Web Spam Identification Through Language Model Analysis

Juan Martinez-Romo
Dpto. Lenguajes y Sistemas Informáticos
UNED
28040 Madrid, Spain
juaner@lsi.uned.es

Lourdes Araujo
Dpto. Lenguajes y Sistemas Informáticos
UNED
28040 Madrid, Spain
lurdes@lsi.uned.es

ABSTRACT

This paper applies a language model approach to different sources of information extracted from a Web page, in order to provide high quality indicators in the detection of Web Spam. Two pages linked by a hyperlink should be topically related, even though this were a weak contextual relation. For this reason we have analysed different sources of information of a Web page that belongs to the context of a link and we have applied Kullback-Leibler divergence on them for characterising the relationship between two linked pages. Moreover, we combine some of these sources of information in order to obtain richer language models. Given the different nature of internal and external links, in our study we also distinguished these types of links getting a significant improvement in classification tasks. The result is a system that improves the detection of Web Spam on two large and public datasets such as `WEBSPAM-UK2006` and `WEBSPAM-UK2007`.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Computing Methodologies**]: Natural Language Processing; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Measurement

Keywords

Web Spam, Content Spam, Language Model approach

1. INTRODUCTION

Nowadays, Web Spam is one of the main problems of the search engines because the quality of their search results has been degraded by the methods used by spammers.

AIRWeb '09, April 21, 2009 Madrid, Spain.

During recent years there have been many advances in detection of these fraudulent pages but, in response, new spam techniques have appeared. It might be said that research in this area is fighting against an adversary who constantly uses more and more sophisticated methods. For this reason, it is necessary to improve anti-spam techniques to get over these attacks.

In this paper we propose several new features based on language models to improve Web Spam detection. Language models[16] are probabilistic methods that have been previously used successfully in areas of speech recognition, machine translation, part-of-speech tagging, parsing and information retrieval and even in some previous works for the detection of Splogs[13] or Nepotistic links[3]. While probabilistic models have been proposed and studied for information retrieval since as early as 1960's, they hadn't really shown clear advantages over the traditional vector space model until around 1998, when Ponte and Croft[16] published a pioneering work which uses a different probabilistic model for retrieval, i.e., the query likelihood scoring method. Statistical language models have been developed to capture linguistic features hidden in texts, such as the probability of words or word sequences in a language. A statistical language model (SLM) is a probability distribution P(s) over strings S that attempts to reflect how frequently a string S occurs as a sentence.

Previous works have proved that language model disagreement techniques are very efficient in tasks such as blocking blog spam[13] or detecting nepotistic links[3]. For this reason, we want to apply these techniques to improve classification in a Web Spam labeled collection[5] widely used. We use an extension of the basic language modeling approach to analyse several sources of information extracted from each web site in the collection. We make a language model from each source of information, and then ask how different these two language models are from each other. These sources of information used are: (i) anchor text, surrounding anchor text and Url terms from the source page, and (ii) title and content from the target page. We apply Kullback-Leibler (KL) divergence between their respective language models. KL divergence is an asymmetric divergence measure originating in information theory, which measures how bad the probability distribution M_q is at modeling M_d.

The remaining of the paper proceeds as follows: section 2 describes the previous works in the Web Spam research area; section 3 enumerates the dataset features and the process of classification; section 4 studies the suitability of different sources of information to provide features based on

divergence measures; section 5 is devoted to describe the methodology we have followed to compute the divergence and classify the datasets; section 6 presents the experiments proposed, as well as the results of applying it to different datasets; Finally, section 7 draws the main conclusions.

2. PREVIOUS WORK

Previous work on Web spam detection has focused mostly on the detection of three types of Web spam: link spam, content spam, and cloaking.

Link spam according to Davison[8] can be defined as "links between pages that are present for reasons other than merit." Link spam consists of the creation of a link structure to take advantage of link-based ranking algorithms, such as PageRank, which gives a higher ranking to a website the more other highly ranked websites link to it. Some of the highlights of this area are: Becchetti et al.[2] used automatic classifiers to detect link-based Spam, Gyöngyi et al.[12] separated useful webpages from spam with TrustRank, Zhou et al.[19] with transductive link spam detection and Benczúr et al.[4] analysed supporting sets and PageRank contributions for building an algorithm to detect link spam.

Content spam includes all techniques involve altering the logical view that a search engine has over the page contents[11], for instance, by inserting keywords that are more related to popular query terms than to the actual content of the pages. Fetterly et al.[9] used simple frequency-based measures for its detection. Ntoulas et al.[14] introduced new features based on checksums and word weighting techniques. Piskorski et al.[15] explored linguistic features to then select the more suitable ones.

Cloaking[11] is a technique in which the content presented to the search engine spider is different to that presented to the browser of the user.

There exist also studies that have combined the detection of different types of spam; Abernethy et al.[1] trained a SVM classifier with content and link data. Castillo et al.[6] combined content and topology information in a cost-sensitive tree. Closest to our research are the works by Mishne et al.[13] that apply language models to Blog spam detection. Here, the authors compare the language models from the original post and each of the comments. Benczúr et al.[3] proposed to detect nepotistic links using language models. In this method, a link is down-weighted if the language models from its source and target page have a great disagreement. We share with this approach, the assumption that pages that are connected by non-nepotistic links must be sufficiently similar. Qi et al.[17] distinguished between qualified links and advertising or spam using six similarity measures. To calculate these measures used methods such as Cosine, Dice or Naive Bayes over the Url terms, anchor texts or content.

