

# UNA REPRESENTACIÓN FUZZY PARA PÁGINAS WEB

**Víctor Fresno Fernández**  
Escuela Superior de Ciencia y Tecnología  
Universidad Rey Juan Carlos  
v.fresno@escet.urjc.es

**Ángela Ribeiro Seijas**  
Instituto de Automática Industrial  
Consejo Superior de Investigaciones Científicas  
angela@iai.csic.es

## Resumen

En el presente artículo se describe un método de representación de páginas web basado en un modelo espacio vectorial, en el que la función de asignación de peso a cada palabra tiene en cuenta características particulares del lenguaje HTML. A partir de varios conjuntos de etiquetas HTML se definen criterios heurísticos que permiten establecer la relevancia de cada palabra dentro de la página. Se proponen dos métodos de combinación de estos criterios: uno analítico y otro fuzzy. Tanto el método como las dos propuestas de combinación de criterios, se evalúan utilizando la representación, en una tarea de aprendizaje y posterior clasificación de páginas web, mostrando que la combinación fuzzy de criterios reporta mejor comportamiento en este tipo de tareas. Una de las principales ventajas de la representación propuesta es que no requiere un conocimiento previo del dominio, haciéndolo muy apropiado al contexto Internet.

**Palabras Clave:** Representación de páginas web, Modelos espacio vectoriales, Clasificación de páginas web, Lógica Fuzzy.

## 1 INTRODUCCIÓN.

Internet se ha convertido desde hace años en la mayor fuente de información en formato digital disponible en el mundo. El crecimiento de esta red de comunicación global es imparable y la capacidad que se nos ofrece de compartir toda esa información dependerá en gran medida de nuestra capacidad de extraer conocimiento útil de la misma.

Clasificar la información contenida en la red de un modo eficiente se ha convertido en una necesidad y es, por ello, que las técnicas de Recuperación de Información (IR) y de Clasificación Automática de Textos (TC) se estén aplicando a gran escala en este dominio.

Fundamentalmente podemos distinguir dos enfoques en la recuperación de la información contenida en Internet: buscadores web y directorios temáticos.

En el caso de los buscadores de páginas web, se posibilita realizar consultas basadas en palabras clave sobre los contenidos de los documentos, mostrándose a continuación los resultados en una lista ordenada en función de un grado de relevancia. Las principales limitaciones de estos sistemas son la ambigüedad y el gran número de resultados que devuelven [1].

En el segundo caso, los directorios temáticos organizan un pequeño subconjunto de todo el material disponible dentro de una jerarquía de categorías. Ejemplos de estos directorios se pueden encontrar en: Altavista, Hotbot y Infoseek<sup>1</sup> con temáticas generales, o en Intomi y Excite<sup>2</sup>, con directorios más restringidos. En cualquiera de los dos casos, las principales limitaciones de estos accesos a la información son: la subjetividad propia de cualquier clasificación humana y el pequeño dominio que representan frente al universo de información potencial disponible. Es precisamente el tamaño de este universo el que imposibilita, tanto en el caso de los buscadores como en el de los directorios temáticos, la actualización y clasificación manual de contenidos.

La combinación de ambos métodos, es decir, la realización de consultas sobre los contenidos de un directorio se ha convertido en una práctica común para la mayoría de los usuarios, que intentan de este modo minimizar las limitaciones de los sistemas de recuperación de información.

De cualquier modo, tanto en un caso como en el otro, la tarea de representar de un modo adecuado los documentos HTML resulta fundamental para el procesamiento automático de las páginas web. Esta representación deberá ser, en primer lugar, fiel al contenido del documento, incluyendo la información necesaria para poder extraer un conocimiento útil y, a la vez, deberá ser compatible con las entradas a los algoritmos que se empleen a continuación.

