WEB SPAM IDENTIFICATION THROUGH LANGUAGE MODEL ANALYSIS

Juan Martinez-Romo and Lourdes Araujo
Natural Language Processing and Information Retrieval Group at UNED * nlp.uned.es

Fifth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'09)
Madrid, Spain. April 21, 2009
Palacio Municipal de Congresos

OUTLINE

- Motivation
- Previous Works
- Language Models
- Sources of Information
- Classification
- Results
- Conclusions and Future Work

MOTIVATION

- Problem Statement
  - Hyperlinks between topically dissimilar pages

- Undeserved PageRank (Spam or Navigational links)
- Links to unrelated pages (links to owners, maintainers)
- Content spam (text with no meaning for humans)
- Malicious unrelated anchor text (deceptive links)

PREVIOUS WORK

- Blocking blog spam with language model disagreement.
  G. Mishne, D. Carmel, and R. Lempel. AIRWeb'05
  - Blog spam detection. Original post and comments.
- Detecting nepotistic links by language model disagreement.
  A. A. Benczúr, I. Bíró, K. Csalogány, and M. Uher. WWW'06
  - Detect nepotistic links. A link is down-weighted if LMs have a great disagreement between anchor text and pointed page.
- Measuring similarity to detect qualified links.
  X. Qi, L. Nie, and B. D. Davison. AIRWeb'07
  - Qualified link analysis. Several similarity methods and sources of information

LANGUAGE MODEL DISAGREEMENT

- Unigram language model for text (D) in collection (C):

$$p(w \mid D) = \lambda \frac{tf(w, D)}{\sum_{v \in D} tf(v, D)} + (1 - \lambda) \frac{tf(w, C)}{\sum_{v \in C} tf(w, C)}$$

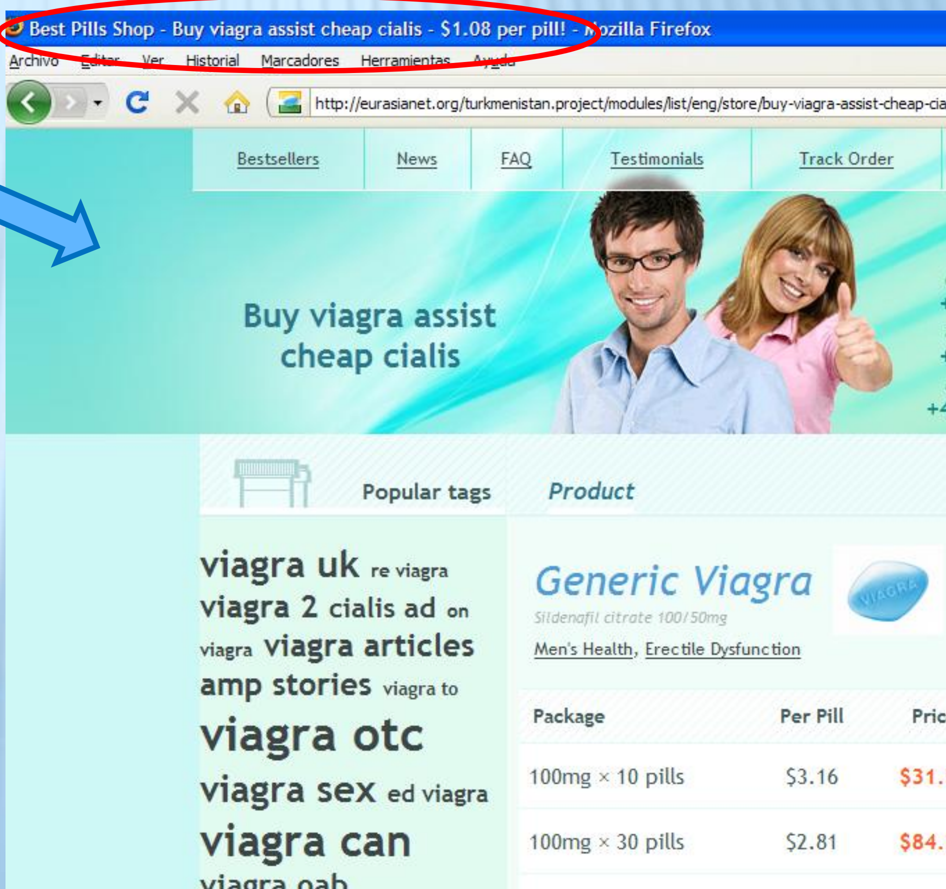
- Kullback-Leibler divergence (KLD) between the language model of two text units from target and source pages:

$$KLD(T_1 \| T_2) = \sum_{t \in T_1} p(t \mid T_1) \log \frac{p(t \mid T_1)}{p(t \mid T_2)}$$

