Procesamiento del lenguaje Natural, nº 41 (2008), pp. 89-96 recibido 7-05-2008; aceptado 16-06-2008

Clasificación de Páginas Web en Dominio Específico

Web Page Classification in Specific Domain

Francisco Manuel Rangel Pardo
Corex Soluciones Informáticas
Grupo Fivasa
C/ Mayor, 79-B-10
46970 Alaquàs (Valencia)
francisco.rangel@corex.es

Anselmo Peñas Padilla
Dpto. Lenguajes y Sistemas Informáticos
C/ Juan del Rosal, 16
48040 Madrid (Madrid)
anselmo@lsi.uned.es

Resumen: El presente trabajo obtiene una representación novedosa y que proporciona un alto rendimiento en la clasificación automática de páginas Web en dominios específicos. Para ello el estudio se centra en obtener una representación formal de la intencionalidad del autor por transmitir información acerca de la página que crea y que se plasma en la meta-información de la misma, en la estructura de enlaces (Links), y en la Url. Se ha construido una colección de pruebas específica del dominio del teatro y la aproximación presentada ha obtenido unas tasas de rendimiento, medidas tanto por el estadístico F como por el intervalo de error cometido, superiores a los métodos existentes en el estado del arte.
Palabras clave: Clasificación Web, Categorización Web, dominios específicos, intención del autor, meta-información, meta-data, cabecera, enlaces, url, H&L&U

Abstract: This paper obtains a novel representation that provides high performance in the automatic classification of web pages in specific domains. For this the study is focused on obtaining a formal representation of the author's intent to convey information about the web page that he creates and that is reflected in the meta-information of the same page, in the structure of links, and in the URL. A dataset has been built in the specific domain of theater and the aproach presented has obtained a performance raiting, measured both by statistical F and by the interval committed error, higher than existing methods in the state of the art.
Keywords: Web Classification, Web Categorization, specific domains, author's intention, meta-information, meta-data, header, links, url, H&L&U

1 Introducción

La evolución y el crecimiento de la Web y la especialización de sus contenidos sugieren la aparición de nuevas problemáticas en clasificación automática.

Las páginas Web se caracterizan principalmente por permitir a su creador estructurar libremente la información, y ello da la posibilidad de jerarquizar desde dominios generales de conocimiento a dominios más específicos, por ejemplo, especializando un sitio Web académico, dominio general, en sus dominios de información concretos, como por ejemplo el procesamiento del lenguaje natural.

Pues bien, cuanto más específico es el dominio de las páginas mayor es su grado de similitud semántico y sintáctico, debido principalmente a que se suele utilizar un estilo sintáctico común para tratar un tema semántico común, lo que dificulta la clasificación automática basada en contenido, presentando problemas como la elevada redundancia, la ambigüedad, el ruido y la alta frecuencia de aparición de las palabras, haciendo más difusa la diferenciación entre categorías.

Ahora bien, partiendo de que toda página está hecha por un autor, quién tiene como objetivo principal comunicar algo, podemos lanzar el supuesto de que ese algo que quiere comunicar quedará enmarcado en sitios concretos de la estructura de la página.

Con ello, la investigación se ha llevado a cabo en la línea de intentar obtener el

ISSN: 1135-5948

máximo poder discriminatorio de las palabras utilizadas para clasificar, acudiendo para ello al lugar dónde los autores de las páginas tienen mayor intención de comunicar el fin de las mismas, intentando romper de este modo el vínculo semántico entre categorías.

El artículo se estructura en un repaso a la situación de la investigación actual en la sección dos. En la tercera sección se da la descripción de la propuesta teórica planteada. En la cuarta sección se describe la creación de la colección de pruebas y sus características y a continuación en la quinta sección la metodología y el marco teórico de experimentación y evaluación de resultados. La sexta sección muestra la manera de obtener las características y formar el modelo de entrenamiento para en la séptima sección presentar los experimentos llevados a cabo y discutir los resultados obtenidos, dando paso a la octava sección que reúne las conclusiones que se extraen y la novena sección que presenta las líneas de investigación futuras. Por último se adjunta la bibliografía consultada para la redacción del artículo.