---

<sup>1</sup> AltaVista., <http://altavista.digital.com> Hotbot, <http://hotbot.com> Infoseek, <http://infoseek.go.com>

<sup>2</sup> Inktomi, <http://www.inktomi.com/products/search/> Excite, <http://excite.com>

La representación de páginas web se ha realizado bajo diferentes enfoques. En la representación por contexto [18][7][8] se explota la información subyacente en los *links* HTML incluyendo también el análisis de material multimedia. En [9] se realiza un análisis de las URL (Universal Resource Location) y de las etiquetas. En este caso, el fin es extraer información sobre la estructura del documento, al igual que en [17], que también busca información de la estructura, pero que explícitamente descarta el hecho de analizar etiquetas de composición o énfasis. El análisis de las Metaetiquetas también ha sido usado como tarea previa a la representación [13], pero este enfoque adolece de poca generalidad, ya que, según estudios realizados por [6], el número de páginas web que disponen de estos metadatos no supera el 30 por ciento del total de páginas analizadas. En la representación por contenido [1] la información se extrae del texto del documento y no explora ni la estructura del mismo ni la topología de los *links*. Cabe destacar dentro de este grupo las representaciones que se basan en la identificación de conceptos, que exploran técnicas con redes neuronales o análisis semántico y lingüístico.

Nuestra propuesta se enmarca dentro de estos métodos de análisis de contenidos y de medidas de similitud, más concretamente, dentro del modelo de espacio vectorial - *vector space model* [16] -. Este modelo, que se explicará con detalle en la sección 2, está basado fundamentalmente en el llamado principio de independencia que considera independientes entre sí las palabras dentro de un mismo documento. Esta suposición inexacta reduce sin embargo drásticamente la complejidad de cómputo brindando resultados, como veremos, muy aceptables. En el modelo de espacio vectorial, los textos quedan representados como un conjunto de pares (término, relevancia). Con el método propuesto la relevancia para cada palabra/término se calcula a partir de la combinación de un conjunto de criterios que quieren evaluar para cada palabra/término la intención del autor a través del etiquetado HTML.

En trabajos anteriores [6] se mostró una mejora en la calidad de la asignación de pesos, al combinar los criterios de un modo analítico respecto a la bolsa de palabras clásica, donde se considera como relevancia la frecuencia de aparición de cada término en el documento. Asimismo en [15] mostrábamos como nuestra propuesta de representación mejoraba el comportamiento de herramientas comerciales como Copernic Summarizer<sup>3</sup>.

El propósito del presente artículo es mostrar como una combinación de criterios a partir de un sistema de reglas fuzzy suministra una representación más exacta de una página web que la combinación analítica. La comparación se realiza usando las representaciones suministradas por ambos métodos de combinación de criterios, en tareas de aprendizaje y clasificación.

Este artículo se estructura como sigue: en la sección 2 se introduce el modelo de espacio vectorial, en el que se basa nuestra representación; en la sección 3, se explicará la estructura de los documentos HTML y los criterios heurísticos empleados en la representación; en la sección 4 se muestra la combinación de criterios tanto de forma analítica como mediante el sistema de reglas fuzzy; en las secciones 5 y 6 se presentan los resultados experimentales y las conclusiones.

## 2 MODELO DE ESPACIO VECTORIAL

Este modelo ha sido empleado en multitud de sistemas dentro del campo de la IR y la TC. Además de asumir el principio de independencia el modelo no tiene en cuenta el orden en que aparecen los términos, por lo que la semántica de un documento queda reducida a la de las palabras que aparecen en él. Estas suposiciones aunque incorrectas reducen drásticamente la complejidad computacional sin afectar en demasía a la calidad de los resultados [11]. Dentro de este modelo se pueden encontrar excepciones a estas asunciones [4] [5] [2].

Este espacio vectorial se genera a partir de un conjunto de términos definidos a priori y que constituyen vectores base. Cada documento se representa como una combinación lineal de estos vectores y los coeficientes de dicha combinación representan la relevancia del término dentro del documento. La representación en este espacio se basa fundamentalmente en la función de asignación de pesos que se utilice.

El modelo de representación más sencillo en este contexto es el conocido como conjunto de palabras – set of words – o modelo de espacio vectorial binario, donde la relevancia es un 0 o un 1 en función de si el término aparece o no en el documento. Dentro de los modelos no binarios tenemos el llamado bolsa de palabras – bag of words [19]– donde el valor de la relevancia representa la frecuencia de aparición del término en el documento.