LANGUAGE MODEL DISAGREEMENT

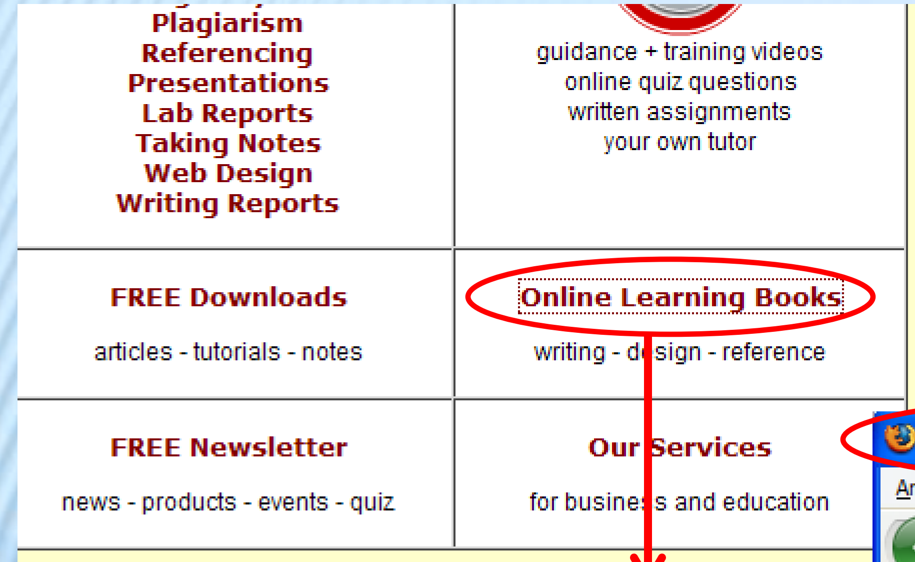
Best UK Reviews

Best Pills Shop - Buy viagra assist cheap cialis - $1.08 per pill!