2 Trabajos relacionados

El estado del arte muestra que se han explorado gran cantidad de alternativas en la extracción de características de la Web para su posterior tratamiento, aprendizaje y modelización.

Se ha investigado en diferentes representaciones formales de las páginas desde el modelo clásico de bolsa de palabras (BoW), basado en obtener características a partir del contenido de la página, hasta modelos que explotan la estructura de enlaces y meta-datos de la web, y la relación entre páginas, representando las mismas como una serie de características referentes no sólo a su contenido.

El modelo BoW será quizás el primer modelo de representación para la clasificación de documentos, y aunque adolece de importantes suele ser utilizado como *baseline* para comparar las nuevas propuestas realizadas.

Una mejora interesante a la clasificación basada en contenido es aquella que utiliza un resumen del mismo para realizar el aprendizaje y la clasificación, de manera que no sólo se reduce la dimensionalidad del problema sino que se acota el mismo sobre un vocabulario más concreto. Estudios como (Shen 2004) muestran cómo mejora significativamente la clasificación de páginas basadas en resúmenes hechos por humanos, consiguiendo mejoras de hasta un 12,9% sobre la baseline BoW.

Pero la clasificación Web añade una serie de problemas a la clasificación clásica de documentos, derivada principalmente de la variedad de autores que escriben las páginas y de la posibilidad de estructurar los contenidos de muy diversas maneras, pero esta posibilidad de estructuración añade nuevas posibilidades; el html permite obtener nuevas características de las páginas en adición a las clásicas basadas únicamente en el contenido.

(Lindemann 2007) propone una clasificación basada exclusivamente en características estructurales, consiguiendo un 92% en la prueba F en clasificación de páginas de dominios de primer nivel.

(Kan 2004) realiza una clasificación a partir de las palabras que aparecen en la URL de la página obteniendo resultados muy interesantes, de manera similar a como lo hace (Shih 2004) para bloquear *spots* publicitarios dependiendo de su URL y quién además utiliza la estructura tabular de las páginas para determinar qué enlaces pueden ser interesantes para el usuario y qué enlaces pueden ser publicitarios, teniendo en cuenta para ello su posición dentro del árbol tabular de la misma.

Pero la mayoría de líneas de investigación actuales se dirigen hacia el análisis contextual de las páginas, que se divide en tres líneas principales, a saber, análisis del hipertexto, análisis de los enlaces y análisis de la vecindad.

El análisis del hipertexto se basa en extraer características del texto de los anclajes, los encabezados, las páginas a las que apuntan, y en general todos aquellos elementos que estén resaltados de algún modo en la estructura html. Así pues, (Sun 2002) utiliza una combinación de características extraídas del título de las páginas y los anclajes, mostrando un incremento del rendimiento frente a los métodos clásicos basados en el contenido.

La aproximación basada en el análisis de enlaces combina el análisis anteriormente

descrito con el análisis textual de las páginas referenciadas. (Calado 2003) consigue aumentar en 46 puntos el F1 sobre el análisis textual de las páginas. (Slattery 2000), hace uso del algoritmo HITS para explorar la topología de hiperenlaces, o (Joachims 2001) que hace uso de la combinación de funciones núcleo de las máquinas de vectores soporte para utilizar conjuntamente la información textual con el análisis de la *co-citation*

(Chakrabarti 1998) y (Oh 2000) hacen uso del análisis de la vecindad, es decir, utilizan la clasificación de las páginas vecinas para determinar la categoría de los nuevos documentos.

En la línea comercial la clasificación automática se ha centrado principalmente en la detección de contenidos perniciosos para su bloqueo y el control parental, aplicaciones por ejemplo como (CYBERsitter), (GFi Web Monitor) o (Spector Soft).