Otras funciones de asignación de pesos no binarias son: el factor de frecuencia inversa del documento (TF-IDF [16] ) – cuantas más veces aparece un término dentro de un documento más representativo es, a la vez que cuanto mayor sea el número de documentos en los que aparezca menos discriminante será – ; Information Gain (IG) ; Mutual Information (MI) ; Chi-square ( $\chi^2$ ); NLG coefficient (NLG) ; Odds Ratio (OR) ; Relevance score (RS) y el GSS coefficient . Las expresiones de todas estas funciones se pueden encontrar en [17] y una comparativa entre ellas en [11][3] y [19]. La mayoría de estas funciones mejoran los resultados de la bolsa de palabras, sin embargo requieren un conjunto de documentos etiquetados a priori.

En el caso de representación de páginas web, sería deseable no necesitar este conocimiento a priori, ya que la representación en sí misma puede considerarse como una tarea de IR, siempre que la representación ayude al

---

<sup>3</sup> Copernic Summarizer. 2001: <http://www.copernic.com>

usuario a determinar si esa página web es o no de su interés.

### 3 ESTRUCTURA DE UNA PÁGINA WEB

Las páginas web son los documentos propios de Internet. Están escritos en lenguaje HTML y se construyen como una combinación de etiquetas e información textual que los exploradores web reconocen y visualizan. La principal característica de este lenguaje de marcado es que sus etiquetas están orientadas principalmente a la visualización, salvo unas pocas metaetiquetas – metadatos como palabras clave, resumen del contenido, etc. – que se introducen como información adicional para los buscadores y que no son visibles al usuario [12].

Hay muchos tipos de etiquetas web: referencias a imágenes o archivos, enlaces a otras páginas o atributos textuales entre otras. El método que presentamos en este artículo tiene su base en estas etiquetas textuales, que de alguna forma incluyen la intención del autor. Si determinadas palabras han sido situadas en el título de una página, o resultan enfatizadas de algún modo se puede extraer como conclusión que la intención del autor es la de destacarlas del resto, y es en esta suposición en la que nos apoyamos para, mediante una selección heurística de algunas de estas etiquetas, extraer una serie de criterios que combinamos para obtener la relevancia de cada término dentro de la página.

Los criterios que consideramos para definir nuestra función de asignación de relevancia son los siguientes:

#### 3.1. TÍTULO

Parece obvio que si determinados términos aparecen situados entre las etiquetas <title> y </title> deberíamos considerarlos con una relevancia elevada dentro del documento, sin embargo esta etiqueta no aparece en la mayor parte de las páginas web analizadas en el estudio descrito en [6]. Incluso en muchos casos lo que aparece es resultado de una generación automática de título para la página que no refleja en ninguna medida el contenido del texto de la página web. Así pues el título debe ser un factor a tener en cuenta al calcular la relevancia pero no el único.

#### 3.2. ÉNFASIS

El lenguaje HTML tiene etiquetas cuya función es la de destacar determinadas partes de un texto de otras, como por ejemplo: <b>...</b>, <u>...</u>, <em>...</em>, <i>...</i> o <strong>...</strong>.

El texto señalado con estas etiquetas llama la atención del usuario y en muchos casos basta con tomar estos fragmentos enfatizados para hacernos una idea sobre el contenido de un documento, aunque en otros casos esta derivación no sea tan directa.

### 3.3. POSICIÓN

La posición dentro de un documento en algunos casos puede resultar muy relevante. Dependiendo de cuál sea el estilo del autor de la página web podemos encontrar el esquema clásico de Introducción-Cuerpo-Conclusión. En estos casos es interesante considerar más relevante un término que aparece en la introducción o en la conclusión, frente a otro que aparece en el cuerpo o desarrollo del discurso. No obstante, parte de los documentos HTML contenidos en Internet no siguen esta estructura ni ninguna otra.