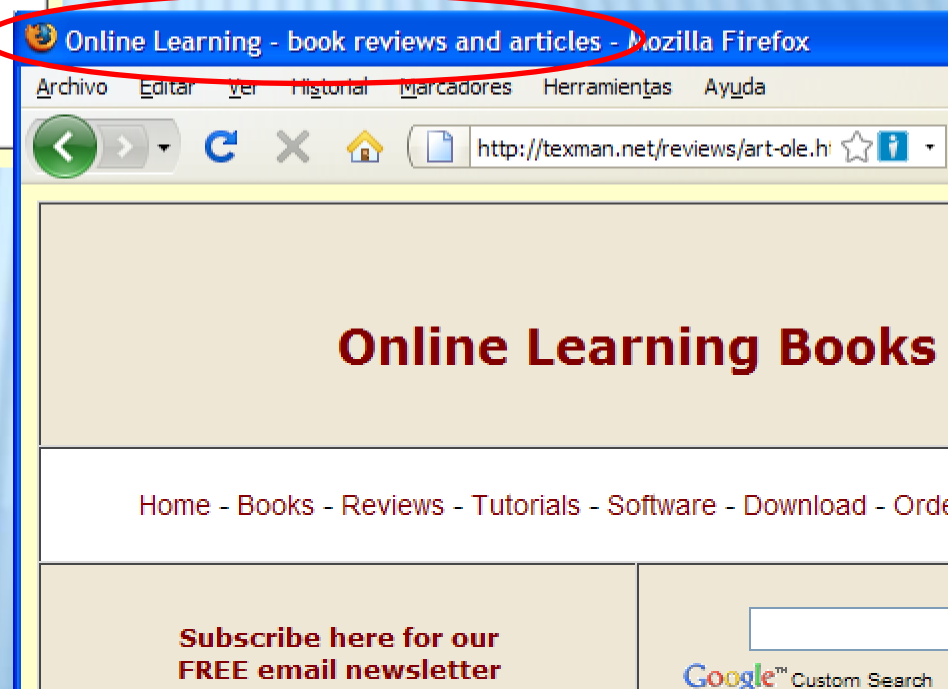


$$KLD(\text{Anchor} \,\|\, Title) = 3.36$$

LANGUAGE MODEL DISAGREEMENT



Online Learning – book reviews and articles

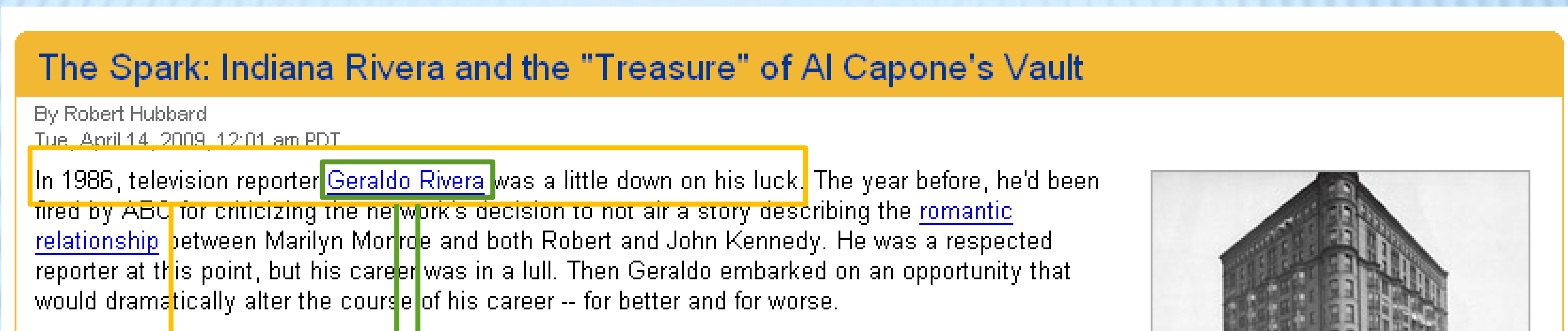


Online Learning Books

$$KLD(\text{Anchor} \| \, Title) = 0.48$$

SOURCES OF INFORMATION

SOURCE PAGE


The Spark: Indiana Rivera and the "Treasure" of Al Capone's Vault

By Robert Hubbard
Tue, April 14, 2009, 12:01 am PDT

In 1986, television reporter Geraldo Rivera was a little down on his luck. The year before, he'd been fired by ABC for criticizing the network's decision to not air a story describing the romantic relationship between Marilyn Monroe and both