3 Clasificación basada en la meta-información: la intención del autor de comunicar información acerca de la página

En estudios como (Lewis 1992) se argumenta que las buenas características para la clasificación de texto deben cumplir las siguientes propiedades:

Ser relativamente pocas
Frecuencia moderada de aparición
Tener poca redundancia
Generar poco ruido
Estar sujetas al ámbito semántico de las clases a las que se van a asignar
No ser ambiguas

Varias de estas condiciones no se cumplirán en una clasificación basada en contenido. Por ello, la investigación se ha llevado en la línea de obtener el máximo poder discriminatorio de las palabras utilizadas para ello, y estas son, aquéllas con las que los autores tienen mayor intención de comunicar el fin de las mismas.

3.1 La meta-información de la cabecera

El html se proporciona una serie de meta-datos que permiten indicar información acerca de la página, para informar a los robots de los motores de búsqueda, a servicios automáticos y/o a otros creadores de páginas Web.

En ellos se informa del nombre del sitio, la descripción del contenido y las palabras clave que lo definen, por lo que contienen básicamente la intención del autor por comunicar información acerca de su página.

El principal problema es que no siempre esta meta-información se encuentra informada, tal y como (Pierre 2000) deja latente en un estudio realizado sobre 29.998 Webs, pero en el caso de existir se cumplirán gran parte de los requisitos enunciados para ser buenas características en un problema de clasificación automática.

3.2 La información de los enlaces

En el caso de no existir o no ser representativos se deberá complementar con alguna otra característica en la línea de comunicar la intención del autor, y para ello se tienen los enlaces.

Los enlaces son la vía facilitada por el hipertexto bien para relacionar la información actual con otra información que la complete, complemente o amplíe, bien con otra sección que la desarrolle y estructure.

Siguiendo las pautas de la usabilidad cualquier creador de páginas Web intentará que sus enlaces muestren una descripción detallada del destino de los mismos, de manera que se invite al usuario a seguirlos, sabiendo hacia dónde se dirige, y proporcionándole la información que necesita, aunque no siempre es así, por ejemplo por los mapas de imágenes, los enlaces con imágenes en lugar de texto, enlaces tipo “pinche aquí...”

3.3 La información de la URL

En contra de la recomendación de inocuidad del W3C en las URLs, generalmente se suele utilizar palabras relevantes en ellas para dotar de estructura y significado a las mismas.

La información de la url nos viene dada por dos vías, por la propia URL de la página,

y por las URLs de los enlaces a otras páginas.

Por lo tanto, enlaces tipo imagen o "pinche aquí" que no aportan información en su texto, pueden aportarlo en su URL si es del tipo "...altacliente.php" o "formalizamatricula.aspx".

La combinación de los tres tipos de información anteriores, obtenidos de la cabecera del documento (head), los enlaces (links) y la (url), determinan el método seguido en la investigación actual y que hemos denominado H&L&U.

4 Colección de pruebas

Hemos construido a partir de un conjunto de sitios Web anotados manualmente en el dominio específico del teatro.

La elección de este dominio se justifica desde dos puntos de vista, por un lado desde el punto de vista de la riqueza semántica de su contenido y la similitud entre diferentes categorías, dificultando por tanto la separación de las mismas, y por otro desde el punto de vista técnico de la dificultad de obtener un conjunto de entrenamiento correctamente clasificado y con cierto grado de calidad, y del que se disponía en este trabajo.

Inicialmente se parte de un conjunto de 167 Webs anotadas en 16 categorías diferentes.

Efectuando un crawl hasta el tercer nivel de cada uno de estos sitios se obtiene el conjunto final de 4801 páginas Web.

La colección, que hemos denominado DS, sigue la siguiente distribución de páginas por categoría:

Clase	Total	%
Asociaciones	26	0,54%
Blogs	553	11,52%
Compañías	2611	54,38%
Festivales	747	15,56%
Formación	290	5,42%
Revistas	75	1,56%
Salas	300	6,25%
Textos	182	3,79%
Resto categorías	17	0,35%
TOTAL	4801	100%

TABLA 1: Colección de pruebas DS

Se realiza un experimento, para verificar la adecuación de la colección, consistente en entrenar clasificadores binarios mediante el método BoW estándar y validación cruzada, y se obtienen los siguientes resultados en estadístico F (la sección 5 determina la metodología y marco de evaluación):

	Expandida
Asociaciones	0,135
Blogs	0,776
Compañías	0,900
Festivales	0,745
Formación	0,736
Revistas	0,296
Salas Alternativas	0,659
Textos	0,936

TABLA 2: Estadístico F BoW std

Los valores anteriores muestran que la colección tiene una calidad elevada para tareas de clasificación en la mayoría de categorías excepto en dos, Asociaciones y Revistas, que serán eliminadas para el resto de experimentos.

