica a la anterior, y (iii) una función de transferencia, que se aplica al valor devuelto por la función de activación y se utiliza para acotar la salida de la neurona.

Tabla 1: Aplicaciones para detectar idiomas.

Nombre	Tipo	Lenguaje de desarrollo	Licencia
Word	Aplicación	C++	Propietario
Lingua-Identify	Librería	Perl	Libre
TextCat	Librería	Perl	Libre
Rosette Language Identifier	Aplicación	No determinado	Propietario
Petamem Language Identifier	Aplicación	No determinado	Propietario
lid	Librería	C/C++	Propietario
Talenknobbel	Aplicación	No determinado	Propietario

En este artículo, se presenta el proyecto que hemos realizado que permite identificar de forma automática el idioma de las páginas web, usando un algoritmo basado en redes neuronales. Este documento está organizado de la siguiente manera: en la Sección 2 se presentan las diferentes aplicaciones y métodos que normalmente son utilizados para la detección del idioma tanto en los documentos textuales, como en las páginas web. La Sección 3 presenta el software desarrollado, así como una breve descripción de las diferentes aplicaciones por las que está compuesto. Los resultados obtenidos se detallan en la Sección 4. Finalmente, la Sección 5 presenta las conclusiones más importantes y el trabajo futuro.

2 DETECCIÓN DEL IDIOMA

Existen diferentes aplicaciones para detectar el idioma de un texto escrito. Por un lado, se puede trabajar con la funcionalidad de una aplicación que use diccionarios ortográficos y gramaticales, reglas de puntuación, además de otros parámetros externos como el idioma del teclado, etc. Un ejemplo sería un procesador de textos como por ejemplo Microsoft Word. Por otro lado, se puede hacer uso de una aplicación externa para identificar el idioma del texto. Ejemplos de esta última forma son: (i) Lingua-Identify [13], (ii) TextCat [9], (iii) lid [7], (iv) PetaMem [12], (v) Rosette [1] y (vi) Talenknobbel [14].

En la Tabla 1 aparecen algunas de las aplicaciones más conocidas. Este tipo de aplicaciones suelen utilizar algunas técnicas entre las que destacamos:

- Técnica de palabras pequeñas. Esta técnica consiste en ir buscando dentro del texto las palabras más comunes de una lengua, como los pronombres, los artículos, etc. Es la opción más recomendada para textos grandes.
- Análisis de Prefijos. Analiza los prefijos más comunes de los idiomas.
- Análisis de Sufijos. Analiza los sufijos más comunes de los idiomas.
- Categorización Ngram, donde n representa las subsecuencias de n elementos. Dada una secuencia de letras se construyen las posibles palabras y se comparan.

A diferencia de la detección de idiomas en textos, para la detección de idiomas de sitios web, se habla de métodos y no de aplicaciones, ya que los buscadores aplican una serie de técnicas para comprobar el idioma de un sitio web y así indexarlo en su sistema.

En la Tabla 2 se muestran las técnicas que tradicionalmente se usan, haciendo especial hincapié en sus ventajas e inconvenientes. De todas ellas, la técnica más usada por los buscadores es la de los diccionarios de palabras.

Cabe destacar, que ninguna de esas técnicas parece lo suficientemente buena, pues todas presentan ciertos inconvenientes, que en algunos casos pueden ser verdaderamente críticos.

Por otro lado, Google ha puesto a disposición de los usuarios dos productos: (i) una API desarrollada en AJAX con la que acceder mediante JavaScript a un conjunto de herramientas de traducción, detección y transliteración de idiomas [3], y (ii) el Google Translator Toolkit [4] un servicio que ofrece un conjunto de herramientas para facilitar las traducciones, permitiendo el trabajo colaborativo, compartir glosarios, etc.

A diferencia de las propuestas existentes, este trabajo presenta un conjunto de aplicaciones que permiten detectar el idioma de las páginas web haciendo uso de una red neuronal basada en los mapas autoorganizados de Kohonen [6], un sistema mucho más cómodo, fácil de mantener y válido, tanto para trabajar con información textual como con información ubicada en sitios web.

Tabla 2: Técnicas de detección de idiomas usadas en Internet.

Técnicas	**Ventajas**	**Inconvenientes**
Cabeceras páginas web	Identificación rápida	Escasa utilización, intervención de las personas
Extensiones del dominio	Identificación rápida	Páginas multiidioma, dominios genéricos (.com, ...)
Direcciones IP	Identificación rápida	Almacenamiento en hosting extranjeros
Diccionarios de palabras	Gran número de aciertos, organización de las páginas en subdominios	Coste temporal

3 DESCRIPCIÓN DE LA APLICACIÓN

Este apartado presenta el proyecto desarrollado para la detección de rasgos de páginas web mediante redes neuronales. La arquitectura software escogida es una arquitectura cliente/servidor de tres capas, que facilita la reducción de tráfico en la red, requiere baja potencia de procesamiento por parte del cliente, y un bajo coste de desarrollo y mantenimiento.