Robert and John Kennedy. He was a respected reporter at this point, but his career was in a lull. Then Geraldo embarked on an opportunity that would dramatically alter the course of his career -- for better and for worse.

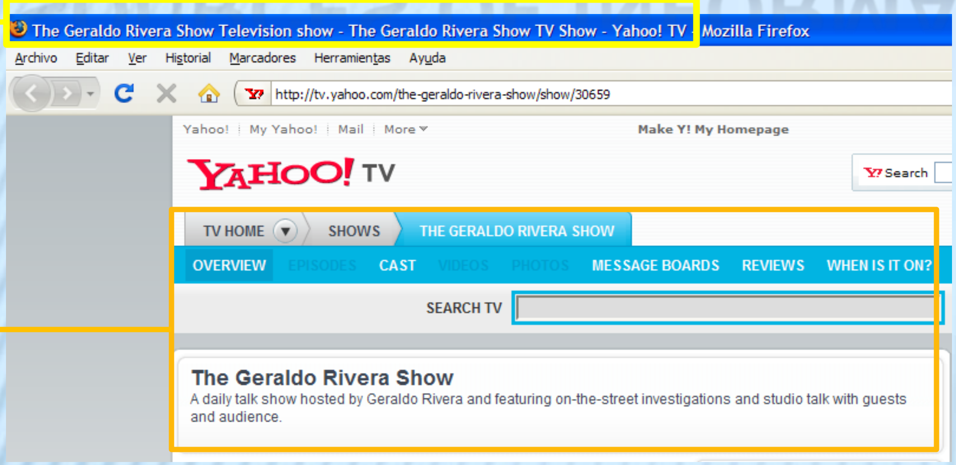
Source of Information	Text
Anchor Text	Geraldo Rivera
URL terms	***tv***.*yahoo.com/the-**geraldo-rivera**-**show**/show/*
Surrounding Anchor Text	In 1986, **television reporter** Geraldo Rivera was a little down on his luck.

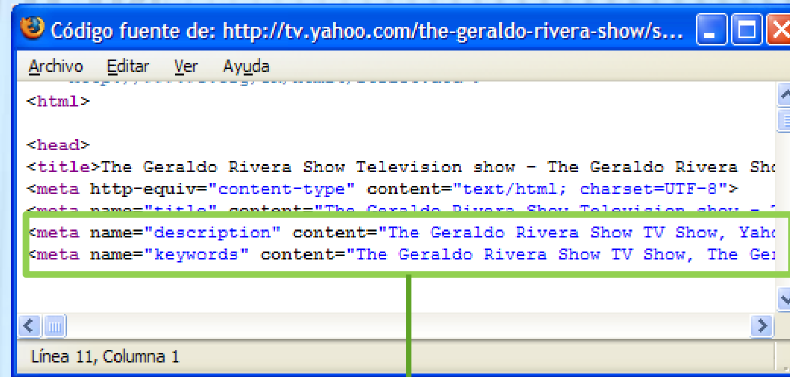
SOURCES OF INFORMATION
SOURCE PAGE

- Why these sources of information?
  - We need small pieces of text because of the computational cost