5 Metodología y marco de evaluación

5.1 Elección del clasificador

Los métodos de aprendizaje inductivo permiten la construcción de modelos que generalizan el comportamiento de los datos dados como evidencia para predecir nuevos datos. En el caso del presente trabajo estos modelos son clasificadores binarios entrenados para discernir cuándo una determinada instancia pertenece a la categoría para la cuál ha sido entrenado.

Para ello se ha elegido el método de aprendizaje menos costoso y que da unos resultados suficientemente buenos, Naïve Bayes, y su implementación concreta en Weka[Weka 2006]

5.2 Técnica de evaluación

En el caso de los clasificadores existen diversas técnicas para evaluar, de las cuales se elige la evaluación de hipótesis basada en precisión, donde se evalúa el porcentaje de error que se comete entre la hipótesis formulada y el valor real, y se guía el aprendizaje para minimizar el número de

errores cometidos, en nuestro caso, a partir del error muestral (Hernández Orallo 2004)

Para realizar la evaluación existen diversas alternativas: uso de la evidencia completa para entrenar y validar, partición entre entrenamiento y pruebas, validación cruzada, etcétera, pero en el caso que nos ocupa el número de páginas que se tiene de cada tipo de categorías, en algunos casos, es muy dependiente del dominio o site al que pertenecen. Es por ello que una validación cruzada puede solapar muchas de ellas, creando un aprendizaje muy ajustado a las mismas, produciéndose algo similar al overfitting, como el problema defendido por autores como (Dietterich 1998) quienes proponen una modificación en la técnica que permita utilizar subconjuntos de entrenamiento independientes. Para limitar la complejidad de esta solución, hacemos una propuesta para validación que llamamos 2x2 y donde se realiza una partición manual entre entrenamiento y validación, de acuerdo a un porcentaje 25/75% y se realiza un doble entrenamiento/validación con cada una de las dos particiones, combinándose el resultado.

5.3 Medidas de evaluación

La evaluación mediante Weka se basa en la precisión. Para ello utiliza una serie de indicativos como TP (*True Positive*), FP (*False Positive*), *Precission*, *Recall* y estadístico F. Se define una matriz de confusión dónde se indican tanto los TP como los FP de las clases evaluadas, y a partir de todos estos datos se puede calcular el intervalo de confianza del error real a partir del error muestral.

La utilización de dos valores, *Precission* y *Recall,* dificulta la comparativa, por lo que se ha optado por una combinación ponderada de ambos, como se efectúa con el estadístico F, eligiendo en este caso la media armónica de ambos:

$$F = \frac{2pr}{p+r}$$

FIGURA 1: Estadístico F

Se efectuará el test t-student de dos colas de las series de estadísticos F de los métodos comparados para un nivel de significación del 95%, dónde la hipótesis nula H_0 será que ambas series son iguales. Si dicho estadístico es superior a un valor tabulado (2,365) que depende del tipo de test (t-student de dos colas), del nivel de significación (95%) y de los grados de libertad (5 en este caso) se rechazará la hipótesis nula, lo que significará que uno de los métodos de representación es superior al otro, y dependiendo del signo, se concluirá cuál.

5.4 Medidas del error

Obtener un intervalo de confianza de la evaluación permite, dada una muestra S de n ejemplos tomada a partir de una función objetivo f con una distribución D, establecer unos intervalos de confianza para el error verdadero (error(h)) de una hipótesis a partir del error de muestra (errorS(h)).

Con ello, a un nivel de confianza de c% se puede determinar el intervalo del error como:

$$errorR(h) = errorS(h) \pm z_c \sqrt{\frac{errorS(h)(1-errorS(h))}{n}}$$

FIGURA 2: Intervalo de error

Dónde z_c se obtiene a partir de la distribución normal y el valor utilizado en las diferentes evaluaciones es de 1,96 equivalente al 95% de certeza.