Definidos y desarrollados los distintos escenarios, se identifican y se organizan las clases relevantes del sistema mediante el diagrama de clases (ver Figura 1), donde se muestra la relación de los objetos modelados que componen cada una de las aplicaciones, describiendo sus características, estructura y comportamiento.

El proyecto desarrollado está compuesto de tres aplicaciones desarrolladas con software libre: (i) la aplicación *Self-Organizing Maps Framewok* (SOMF), (ii) la aplicación *Detección de Rasgos de Páginas Web (DRPW)*, y (iii) una aplicación que permite recoger direcciones web en diferentes idiomas para utilizarlas como ficheros de entrada de datos. A continuación se presentan más en detalle.

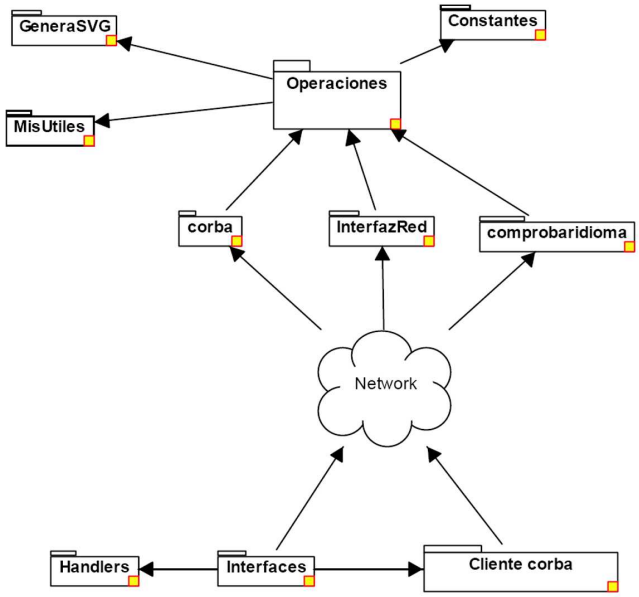

Figura 1: Paquetes que componen la aplicación.

3.1 Aplicación SOMF

El principal objetivo de la aplicación SOMF es el de proporcionar una herramienta software, que facilite el manejo de redes neuronales utilizando como base el paquete SOM-PAK (Self-Organizing Maps Program Package) [8].

La aplicación SOMF consta de cuatro operaciones principales: (i) la opción operaciones de red, donde el usuario podrá crear, entrenar y visualizar una red en formato SVG, (ii) la opción operaciones de ficheros, donde el usuario podrá visualizar los diferentes archivos generados por la aplicación, ya sean

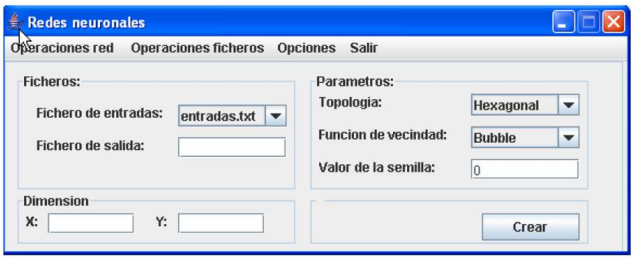

Figura 2: Interfaz de la aplicación SOMF. Crear una red neuronal.

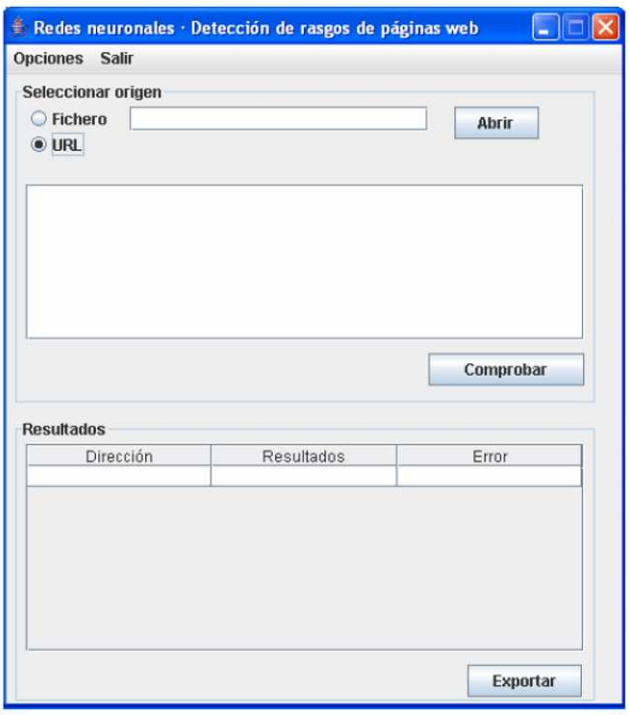

Figura 3: Interfaz de la aplicación DRPW. Pantalla principal.

de texto o visuales, (iii) el apartado de opciones, donde se pueden configurar aspectos tales como la sincronía, la configuración de parámetros de conexión, etc., y (iv) la opción que nos permite salir de la aplicación de forma correcta cerrando todos los hilos de ejecución. La Figura 2, presenta un ejemplo de la interfaz.