Anchor Text	• Relevant and summarized information
Surrounding Anchor Text	• Sometimes anchor provide little or no descriptive value ("click here") • 7 terms per side (left and right) • Taking into account HTML block-level elements and punctuation.
URL Terms	• Relevant terms to match against queries in search engines • Extract terminology (top 60%) with KLD and ODP

SOURCES OF INFORMATION





Source of Information	Text
Title	The Geraldo Rivera Show Television show - The Geraldo Rivera Show TV Show - Yahoo! TV
Page Content	*TV Home* > *Shows* > *The Geraldo Rivera Show* Episodes Cast Videos Photos Message Boards Reviews The Geraldo Rivera Show A daily talk show hosted by Geraldo Rivera and featuring on-the-street investigations and studio talk with guests and audience. [...]
Meta Tags	The Geraldo Rivera Show TV Show, Yahoo! TV is your reference guide to The Geraldo Rivera Show Show. Episode guide, photos, videos, cast and crew information, forums, reviews and more The Geraldo Rivera Show TV Show, The Geraldo Rivera Show Television Show, The Geraldo Rivera Show Episode guide, The Geraldo Rivera Show photos, The Geraldo Rivera Show videos, The Geraldo Rivera Show cast, The Geraldo Rivera Show crew , The Geraldo Rivera Show [...]

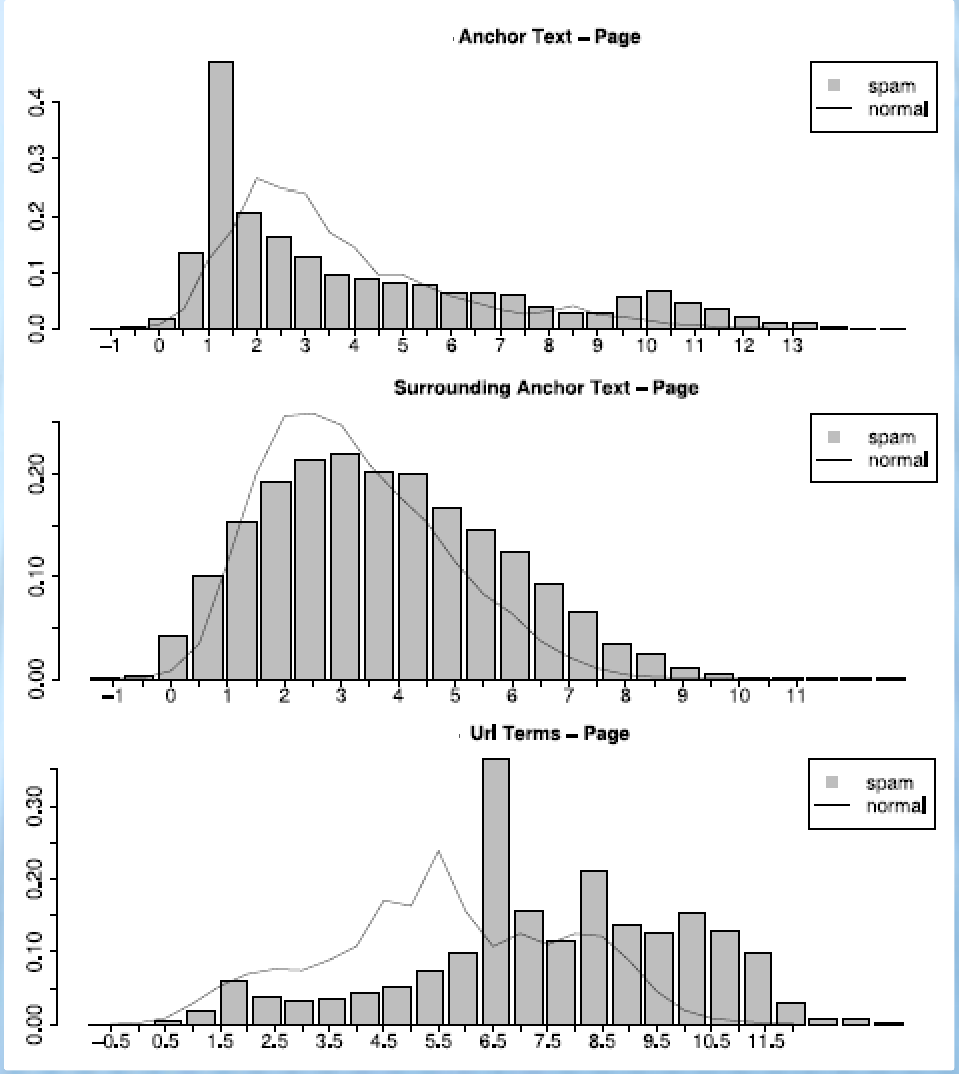
SOURCES OF INFORMATION
TARGET PAGE

- Why these sources of information?
  - We need a descriptive language model from target page