### 3.4. FRECUENCIA

El número de veces que aparece un término en un documento debe ser un factor determinante a la hora de establecer su relevancia. Sin embargo, al igual que el resto no se debe considerar como criterio único ya que esto potenciaría a aquellas palabras comodín, es decir palabras muy utilizadas en el desarrollo de un discurso pero que sin embargo no permiten distinguir claramente los contenidos de documentos con temática diferente.

## 4 COMBINACIÓN DE CRITERIOS

Una vez definidos cada uno de los criterios que se van a tener en cuenta para el cálculo de la función de relevancia, en los siguientes puntos se detalla para la combinación de criterios, tanto un método analítico, como uno basado en un sistema de reglas fuzzy.

### 4.1. APROXIMACIÓN ANALÍTICA

En esta aproximación la función de asignación de pesos se construye como una combinación lineal de los criterios anteriores. A continuación se describen las funciones empleadas.

#### 4.1.1. Función de frecuencia en la página

$$f_f(i) = n_f(i) / N_{tot} \quad (1)$$

Donde  $n_f(i)$  representa la frecuencia de aparición del término  $i$  en el documento, y  $N_{tot}$  es el número total de palabras aparecidas en la página.

#### 4.1.2. Función de frecuencia en el título

$$f_t(i) = n_t(i) / N_{tit} \quad (2)$$

Donde  $n_t(i)$  representa el número de veces que aparece el término  $i$  en el título del documento, y  $N_{tit}$  es el número total de palabras aparecidas en el título.

### 4.1.3. Función de enfatizado

$$f_e(i) = n_e(i) / N_{enf} \quad (3)$$

Donde  $n_e(i)$  representa el número de veces que el término  $i$  aparece enfatizado, y  $N_{enf}$  es el número total de palabras enfatizadas en la página.

### 4.1.4. Función de posición

Para computar la relevancia de un término en función del criterio posición, dividimos la página web en cuatro partes iguales, tomando la primera y la cuarta como introducción y conclusión respectivamente, y la segunda y tercera partes como cuerpo del documento.

$$f_p(i) = (3/4 * n_{1,4} + 1/4 * n_{2,3}) / N_{tot} \quad (4)$$

Donde  $n_{1,4}(i)$  y  $n_{2,3}(i)$  representan el número de veces que el término  $i$  aparece en la introducción o conclusión – partes primera y cuarta – y en el cuerpo del documento – parte segunda y tercera - respectivamente.

### 4.1.5. Cálculo de la relevancia. Método analítico

Una vez definidas las funciones, se realizó un estudio estadístico [6] para determinar cuales eran los coeficientes de la combinación lineal más adecuados y los resultados se muestran en la Tabla 1:

$$r(i) = C_1 f_1(i) + C_2 f_2(i) + C_3 f_3(i) + C_4 f_4(i) \quad (5)$$

Tabla 1: Coeficientes de la combinación analítica.

$C_1$ (Frecuencia)	0.30
$C_2$ (Título)	0.15
$C_3$ (Énfasis)	0.25
$C_4$ (Posición)	0.30

Así, la función de asignación de relevancia analítica queda como

$$r(i) = 0.3 f_1(i) + 0.15 f_2(i) + 0.25 f_3(i) + 0.30 f_4(i) \quad (6)$$

Un estudio más detallado sobre estas funciones se puede encontrar en [6].

## 4.2. APROXIMACIÓN BORROSA

La combinación analítica de criterios mejora la representación frente a la bolsa de palabra [6], pero a menudo un criterio toma verdadera importancia al aparecer unido a otro. Esto es debido a las limitaciones de los criterios ya descritas a lo largo de la sección 3. Como ejemplo, un término que aparece en el título no tendrá la misma relevancia si además tiene un alto grado de énfasis que otro que sólo aparece en el título.

Los sistemas de razonamiento borrosos representan el marco más adecuado para capturar el conocimiento experto humano cualitativo y resolver la ambigüedad inherente a cualquier proceso de razonamiento, fusionando el conocimiento y la experiencia en un conjunto de expresiones lingüísticas que tratan con palabras en lugar que con datos numéricos.

Como consecuencia planteamos un sistema basado en reglas borrosas como función de asignación de pesos dentro de un modelo de espacio vectorial para la representación de páginas web.