Con esto se define el marco teórico y formal sobre el que se realizarán las validaciones de los modelos y que permitirá comparar las diversas técnicas.

6 Obtención de características

La representación H&L&U es una representación BoW a partir del corpus obtenido desde la cabecera, los enlaces y la url del documento.

Este corpus es tratado con un proceso de stem, filtrado por stop word list y selección de características, obteniéndose un corpus de aproximadamente 700 palabras.

Por cada palabra p del corpus se crean tres características p_HEAD, p_LINKS, p_URL, dependiendo de los tres sitios dónde se va a buscar, cabecera, enlaces y url.

Se obtiene mediante expresiones regulares las palabras de los meta-datos de la cabecera, se calcula su frecuencia de

aparición y se apunta en la característica p_HEAD correspondiente.

Se obtiene las palabras que aparecen en la URL, se calcula su frecuencia de aparición y se apunta en la característica p_URL correspondiente.

Para los enlaces se realiza la obtención de la frecuencia desde dos propiedades de los mismos: su texto y su url. Se suman ambas frecuencias y se apuntan en la característica p_LINK correspondiente.

De este modo queda determinada la representación formal del documento y que será utilizada para el entrenamiento y validación de los clasificadores.

7 Resultados experimentales

El experimento consiste en obtener la representación H&L&U de las páginas del repositorio DS y realizar la validación 2x2 para comparar los resultados con los obtenidos por el método *baseline* elegido.

La hipótesis de partida H_0 es que la adición de características obtenidas de la meta-información de la página obtiene un incremento significativo del rendimiento de los clasificadores.

Se validará H_0 (aceptándola o rechazándola) mediante la prueba t-student de las series de valores para la prueba F de cada clasificador.

Se calculará el intervalo de confianza al 95% del error cometido.

En ambos casos se separan los resultados en caso de pertenencia y no-pertenencia a la clase.

Los primeros experimentos se orientaron a comprobar la hipótesis de que los métodos tradicionales de clasificación automática existentes en el estado del arte no obtenían resultados significativamente diferentes unos de otros cuando se utilizan en un dominio específico como el teatro.

Se evalúan el BoW estándar, el BoW mejorado con características contextuales, y el BoW de las Urls.

Los resultados de las pruebas t-student de las series de estadísticos F son:

	BoW std vs. BoW improv
PERTENENCIA A LA CATEGORÍA	t = 0,185 < 2,365
NO PERTENENCIA A LA CATEGORÍA	t = 1,438 < 2,365

TABLA 3: Prueba t-student en la clasificación BoW std vs. BoW improv

	BoW std vs. BoW URL
PERTENENCIA A LA CATEGORÍA	t = 0,081 < 2,365
NO PERTENENCIA A LA CATEGORÍA	t = 0,231 < 2,365

TABLA 4: Prueba t-student en la clasificación BoW std vs. BoW url

Los resultados anteriores demuestran que todos los clasificadores tienen un rendimiento medio similar (H_0). Esto se deduce de que los valores t-student obtenidos son inferiores en todos los casos al valor tabulado para los parámetros fijados, lo que indica que las medias de las series se pueden considerar iguales, por lo que se elige el método BoW como *baseline* para las comparaciones.

Se efectúa el experimento comparativo de la *baseline* con la propuesta H&L&U y los resultados se muestran a continuación.

En primer lugar se tabula los resultados de las pruebas F obtenidas para el caso de pertenencia y no-pertenencia a la clase por ambos métodos:

PERTENENCIA A LA CATEGORÍA	**BoW std**	**BoW h&l&u**
Blogs	0.299	**0.663**
Compañías	0.660	**0.825**
Festivales	0.084	**0.740**
Formación	0.157	**0.356**
Salas Alternativas	0.185	**0.406**
Textos	0.814	**0.868**

TABLA 5: Prueba F en la clasificación de pertenencia BoW std vs. BoW h&l&u

No-PERTENENCIA A LA CATEGORÍA	**BoW std**	**BoW h&l&u**
Blogs	0.707	**0.947**
Compañías	**0.706**	0.699
Festivales	0.760	**0.939**
Formación	0.761	**0.909**
Salas Alternativas	0.795	**0.941**
Textos	0.991	**0.994**

TABLA 6: Prueba F en la clasificación de no-pertenencia BoW std vs. BoW h&l&u

En negrita se ha mostrado los mejores valores de las clasificaciones aparejadas.