Title	• Relevant and summarized information
Meta Tags	• Provide structured metadata about a Web page • Attributes "description" and "keywords" • Not always available (30-40%)
Page Content	• Always available • At least a minimum amount of text

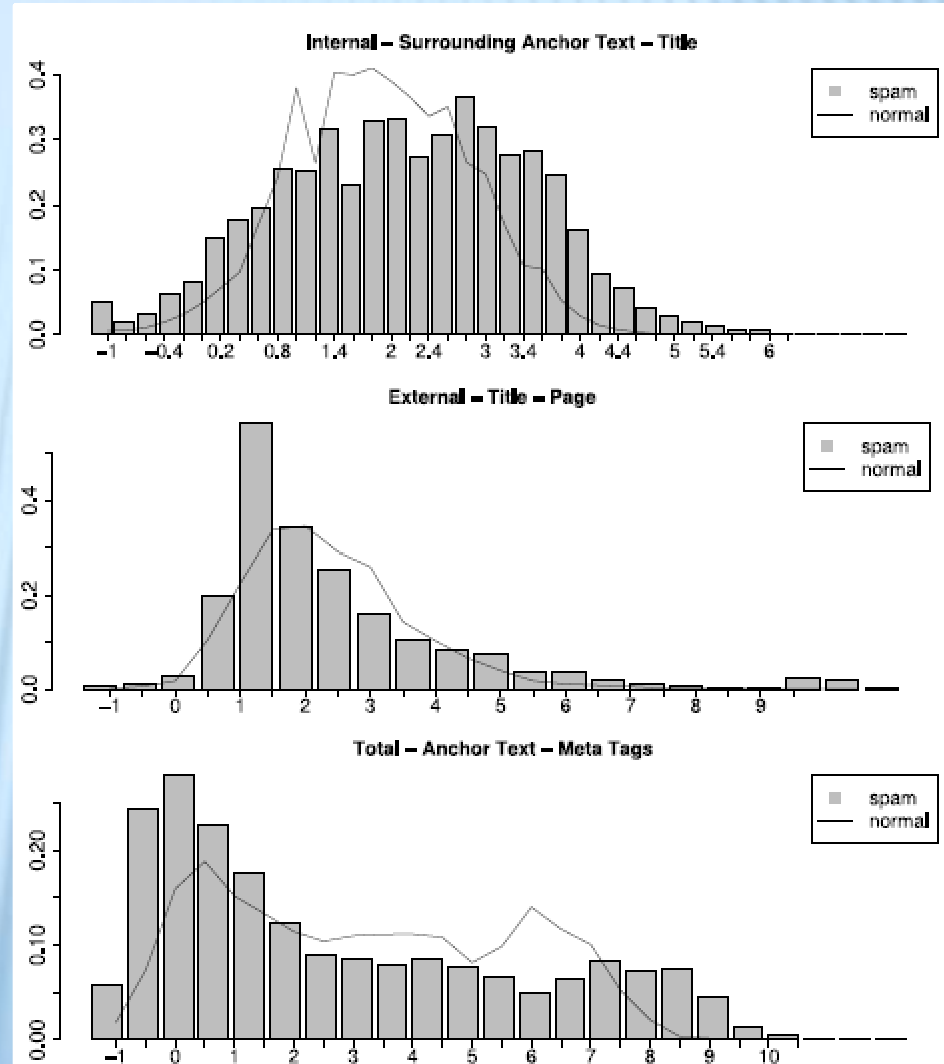
SOURCES OF INFORMATION
SOURCE PAGE

- Anchor Text
- Surrounding Anchor Text
- Url Terms



SOURCES OF INFORMATION
TARGET PAGE

- Title
- Page Content
- Meta Tags



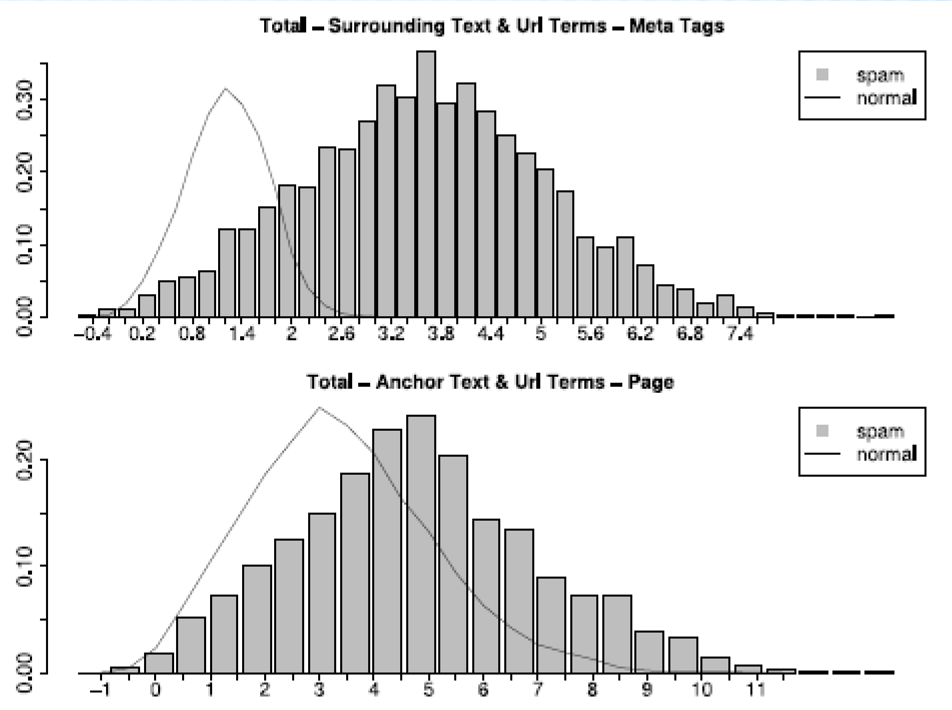
SOURCES OF INFORMATION

- Language models with more information
  - Combining some sources of information - richer language models
  - Take into account the computational cost and the limited relationship between elements
  - Sometimes a small amount of terms in Anchor or URL
  - Combination of sources of information (AU & SU)
  - 14 features

Combination of different Sources of Information	
Content Page (P)	
Anchor Text ($A \rightarrow P$)	Surrounding Anchor Text ($S \rightarrow P$)
Url Terms ($U \rightarrow P$)	Anchor Text ∪ Url Terms ($AU \rightarrow P$)
Surrounding Anchor Text ∪ Url Terms ($SU \rightarrow P$)	
Title (T)	
Anchor Text ($A \rightarrow T$)	Surrounding Anchor Text ($S \rightarrow T$)
Url Terms ($U \rightarrow T$)	Title vs Page ($T \rightarrow P$)
Surrounding Anchor Text ∪ Url Terms ($SU \rightarrow T$)	
Meta Tags (M)	
Anchor Text ($A \rightarrow M$)	Surrounding Anchor Text ($S \rightarrow M$)
Surrounding Anchor Text ∪ Url Terms ($SU \rightarrow M$)	Meta Tags vs Page ($M \rightarrow P$)

SOURCES OF INFORMATION

- Combination of sources of information obtains the best divergence between spam and non-spam pages
  - Surrounding Anchor Text + URL Terms vs Meta Tags
  - Anchor Text + URL Terms vs Page Content

