

# Clasificación de páginas Web en dominio específico

Web Page Classification in Specific Domain



D. Francisco Manuel Rangel Pardo  
Director: Dr. D. Anselmo Peñas Padilla



# Introducción

---

- Qué son “Dominios genéricos”
  - Académicas, Blogs, Corporativas, Información, Entretenimiento, Personales, Tiendas...
- Qué son “Dominios específicos”
  - Entretenimiento / Teatro / (revistas, compañías, festivales, salas...)
- Motivación
  - > Similitud semántica y sintáctica
  - Métodos tradicionales no son eficientes



# Estructura

---

- ➡ Definición del problema y objetivos
  - Experimentos de clasificación Web
  - Conclusiones y líneas de investigación futuras. Posibles aplicaciones

# Definición del problema y objetivos

## La Web y las necesidades de clasificación

Gran repositorio de información

Muy dinámica

Muy utilizado para consultar



Necesidad de dar orden a toda esta información



Necesidad de clasificación de manera automática



Necesidad de proteger ciertos grupos sociales de contenidos perniciosos



Acercamiento tecnológico a la población

# Definición del problema y objetivos



## La Web 2.0 y las tendencias de futuro

---

Colaboración entre usuarios  
Servicios avanzados

....

Blogs, Wikis...



Buscadores de contenido específico



Necesidad de clasificación de manera automática

# Definición del problema y objetivos



## Objetivos de la investigación

---

- Crear colección de pruebas en dominio específico
  - Determinar marco de evaluación
  - Determinar necesidad de preprocesado lingüístico
  - Fijar marco comparativo con representaciones actuales
- 
- **Proponer representación autocontenida para las páginas Web**



# Estructura

---

- Definición del problema y objetivos
- ➡ Experimentos de clasificación Web
- Conclusiones y líneas de investigación futuras. Posibles aplicaciones

# Experimentos de clasificación Web

## La colección de pruebas

- **Objetivo:** Crear una colección de pruebas para el dominio específico del teatro
- Se parte de 167 sitios anotados en 16 categorías
- Se realiza un crawl para obtener 4801 páginas

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%

# Experimentos de clasificación Web

## La colección de pruebas

- **Objetivo:** Crear una colección de pruebas para el dominio específico del teatro
- Se parte de 167 sitios anotados en 16 categorías
- Se realiza un crawl para obtener 4801 páginas

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%

# Experimentos de clasificación Web

## Determinar la necesidad del crawl

- **Hipótesis:** Necesidad de crawl de las Urls anotadas
- **Método:**
  - Se realiza el crawl
  - Se obtiene Bag of Words
  - Validación cruzada
  - Se comparan tasas de error y estadístico F
- **Conclusión:**
  - Necesidad de crawl

	Sin expandir	Expandida
Asociaciones	0,000	0,135
Blogs	0,211	0,776
Compañías	0,136	0,900
Festivales	0,250	0,745
Formación	0,000	0,736
Revistas	0,000	0,296
Salas Alternativas	0,000	0,659
Textos	0,286	0,936

**Sin expandir:**  $\text{ErrorR}(S) = 0,907 \pm 0,058$

**Expandido:**  $\text{ErrorR}(S) = 0,227 \pm 0,012$

# Experimentos de clasificación Web



## La división de la colección de pruebas

---

- **Objetivo:** Obtener un método de evaluación independiente de los datos
- Se divide colección en dos repositorios independientes por dominio
- Se mantiene un porcentaje aproximado 25/75% de páginas por división
- Algunas categorías como *compañías* tienen una distribución muy desigual
- Se propone una evaluación 2x2 frente a validación cruzada

# Experimentos de clasificación Web

## Determinar el marco de evaluación

- Validación basada en precisión
  - Matriz de confusión (*True Positive, False Positive*)

Clase A	Clase B	<- Clasificado como
TPA	FPA	Clase A
FPB	TPB	Clase B

- Precision y Recall  $p = \frac{TPc}{TPc + FPnc}$   $r = \frac{TPc}{TPc + FPc}$
- Prueba F  $F = \frac{2pr}{p+r}$  t-student 95%
- Intervalos de confianza del error real

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

# Experimentos de clasificación Web



## Determinar el marco de evaluación

---

- Validación cruzada
  - K veces (K-1 entrenamiento / 1 validación)
  - K Subconjuntos de prueba independientes
  - Subconjuntos de entrenamiento con datos compartidos
  - Posibles problemas de sobre-ajuste
- Métodos alternativos
  - Dietterich 5x2
  - Bouckaert NxM
  - Propuesta 2x2