En este caso las variables lingüísticas que vamos a considerar son: *text\_frequency*, *title\_frequency*, *emphasis* y *global\_position*.

### 4.2.1. Variables lingüísticas

Estas variables lingüísticas representan los mismos criterios que en el caso analítico y servirán de entrada al sistema de reglas borrosas. Como salida del sistema tendremos una única variable, la *relevance*.

Para hacer las reglas independientes del tamaño del documento, las entradas a la *text\_frequency* estarán normalizadas a la mayor de las frecuencias aparecidas en el documento. En el caso de la *title\_frequency* a la mayor de las frecuencias aparecidas en el título y para el *emphasis* a la mayor de las frecuencias aparecidas en los enfatizados. Las Figuras 1, 2 y 3 muestran estas variables lingüísticas.

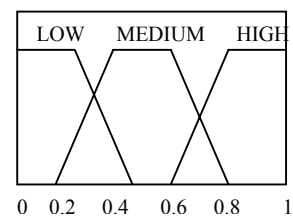


Figura 1: *text-frequency*

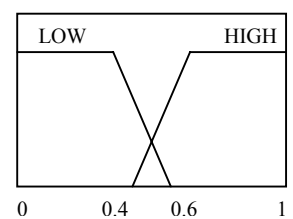


Figura 2: *title-frequency*

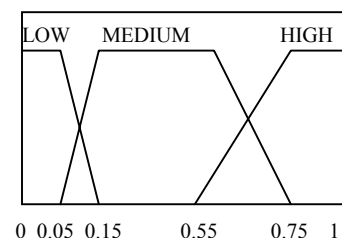


Figura 3: *emphasis*

En el caso de la posición, el valor de entrada a la variable lingüística *global\_position* deberá calcularse de la contribución de todas las posiciones del término en la página. Este nuevo sistema tendrá como variable lingüística de entrada la *line\_position* y como variable de salida la *global\_position*. La contribución de la ocurrencia *o* del término *i* se expresa como:

$$c_p(i,o) = n_p(i,o) / N_{tot} \quad (7)$$

Donde  $n_p(i,o)$  representa el número de línea de la ocurrencia *o* del término *i*, y  $N_{tot}$  el número total de líneas en la página. Por tanto la ecuación 7 se deberá calcular para cada ocurrencia del término o palabra.

Mediante el proceso de *fuzzificación*, a cada ocurrencia se le asigna un grado de pertenencia dentro del conjunto de etiquetas lingüísticas que se muestran en la Figura 4.

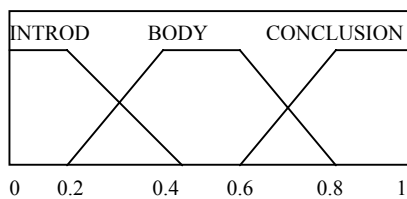


Figura 4: *line-position*

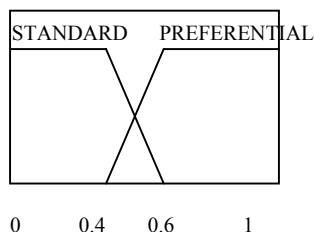


Figura 5: *global-position*

La salida de este sistema auxiliar son las dos etiquetas que se muestran en la figura 5. El conjunto de reglas de este sistema puede encontrarse [15] y en definitiva expresan que si la posición del término de entrada se corresponde con la introducción o la conclusión, entonces la salida es preferente, mientras que si el término se encuentra en el cuerpo del documento la salida es standard

La variable de salida *relevance* del sistema fuzzy general se muestra en la Figura 6.