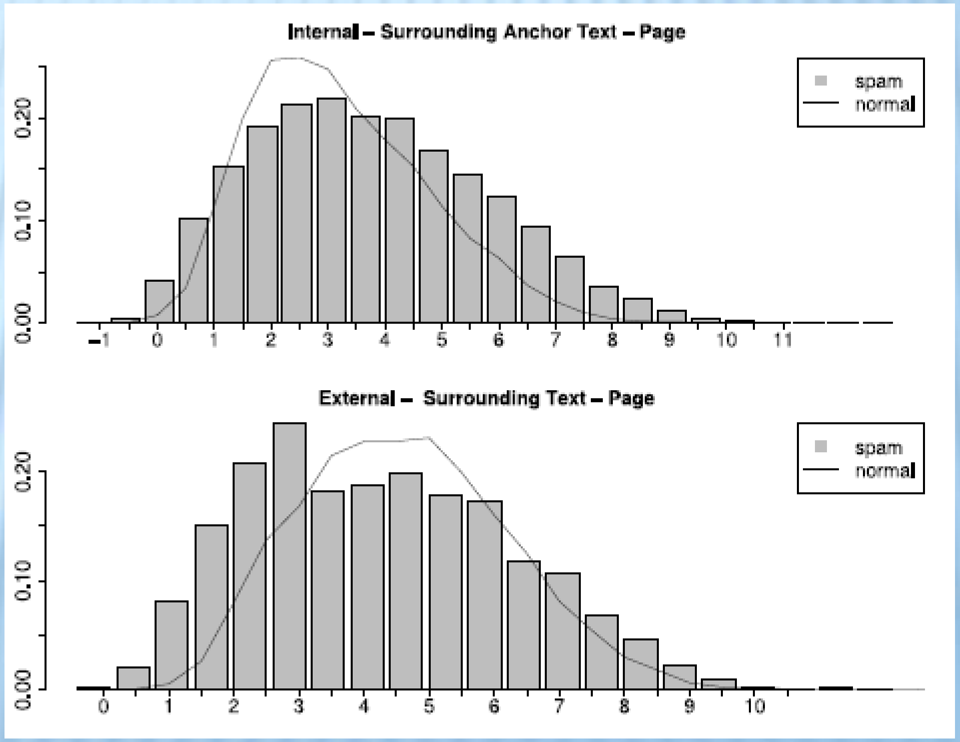


SOURCES OF INFORMATION

- External and Internal Links
  - Articles in SEO Websites and Blogs
  - Ratio between number of such links
  - Triple features (14 x 3)

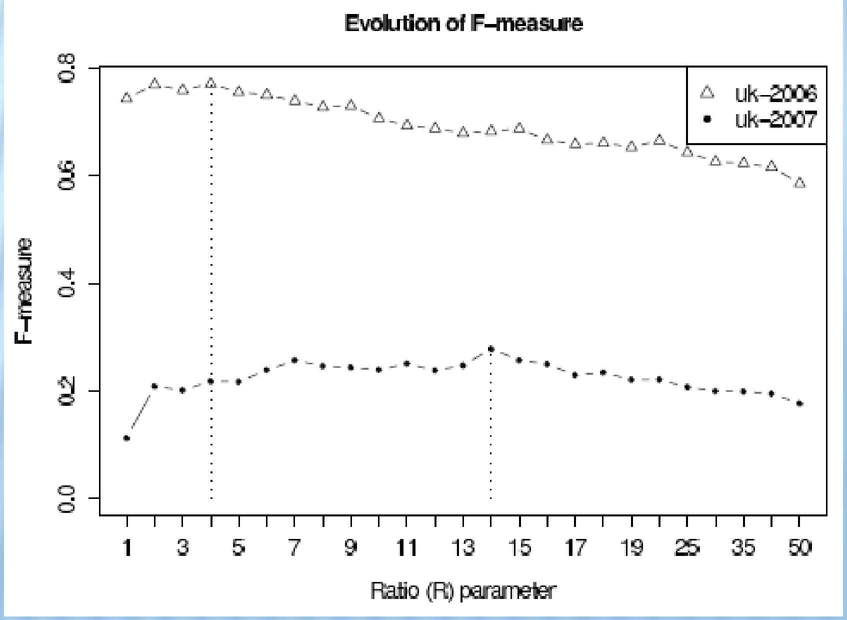
Internal Links

External Links



DATASET AND CLASSIFICATION

- Datasets
  - WEBSPAM-UK2006 and WEBSPAM-UK2007
- Classification
  - Weka
  - Algorithm based on a cost-sensitive decision tree with bagging
    - Misclassify spam pages as normal R times higher



EVALUATION

- Set of features
  - Pre-Computed features from Datasets
    - Content based features (98)
    - Transformed link based features (139)
  - Language model based features (42)
- Performance measures
  - True Positive or Recall (TP)
  - False Positive (FP)
  - F-Measure (combines Precision and Recall)
  - Focus on the F-measure
- Ten-fold cross validation

RESULTS

- WEBSPAM-UK2006

WEBSPAM-UK2006					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Models(LM)	42	0.43	0.05	0.55	0.76
$C \cup L$	237	0.84	0.14	0.75	0.85
$C \cup LM$	140	0.58	0.09	0.61	0.81
$L \cup LM$	181	0.84	0.20	0.69	0.83
$C \cup L \cup LM$	279	0.87	0.11	0.81	0.86

RESULTS

- WEBSPAM-UK2006

Pre-computed

WEBSPAM-UK2006					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Models(LM)	42	0.43	0.05	0.55	0.76
C ∪ L	237	0.84	0.14	0.75	0.85
C ∪ LM	140	0.58	0.09	0.61	0.81
L ∪ LM	181	0.84	0.20	0.69	0.83
C ∪ L ∪ LM	279	0.87	0.11	0.81	0.86

RESULTS

- WEBSPAM-UK2006

LM

WEBSPAM-UK2006					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Models(LM)	42	0.43	0.05	0.55	0.76
C ∪ L	237	0.84	0.14	0.75	0.85
C ∪ LM	140	0.58	0.09	0.61	0.81
L ∪ LM	181	0.84	0.20	0.69	0.83
C ∪ L ∪ LM	279	0.87	0.11	0.81	0.86