# Experimentos de clasificación Web

## Determinar el marco de evaluación

- **Hipótesis:** La validación cruzada introduce un sesgo en la evaluación
- **Método:** Se validan los modelos por validación cruzada y validación 2x2
- **Conclusión:** Los resultados muestran la necesidad de la evaluación 2x2 para una evaluación más ajustada

	Validación cruzada	2x2
Blogs	0,776	0,303
Compañías	0,900	0,048
Formación	0,736	0,107
Salas Alternativas	0,659	0,228

Validación cruzada: 0,227 +- 0,012

2x2: 0,917 +- 0,008

# Experimentos de clasificación Web



## Determinar necesidades de pre-procesamiento

---

- **Hipótesis:** La selección de corpus y el tratamiento lingüístico reducen la dimensionalidad y aumentan el rendimiento
- **Método:**
  - Bag of Words corpus general y corpus por categoría
  - Bag of Words corpus con stem y sin stem
  - Validación cruzada

# Experimentos de clasificación Web



## Determinar necesidades de pre-procesamiento

---

### ■ **Resultados:**

- Palabras
  - Corpus: 7019 vs. Aprox. 4000
  - Stem: 51292 vs. 7019
- Resultados t-student 95%
  - Corpus:  $1,846 < 2,365$  y  $0,351 < 2,365$
  - Stem:  $-3,098 > 2,365$

### ■ **Conclusiones:**

- Ambas disminuyen la dimensionalidad
- Corpus específico no mejora rendimiento
- Stem empeora rendimiento

# Experimentos de clasificación Web



## Determinar la baseline

---

- **Hipótesis:** La adición de información contextual y url mejora la baseline basada en Bag of Words
- **Método:**
  - Se obtiene la representación BoW
  - Se obtiene la representación BoW mejorada
  - Se obtiene la representación BoW de las urls
  - Se validan 2x2



## Determinar la baseline

---

### ■ **Resultados:**

#### ■ BoW mejorado

- T-student 95%:  $t = 0,185 < 2,365$  y  $t = 1,438 < 2,365$

#### ■ BoW Url

- T-student 95%:  $t = 0,081 < 2,365$  y  $t = 0,231 < 2,365$
- Casos concretos F: 0.084 vs. 0.664

### ■ **Conclusiones:**

- BoW mejorado no mejora la baseline
- BoW Url no mejora la baseline
- BoW Url mejoras considerables puntuales

# Experimentos de clasificación Web



## Propuesta basada en meta-información h&l&u

---

- **Hipótesis:** Combinación meta-información para mejorar la representatividad en dominios específicos
- **Método:**
  - Se obtiene corpus de cabecera y enlaces
  - Se crea una característica por palabra
  - Se triplica la característica para cabecera, enlaces y url
  - Validación 2x2

# Experimentos de clasificación Web

## Propuesta basada en meta-información h&l&u

### ■ Resultados:

PERTENENCIA A LA CATEGORÍA	BoW std	BoW h&l&u
Asociaciones	0	<b>0.016</b>
Blogs	0.299	<b>0.663</b>
Compañías	0.660	<b>0.825</b>
Festivales	0.084	<b>0.740</b>
Formación	0.157	<b>0.356</b>
Revistas	<b>0.036</b>	0.010
Salas Alternativas	0.185	<b>0.406</b>
Textos	0.814	<b>0.868</b>

FIGURA 3.55: 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
Asociaciones	<b>0.958</b>	0.788
Blogs	0.707	<b>0.947</b>
Compañías	<b>0.706</b>	0.699
Festivales	0.760	<b>0.939</b>
Formación	0.761	<b>0.909</b>
Revistas	<b>0.959</b>	0.506
Salas Alternativas	0.795	<b>0.941</b>
Textos	0.991	<b>0.994</b>

FIGURA 3.56: Prueba F en la clasificación de no-pertenencia BoW std vs. BoW h&l&u

- Pertenencia:  
 $t = 3,31 > 2,365$
- No-pertenencia:  
 $t = 2,920 > 2,365$

# Experimentos de clasificación Web

## Propuesta basada en meta-información h&l&u

### ■ Resultados:

PERTENENCIA A LA CATEGORÍA	BoW std	BoW h&l&u
Asociaciones	0	0.016
Blogs	0.299	0.663
Compañías	0.660	0.825
Festivales	0.084	0.740
Formación	0.157	0.356
Revistas	0.036	0.010
Salas Alternativas	0.185	0.406
Textos	0.814	0.868