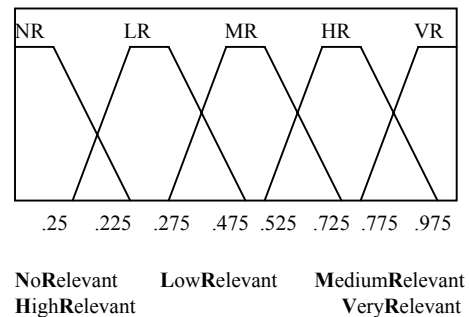


Figura 6: *relevance*

El conjunto de reglas del sistema general, función de asignación de relevancia, se puede encontrar en [15]. Para entender bien la base de reglas IF-THEN debemos tener en cuenta los siguientes puntos:

- El título sólo es relevante si aparece acompañado de algún otro criterio
- La frecuencia en texto pasa a ser muy relevante cuando se combina también con algún otro criterio
- El énfasis tiene una salida generalmente más alta que el resto de los criterios, podríamos decir por tanto que es el criterio más relevante
- Un término con mucha frecuencia de aparición puede significar que es una palabra comodín, en el sentido de que su significado es poco discriminante.

## 5 RESULTADOS EXPERIMENTALES

Las representaciones obtenidas por medio de las aproximaciones analítica y borrosa se usan como entradas en un proceso de aprendizaje supervisado que permite la extracción de las representaciones de dos clases, a partir de un conjunto de páginas web seleccionadas manualmente entre Medicina-Farmacología y Tecnología Aeroespacial. Una vez realizado el aprendizaje se lleva a cabo un proceso de clasificación con el que se estima la calidad de las representaciones de clase obtenidas. En los siguientes apartados se presentan brevemente los principales aspectos del aprendizaje y la clasificación, así como un resumen de los resultados.

### 5.1 FASE DE APRENDIZAJE

Las representaciones de clase se obtienen a partir de un método de aprendizaje supervisado en el que se asume el teorema central del límite, de forma que la clase quedaría representada por una matriz  $3 \times N$  dimensional donde las primeras componentes se corresponderían con los términos que aparecen en los documentos de entrenamiento y la otras dos serían los parámetros de una distribución Normal  $(\mu, \sigma)$ .

La ecuación 8 muestra la función  $f_i$  densidad de probabilidad de un término  $i$ , con relevancia  $r$ , para una determinada clase.

$$f_i(r_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r_i - \mu)^2}{2\sigma^2}} \quad (8)$$

La media  $\mu$  y la varianza  $\sigma$  se obtienen a partir de un conjunto de ejemplos seleccionados para cada clase por estimación de máxima verosimilitud. Para más detalles referirse a [15]

## 5.2 FASE DE CLASIFICACIÓN

La clasificación se lleva a cabo mediante un algoritmo Naive-Bayes, que resultan ser los más efectivos en las tareas de clasificación de textos [10]. La expresión final del algoritmo se expresa en la ecuación 9.

$$c_{ML} = \arg \max_{c_j \in C} \left( \sum_i \ln(s_i) \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{1}{2\sigma_{ij}^2}(s_i - \mu_{ij})^2} + 1 \right) \quad (9)$$

Para ver detalles de la aplicación de este algoritmo [15].

En la Tabla 2 se muestran los resultados de la clasificación obtenidos para dos clases con un conjunto de ejemplos de 200 páginas web seleccionadas manualmente. La selección del conjunto de prueba se hizo tomando páginas de diferentes tamaños y distintos servidores web. Se evitaron también los directorios temáticos para no sesgar los resultados de la clasificación. Por otro lado, se tomó el 65% del conjunto de entrenamiento para la fase de aprendizaje y el resto para la evaluación de la clasificación. Los resultados muestran una mejora en el caso de la representación con combinación de criterios a través de un sistema de reglas fuzzy.

**Tabla 2.** Principales resultados de la clasificación

	Medicina-Farmacología		Tecnología-Aerospacial		Valores medios	
	Aciertos (%)	Fallos (%)	Aciertos (%)	Fallos (%)	Aciertos (%)	Fallos (%)
ANALITICA	82.45	17.55	71.93	28.07	<b>77.19</b>	<b>23.61</b>
FUZZY	84.21	15.79	89.47	10.53	<b>86.84</b>	<b>13.16</b>

## 6 CONCLUSIONES

En este artículo se ha presentado un nuevo método de representación de páginas web y dos modos distintos de obtenerlo. La representación se basa en el modelo de espacio vectorial y la función de asignación de pesos toma algunas características propias del lenguaje HTML y otros atributos más “clásicos”, a menudo utilizados en técnicas de IR. La característica principal es que analiza el texto visible al usuario, y se demuestra que con la fusión de criterios extraídos de la propia naturaleza del lenguaje HTML mejoran las representaciones y más aún si la fusión se realiza de un modo borroso.