3.2 Aplicación DRPW

El principal objetivo de la aplicación DRPW, es el de proporcionar una herramienta software que facilite la identificación del idioma de una página web a través de una red neuronal. A través de la interfaz de esta aplicación (ver Figura 3) el usuario podrá introducir una o varias direcciones URL para que la red neuronal detecte el idioma de la página web. Estas direcciones se podrán introducir de dos maneras diferentes: desde un fichero generado previamente por el usuario, o escribiéndola directamente en la interfaz de la aplicación. Una vez especificada esta información, el programa automáticamente mostrará los resultados al usuario dando la posibilidad de exportar estos resultados a un fichero con formato CSV (Comma Separated Value File Format) [2].

Figura 4: Frecuencia de aparición de los caracteres en castellano (en verde) y japonés (en rojo).

3.3 Aplicación de Recogida de URLs

El principal objetivo de la aplicación de recogida de URLs es el de encargarse de recoger direcciones web en diferentes idiomas, para después poder generar los ficheros de entrada y ahorrarle al usuario la tediosa tarea de ir introduciendo las direcciones, una a una, a través de la interfaz.

En la Tabla 3 se muestra dónde estarían ubicados en la arquitectura Cliente/Servidor, los componentes que forman cada una de las aplicaciones desarrolladas y detalladas con anterioridad.

4 RESULTADOS

Para poder realizar las pruebas, en primer lugar ha sido necesario construir un fichero para poder entrenar la red neuronal. En nuestro caso utilizamos cien páginas por cada idioma. Juntando las estadísticas obtenidas con dos idiomas (castellano y japonés) para el primer juego de pruebas, construimos el fichero de entrada. En la Figura 4 se muestra gráficamente la frecuencia normalizada de aparición de los distintos caracteres en las páginas en castellano (en verde) y las páginas en japonés (en rojo).

En la Tabla 4 aparecen los valores introducidos para ajustar los parámetros de configuración, con la finalidad de establecer un juego de pruebas y entrenamientos válidos, cuyos resultados quedan reflejados en la Tabla 5.

Tras varios entrenamientos, se detectaron algunos problemas a resolver:

Tabla 3: Ubicación de las aplicaciones.

Aplicación	Localización	Función
Interfaz DRPW	Cliente	Se ocupa de interactuar con el usuario y de comunicarse con el servidor DRPW
Interfaz SOM	Cliente	Se encarga de interactuar con el usuario y de comunicarse con el servidor SOMF
Interfaz DRPW	Servidor	Se ocupa de atender las operaciones solicitadas a través de la interfaz DRPW (lógica de la aplicación, manejo de redes SOMF)
Interfaz SOM	Servidor	Se ocupa de atender las operaciones solicitadas a través de la interfaz SOMF (lógica de la aplicación, calcula las estadísticas y proyecta la nueva entrada sobre una red)
Recogida URLs	Servidor	Esta aplicación es la encargada de recoger direcciones web en diferentes idiomas para después poder generar los ficheros de estadísticas
Cálculo de estadísticas	Servidor	A través de este grupo de aplicaciones se generan los diferentes archivos necesarios para el entrenamiento de la red neuronal

- Si no se hace nada para evitarlo, se tienen en cuenta caracteres que no son representativos para la deteccion del idioma de una página web. Estos caracteres son el salto de línea y el salto de párrafo. Tan sólo están en el texto para mejorar el aspecto de la página, por lo que deben ser eliminados.

- Tal y como se observa en la Figura 4, existe un gran número de caracteres que tienen unos valores muy pequeños por lo que la red no los estará teniendo en cuenta, ya que no conseguirá diferenciarlos. Para ello, aplicamos el logaritmo a los valores para que la red sea capaz de distinguirlos y así aumentar los aciertos.

- Gracias a un proceso de refinado de parámetros, se han ido obteniendo resultados cada vez más precisos comparando distintas lenguas romances, germánicas y aglutinantes (japonés), consiguiendo un porcentaje de aciertos que supera el 90%.

Tabla 4: Parámetros usados para realizar las pruebas.