FIGURA 3.55: 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
Asociaciones	0.958	0.788
Blogs	0.707	0.947
Compañías	0.706	0.699
Festivales	0.760	0.939
Formación	0.761	0.909
Revistas	0.959	0.506
Salas Alternativas	0.795	0.941
Textos	0.991	0.994

FIGURA 3.56: Prueba F en la clasificación de no-pertenencia BoW std vs. BoW h&l&u

- Pertenencia:  
 $t = 3,31 > 2,365$
- No-pertenencia:  
 $t = 2,920 > 2,365$

# Experimentos de clasificación Web

## Propuesta basada en meta-información h&l&u

### ■ Resultados:

PERTENENCIA A LA CATEGORIA	BoW std	BoW H&L&U
Asociaciones	1 +- 0	0,409+-0,205
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
Revistas	0,887 +- 0,079	0,754+-0,014
Salas Alternativas	0,391 +- 0,057	0,612+-0,057
Textos	0,044 +- 0,030	0,033+-0,026

FIGURA 3.57: Intervalo de error en la clasificación de pertenencia BoW std vs. BoW h&l&u

- Pertenencia:  
 $t = 3,31 > 2,365$
- No-pertenencia:  
 $t = 2,920 > 2,365$

No-PERTENENCIA A LA CATEGORIA	BoW std	BoW H&L&U
Asociaciones	0,078 +- 0,008	0,348+-0,014
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
Revistas	0,067 +- 0,007	0,658+-0,014
Salas Alternativas	0,323 +- 0,014	0,090+-0,009
Textos	0,016 +- 0,004	0,011+-0,003

FIGURA 3.58: Intervalo de error en la clasificación de no-pertenencia BoW std vs. BoW h&l&u

# Experimentos de clasificación Web

## Propuesta basada en meta-información h&l&u

### ■ Resultados:

PERTENENCIA A LA CATEGORIA	BoW std	BoW H&L&U
Asociaciones	1 +- 0	0,409+-0,205
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
Revistas	0,887 +- 0,079	0,754+-0,014
Salas Alternativas	0,391 +- 0,057	0,612+-0,057
Textos	0,044 +- 0,030	0,033+-0,026

FIGURA 3.57: Intervalo de error en la clasificación de pertenencia BoW std vs. BoW h&l&u

- Pertenencia:  
 $t = 3,31 > 2,365$
- No-pertenencia:  
 $t = 2,920 > 2,365$

No-PERTENENCIA A LA CATEGORIA	BoW std	BoW H&L&U
Asociaciones	0,078 +- 0,008	0,348+-0,014
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
Revistas	0,067 +- 0,007	0,658+-0,014
Salas Alternativas	0,323 +- 0,014	0,090+-0,009
Textos	0,016 +- 0,004	0,011+-0,003

FIGURA 3.58: Intervalo de error en la clasificación de no-pertenencia BoW std vs. BoW h&l&u

# Experimentos de clasificación Web



## Propuesta basada en meta-información h&l&u

---

- **Conclusiones:**

- Obtiene mejoras significativas sobre la baseline
- Método sensible a entrenamiento/validación “defectuosa”



# Estructura

---

- Definición del problema y objetivos
- Experimentos de clasificación Web
- ➔ Conclusiones y líneas de investigación futuras. Posibles aplicaciones

# Conclusiones y líneas de investigación futuras



## Conclusiones

---

- **Se ha fijado:**
  - Colección de pruebas, marco de evaluación y Baseline
- **Se ha propuesto el método h&l&u:**
  - Mejora baseline hasta en 70 puntos de F
  - Sensibilidad a determinados entrenamientos/validaciones

# Conclusiones y líneas de investigación futuras



## Líneas de investigación futuras

---

- **H&L&U:**

- Extensión del método h&l&u fuera del dominio del teatro.
- Nuevas características para mayor estabilidad
- Mejorar tratamiento lingüístico
- ¿Análisis de las imágenes?

# Conclusiones y líneas de investigación futuras



## Posibles aplicaciones

---

- Corex/Grupo Fivasa:
  - Restricción de los empleados a determinados contenidos
  - Búsqueda de información en la Web por especialidad médica
- Otras:
  - Pederastía: Clasificación de páginas pornográficas ilegales
  - Combinación con análisis de página: antispam

# Clasificación de páginas Web en dominios específicos



Muchas gracias por su atención

¿Preguntas?

[francisco.rangel@corex.es](mailto:francisco.rangel@corex.es)