Es importante también hacer notar que con el método propuesto la dimensión del vector que representa la página web se ve drásticamente reducida.

### Agradecimientos

Este trabajo ha sido financiado completamente a través de contrato directo con la empresa Innovatec S.A.

### Referencias

- [1] Attardi, G., Gulli A., Sebastiani F. “Automatic Web Page Category Categorization by Link and Context Analysis”. Manuscript
- [2] Billahardt, H.”A Context vector Model for information Retrieval”. Journal of the American Society and technology 53, 3, 2002
- [3] Caropreso, M. F, Matwin S, Sebastiani F. A learner-independent evaluation of the usefulness of statistical phrases for automated text classification. In Text database and Document Management: theory and practice, A. G. Chin, ed. Idea Group Publishing, Hershey, PA, 78-102, 2001
- [4] Cohen, W. W.”Learning to classify English text with ILP methods”. In Advances in Inductive Logic Programming, L. De Raedt, ed IOS Press, Amsterdam, The Netherlands, 124-143, 1995
- [5] Cohen, W. W., Singer, Y.”Context Sensitive Learning methods for Text Categorization”. ACM Trans. Inform. System 17, 2, 141-173, 1999
- [6] Fresno, V., Ribeiro, A.”Feature Selection and dimensionality reduction in Web pages representation” International ICSC Congress on Computational Intelligence: Methods & Applications. Bangor, Wales (U.K.), 416-421
- [7] Guglielmo, E. J. Rowe, N.”Natural Language retrieval of images based on descriptive captions”, ACM Transactions on Information Systems, 14(39), 237-267, 1996

- [8] Harmadas , V., Sanderson, M.,Dunlop, M. D.:”Images retrieval by hypertext links”Proceeding of SIGIR-97, 20<sup>th</sup> ACM International Conference on Research and Development in Information Retrieval, Philadelphia, US, 296-303, 1997.
- [9] Merkl ,D. Text data minig. A handbook of Natural languages Processing techniques and Applications for the Procesising of Languages as Text.dale, R., Moisl H. Y Somers H.(Eds). New York: Marcel Dekker, 1998.
- [10] Mitchell T. M.. Machine Learning. Mc Graw-Hill International Editions.
- [11] Mladenic´, D.”Feature subset selection in text learning”. In Proceedings of ECML-98, 10<sup>th</sup> European Conference on Machine Learning (Chemnitz, Germany, 95-100, 1998
- [12] Musciano C. and Kennedy B., “HTML: The Complete Guide” McGraw-Hill, 1997
- [13] Pierre M.,:”On the Automated Classification of Web Sites”. Linköping Electronic Articles in Computer and Information Science, 6. Linköping University Electronic Press. Linköping, Sweden
- [ 14] Ribeiro, A., Fresno, V.”A Multi Criteria Function to Concept Extraction in HTML Environment” International Conference on Internet Computing IC’2001, Las Vegas (USA),1, 1-6, 2001
- [15] Ribeiro A., Fresno V., M.C.García-Alegre, D.Guinea “A fuzzy system for a web page representation” Intelligent Exploration of the Web (P.S.Szczepaniak, J.Segovia, J.Kacprzyk,, L.A.Zadeh, Eds). Springer-Verlag. Alemania. Publicación a finales de 2002
- [16] Salton G., Buckley C.,”Term weighting Approaches in Automatic text Retrieval”, 1987
- [17] Sebastiani, F.: “Machine Learning in Automated Categorization” ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp 1-47
- [18] Srihari, R. K., “Automatic Indexing and content-based retrieval of captioned iages”, Computer, 28(9), 49-56,1995
- [19] Yang Y., Pedersen J.I.”A comparative study on feature selection in text categorization” In Proceeding ICLM-97, 14<sup>th</sup> Intenational Conference on machine Learning. Nashville, TN.412-420, 1997