Idiomas estudiados	Castellano, Japonés Francés, Alemán
Dimensión de la red (X)	5, 6, 10
Dimensión de la red (Y)	5, 6, 10
Valor de la semilla	valores aleatorios
Función de vecindad	Bubble
Topología	Hexa
Número de iteraciones	1000, 6000 10000, 15000
Factor de aprendizaje	0.02, 0.05
Radio de aprendizaje	2, 3, 4, 5, 8

5 CONCLUSIONES Y TRABAJO FUTURO

Para la realización de este trabajo se han programado en Java los algoritmos que permiten la detección del idioma de una página web a través de una red neuronal. Una vez concluida la fase de desarrollo de las aplicaciones necesarias para la captura de direcciones url, el cálculo de estadísticas, la aplicación gráfica donde se visualiza el entrenamiento y los resultados, la aplicación DRPW y la aplicación SOMF, se entrenó la red neuronal con diferentes parámetros para buscar los más adecuados, consiguiendo un porcentaje de aciertos mayor del 90%.

El presente proyecto ha sido desarrollado para su aplicación en un modelo de calidad orientado a la evaluación de software educativo [10], así como la accesibilidad y usabilidad de sitios web educativos [11]. Debido a los buenos resultados obtenidos en esta primera fase de desarrollo del proyecto, la principal línea de trabajo futuro es realizar una mejora a esta aproximación funcional del soft-computing, basada en redes neuronales, para su adaptación a un sistema híbrido basado en lógica fuzzy que permita predecir los valores de los parámetros de configuración, poder trabajar con más tipos de idiomas, conseguir separar la capa de presentación de los sitios web de la del contenido en aquellos casos que sea imprecisa, para así evitar tener en cuenta caracteres que no sean representativos y facilitar la detección de valores pequeños que sean relevantes y deban ser tenidos en cuenta.

Tabla 5: Resultados obtenidos en las pruebas.

Prueba 1	
Aciertos	Japonés: 58.0882% Castellano: 98.1595% Francés: 97.1014%
	Total: 85.3547%
Fallos	Japonés: 41.9118% Castellano: 1.8405% Francés: 2.8986%
	Total: 14.6453%
Prueba 2	
Aciertos	Japonés: 97.9381% Castellano: 100.0000%
	Total: 98.9691%
Fallos	Japonés: 2.0619% Castellano: 0.0000%
	Total: 1.0309%
Prueba 3	
Aciertos	Alemán: 98.3051% Castellano: 96.6387% Francés: 91.5966%
	Total: 95.5056%
Fallos	Alemán: 1.6949% Castellano: 3.3613% Francés: 8.4034%
	Total: 4.4944%

Referencias

[1] Basis Technology (2009) Rosette Language Identifier. [Online]. Available: http://www.basistech.com/language-identification/

[2] Creativyst Software. (2007) The Comma Separated Value (CSV) File Format. [Online]. Available: http://www.creativyst.com/Doc/Articles/CSV/CSV01.htm

[3] Google. (2009) Google AJAX Language API. [Online]. Available: http://www.google.com/uds/samples/language/detect.html

[4] Google. (2009) Google Translation Toolkit. [Online]. Available: http://translate.google.com/toolkit/

[5] L. D. Jackel, R. E. Howard, H. P. Graf, B. Straughn, J. S. Denker. Artificial neural networks for computing. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 4, no. 1, pp. 61-63, 1986.

[6] Kohonen, T. Exploration of very large databases by self-organizing maps. *International Conference on Neural Networks*, vol. 1, 1997.

[7] Lingua Systems. (2009) lid Language Identifier. [Online]. Available: http://www.lingua-systems.com/products.html

[8] Neural Networks Research Centre. (2009) SOM_PAK and LVQ_PAK. [Online]. Available: http://www.cis.hut.fi/research/som_lvq_pak.shtml

[9] G. van Noord. (2009) TextCat. [Online]. Available: http://www.let.rug.nl/v̈annoord/TextCat/

[10] I. Plaza, R. Igual, J.J. Marcuello, S. Sánchez, F. Arcega. Proposal of a Quality Model for Educational Software. *XX Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE09)*, Valencia, España, junio 2009

[11] I. Plaza, F. Naranjo, F. Arcega, P.Garrido, S. Castellote. Calidad en grupos universitarios de investigación como nexo de unión empresa-universidad-sociedad. *XVII Congreso Universitario de Innovación Educativa en las Enseñanzas Técnicas*, Valencia, España, septiembre 2009.

[12] PetaMem (2009) Language identifier by Petamem. [Online]. Available: http://nlp.petamem.com/langident.cgi

[13] A. Simoes. (2009) Lingua-Identify. [Online]. Available: http://search.cpan.org/ ambs/Lingua-Identify-0.23/lib/Lingua/Identify.pm.

[14] (2009) Talenknobbel. [Online]. Available: http://www.fuzzums.nl/j̃oost/talenknobbel/