RESULTS

- WEBSPAM-UK2006

LM < C
LM < L

WEBSPAM-UK2006					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Models(LM)	42	0.43	0.05	0.55	0.76
$C \cup L$	237	0.84	0.14	0.75	0.85
$C \cup LM$	140	0.58	0.09	0.61	0.81
$L \cup LM$	181	0.84	0.20	0.69	0.83
$C \cup L \cup LM$	279	0.87	0.11	0.81	0.86

RESULTS

- WEBSPAM-UK2006

WEBSPAM-UK2006					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Models(LM)	42	0.43	0.05	0.55	0.76
C ∪ L	237	0.84	0.14	0.75	0.85
C ∪ LM	140	0.58	0.09	0.61	0.81
L ∪ LM	181	0.84	0.20	0.69	0.83
C ∪ L ∪ LM	279	0.87	0.11	0.81	0.86

C + LM < C

• Focus on content spam
•Disagreement in spam cases

RESULTS

- WEBSPAM-UK2006

WEBSPAM-UK2006					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Models(LM)	42	0.43	0.05	0.55	0.76
C ∪ L	237	0.84	0.14	0.75	0.85
C ∪ LM	140	0.58	0.09	0.61	0.81
L ∪ LM	181	0.84	0.20	0.69	0.83
C ∪ L ∪ LM	279	0.87	0.11	0.81	0.86

L + LM > L

- Focus on link and content spam
- Complementary features

RESULTS

- WEBSPAM-UK2006

Baseline

WEBSPAM-UK2006					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Models(LM)	42	0.43	0.05	0.55	0.76
$C \cup L$	237	0.84	0.14	$\boxed{0.75}$	0.85
$C \cup LM$	140	0.58	0.09	0.61	0.81
$L \cup LM$	181	0.84	0.20	0.69	0.83
$C \cup L \cup LM$	279	0.87	0.11	0.81	0.86

RESULTS

- WEBSPAM-UK2006

WEBSPAM-UK2006					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Models(LM)	42	0.43	0.05	0.55	0.76
$C \cup L$	237	0.84	0.14	0.75	0.85
$C \cup LM$	140	0.58	0.09	0.61	0.81
$L \cup LM$	181	0.84	0.20	0.69	0.83
$C \cup L \cup LM$	279	0.87	0.11	0.81	0.86

C + LM < bas.
L + LM < bas.

- Although with fewer features than Baseline

RESULTS

- WEBSPAM-UK2006

WEBSPAM-UK2006					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Models(LM)	42	0.43	0.05	0.55	0.76
$C \cup L$	237	0.84	0.14	$\boxed{0.75}$	0.85
$C \cup LM$	140	0.58	0.09	0.61	0.81
$L \cup LM$	181	0.84	0.20	0.69	0.83
$C \cup L \cup LM$	279	0.87	0.11	$\boxed{0.81}$	0.86

C + L + LM
>
Baseline

- 6% of improvement in F-measure

RESULTS

- WEBSPAM-UK2007

WEBSPAM-UK2007					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.33	0.04	0.30	0.72
Link (L)	139	0.39	0.12	0.20	0.68
Lang. Models(LM)	42	0.24	0.04	0.24	0.72
$C \cup L$	237	0.31	0.03	0.31	0.73
$C \cup LM$	140	0.37	0.05	0.30	0.72
$L \cup LM$	181	0.42	0.12	0.22	0.70
$C \cup L \cup LM$	279	0.33	0.03	0.33	0.75

RESULTS

- WEBSPAM-UK2007

WEBSPAM-UK2007					
Feature Set	Features	TP	FP	F	AUC
Content (C)	98	0.33	0.04	0.30	0.72
Link (L)	139	0.39	0.12	0.20	0.68
Lang. Models(LM)	42	0.24	0.04	0.24	0.72
C ∪ L	237	0.31	0.03	0.31	0.73
C ∪ LM	140	0.37	0.05	0.30	0.72
L ∪ LM	181	0.42	0.12	0.22	0.70
C ∪ L ∪ LM	279	0.33	0.03	0.33	0.75

C + L + LM
>
Baseline

  - Similar results
  - Only a improvement of 2% in F-measure
  - Dataset has a lower ratio of spam pages to learn and classify

CONTRIBUTIONS

- Use of new different sources of information such as URL, Title, Meta Tags
- Combination of sources of information to build richer language models
- Analysis of different features for external and internal links
- Application of language model based features in a public dataset of Web Spam

CONCLUSIONS

New methodology that takes advance of statistical models and NLP

Kullback–Leibler divergence is an efficient measure to detect disagreement between two Web pages

Language model based features improve a 6% in UK2006 dataset, and 2% in UK2007 dataset.

FUTURE WORKS

Analyze the relationship between a page and those that point to it

Extract topics with LDA or LSI to build new language models

Combine language model features with linguistic or new link features

Analyze n-gram models

Thanks!