Como se puede apreciar, el clasificador H&L&U es superior en la mayoría de ellas por una diferencia considerable. Se realiza el

test t-student a un nivel de significación del 95% para corroborar estadísticamente esta afirmación y se muestra a continuación:

	BoW std vs. H&L&U
PERTENENCIA A LA CATEGORÍA	t = 3,310 > 2,365
NO PERTENENCIA A LA CATEGORÍA	t = 2,920 > 2,365

TABLA 7: Prueba t-student en la clasificación BoW std vs. H&L&U

Dado que el valor obtenido en ambos casos es mayor que el valor tabulado, se demuestra que el clasificador H&L&U, obtiene tasas significativamente superiores a la base BoW estándar, y por lo tanto, debido a los test anteriormente efectuados, superior al resto de métodos estudiados.

El análisis de los intervalos de error de cada clasificador se muestra a continuación:

PERTENENCIA A LA CATEGORÍA	BoW std	BoW H&L&U
Blogs	0,256 +- 0,036	0,261+-0,037
Compañías	0,315 +- 0,013	0,221+-0,012
Festivales	0,891 +- 0,022	0,098+-0,009
Formación	0,409 +- 0,058	0,213+-0,050
Salas Alternativas	0,391 +- 0,057	0,612+-0,057
Textos	0,044 +- 0,030	0,033+-0,026

TABLA 7: Intervalo de error en la clasificación de pertenencia BoW std vs. BoW h&l&u

No-PERTENENCIA A LA CATEGORÍA	BoW std	BoW H&L&U
Blogs	0,434 +- 0,015	0,068+-0,008
Compañías	0,162 +- 0,016	0,427+-0,021
Festivales	0,284 +- 0,014	0,104+-0,010
Formación	0,370 +- 0,014	0,156+-0,011
Salas Alternativas	0,323 +- 0,014	0,090+-0,009
Textos	0,016 +- 0,004	0,011+-0,003

TABLA 8: Intervalo de error en la clasificación de no-pertenencia BoW std vs. BoW h&l&u

Como se puede apreciar los intervalos obtenidos para la representación H&L&U son significativamente inferiores a los obtenidos por el método estándar e inferiores en la mayoría de casos al 10%.

8 Conclusiones

El presente trabajo se ha enmarcado en las líneas de investigación actuales para proponer una nueva representación autocontenida de los documentos.

Se ha creado una colección de pruebas en el dominio específico del teatro y se ha demostrado su calidad para la utilización en tareas de clasificación.

Se ha demostrado estadísticamente que los métodos estudiados en el estado del arte no producen mejora significativa sobre la *baseline* en la clasificación en dicho dominio.

Se ha demostrado que la representación propuesta obtiene en el estadístico F, para el dominio específico del teatro, al menos 20 puntos más que la *baseline,* y en ocasiones hasta 60 puntos dependiendo de la categoría, así como reduce el intervalo de error a la mitad en la mayoría de casos y siempre por debajo del 10%

La principal conclusión del trabajo es que los métodos tradicionales de clasificación automática de páginas Web, que bien funcionan incrementando el rendimiento de la *baseline* en dominios más generales, no son válidos para la clasificación en el dominio específico del teatro y que la representación propuesta consigue unos resultados significativamente superiores, dejando latente el interés de seguir investigando y trabajando en esta línea de intentar comprender la intención del autor de la página para clasificarla correctamente.

9 Trabajo futuro

Demostrar la adecuación del método a otros dominios específicos diferentes obtenidos por ejemplo de subconjuntos específicos de colecciones estándar como Cade, WebKB o dmoz ODP.

Fortalecer el método con características complementarias, por ejemplo análisis del texto alternativo.

Incluir tratamiento multilingüe, o investigar en clasificación separada por idiomas.

Agradecimientos

Este trabajo ha sido subvencionado parcialmente por el proyecto QEAVis-Catiex (TIN2007-67581-C02-01) del Ministerio de Ciencia e Innovación.

La línea I+D+i de Corex Soluciones Informáticas ha aportado tecnología y personal en la investigación de este trabajo.

Bibliografía

Attardi, Giuseppe; Gulli, Antonio; Sebastiani, Fabrizio. *Automatic Web Page Categorization by Link and Context Analysis*

Bouckaert, R. *Estimating Replicability of Classifier Learning Experiments* (2004)

Calado, Pável; Cristo, Marco; Moura, Edleno; Ziviani, Nivio; Ribeiro-Neto, Berthier; Adré Goncalves, Marcos. *Combining Link-based and Content-based Methods for Web Document Classification*

Chakrabarti S, Dom B, Indyk P. *Enhaced Hypertext Categorization Using Hyperlinks*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 307-318, Seattle, Washington, June 1998

Cristo, Marco; Calado, Pável; Silva de Moura, Edleno; Ziviani, Nivio; Berthier, Ribeiro-Neto. *Link Information as a Similarity Measure in Web Classification*

Dietterich, T. G. *Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*, 1998

Forman, George. *An Extensive Empirical Study of Feature Selection Metrics for Text Classification*, 2002

Furnkranz J. *Exploiting Structural Information for Text Classification on the WWW*. In Intelligent Data Analysis, pages 487-498, 1999

Glover E.J, Tsioutsiouliklis K., Lawrence S, Pennock, D.M., Flake G.W.. *Using Web Structure for Classifying and Describing Web Pages*. In Proceedings of WWW-02. International Conference on the World Wide Web, 2002

Hernández Orallo, José; Ramírez Quintana, Mª José, Ferri Ramírez, César. *Introducción a la Minería de Datos*, 2004

Joachims T., Cristianini N., Shawe-Taylor J. *Composite kernels for hypertext categorisation*. In C. Broodley and A.Daniluk, editors, Proceedings of ICML-01, 18th International Conference on Machine Learning, pages 250-257, Williams College, US, 2001. Morgan Kaufmann Publishers, San Francisco, US

Joachims. *Learning to Classify Text Using Support Vector Machines. Methods, Theory and Algorithms*. 2002

Kan, Min-Yen. *Web Page Classification Without the Web Page*

Lindemann, Christoph; Littig, Lars. *Classifying Web Sites*.

Oh H.J, Myaeng S.H., Lee M.H. *A practical Hypertext Categorization Method Using Links and Incrementally Available Class Information*. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 264-271. ACM Press, 2000

O'Neill T., Lavoie Brian F., Bennet Rick. *Trends in the Evolution of the Public Web*, 1998-2002.

Pierre**,** John M**.** *Practical Issues for Automated Categorization of Web Sites*. Metacode Technologies, Inc. 2000

Shen, Dou; Chen, Zheng; Yang, Qiang; Zeng, Hua-Jun; Zhang, Benyu; Lu, Yuchang; Ma, Wei-Ting. *Web Page Classification Through Summarization*

Shih, L.K; Karger, D.R. *Using Urls and Table Layout for Web Classification Tasks*

Slattery S., Craven M. *Discovering Test Set Regularities in Relational Domains*. In P. Langley, editor, Proceedings of ICML-00, 17th International Conference on Machine Learning, pages 895-902, Stanford, US, 2000. Morgan Kaufmann Publishers, San Francisco, US

Sun, Aixin; Lim Ee-Peng; Ng, Wee-Keong. *Web Classification Using Support Vector Machine*. , WIDM 2002 Virginia

Weka, *Weka Machine Learning Project*, The University of Waikato, Release 3.4 2006

Zheng, Z; Wu, X, Srihari, R. *Feature Selection for Text Categorization on Imbalanced Data*. ACM SIGKDD Explorations Newsletter, 2004