3. DATASET AND CLASSIFICATION

We use two publicly available Web Spam collections[5] based on crawls of the .uk Web domain done in May 2006 and May 2007 respectively. `WEBSPAM-UK2006` include 77.9 million pages and over 3 billion links about 11,400 hosts. `WEBSPAM-UK2007` include 105.9 million pages and over 3.7 billion links about 114,529 hosts. These reference collections are tagged by a group of volunteers labeling hosts as "normal", "spam" or "borderline".

In our experiments, we restricted the datasets using only hosts labeled at least by two persons independently, and for which all assessors agreed. Moreover Open Directory Project (*ODP*) labels [5] are not taken into account. We made this decision because in some cases mislabeling may mislead classifier, as we can see in the following example (Figure 1) taken from the labels file in the `UK2006-WEBSPAM` collection.

```
4road.co.uk spam 0.66667 j20:S,j7:S,odp:N
www.4road.co.uk normal 0.00000 odp:N
```

Figure 1: Example of mislabelling. A site is labeled as spam and normal in the collection.

Thus, the subset of `WEBSPAM-UK2006` used in our experiments has got 3394 hosts, 1993 of these are labeled as "normal" and 1401 as "spam". Moreover, the subset of `WEBSPAM-UK2007` has a size of 4775, 4593 of these are labeled as "normal" and 182 as "spam".

3.1 Classification

For the classification tasks, we have used the Weka[18] software because it contains a whole collection of machine learning algorithms for data mining tasks. In particular we have chosen a classification algorithm based on a cost-sensitive decision tree with bagging[18] because we have carried out several experiments, and this algorithm works better than the other methods used. As baseline for our experiments we selected the pre-computed content and link features in a combined way to detect different types of Web Spam pages. These features were previously presented in [2, 14].

We have adopted a set of well-known[6] performance measures in Web Spam research: true positive (TP or recall), false positive rate (FP) and *F-measure*. *F-measure* combines precision P and recall R by $F = 2\frac{PR}{P+R}$. For evaluating the classification algorithms, we focus on the *F-measure* as it is a standard measure of summarising both precision P and recall R. The evaluation of the learning schemes used in all the predictions of this paper was performed by a ten-fold cross-validation. For each evaluation, the dataset is split into 10 equal partitions and is trained 10 times. Every time the classifier train with 9 out of the 10 partitions and use the tenth partition as test data.

4. LANGUAGE MODELS AND FEATURES

One of the most successful methods based on term distribution analysis uses the concept of Kullback-Liebler Divergence [7] (KLD) to compute the divergence between the probability distributions of terms of two particular documents considered. We have applied KLD to measure the divergence between two text units of the source and target pages. In Figure 2 there are shown two examples of KLD applied to the anchor text of a link and the title of the page pointed by this link.

$$KLD(T_1||T_2) = \sum_{t \in T_1} P_{T_1}(t) log \frac{P_{T_1}(t)}{P_{T_2}(t)} \quad (1)$$

where $P_{T_1}(t)$ is the probability of the term t in the first text unit, and $P_{T_2}(t)$ is the probability of the term t in the second text unit.

KLD(*Free Ringtones || Free Ringtones for Your Mobile Phone from PremieRingtones.com*) = 0.25

KLD(*Best UK Reviews || Findabmw.co.uk - BMW Information Resource*) = 3.89

Figure 2: Divergence computed applying KLD between the anchor text of a link and the title of the page pointed by this link. Examples extracted from WEBSPAM-UK2006.

The language models that we use estimate maximum likelihood of the unigram occurrence probabilities. In preliminary experiments, we used Jelinkek-Mercer smoothing as Mishne et al.[13], which interpolates the language model of two sources of information.

We test this smoothing approach because it is well-known that it works better with long queries than the divergence without smoothing. To study the impact of smoothing on the results we used two different collections of web pages indexed with Lucene as reference collections:

- Enwiki. This collection contains articles, templates, image descriptions, and primary meta-pages from the English Wikipedia dumps. The size of this collection is around 3.6 million of documents.
- Dmoz. This collection is the result of a crawling process on a random set of Urls from the DMOZ Open Directory Project (ODP)[1]. The whole set is around 4.5 million of sites, but we set the crawling depth to zero, so just a document has been retrieved from each site (homepage site).

Results showed that smoothing improved the results although the difference was quite small. In addition the computation time increased substantially. For these two reasons, we decided not to use a smoothing approach for language models in this work.

4.1 Features

In this paper, we try to characterise the relationship between two linked Web pages according to different values of divergence. These values are obtained by calculating the KL divergence between one or more sources of information from each page. In particular we consider 3 sources of information from the source page: (i) Anchor Text, (ii) Surrounding Anchor Text and (iii) Url terms. We also get 3 sources of information from the target page: (i) Title, (ii) Content Page and (iii) Meta Tags.

Many combinations of these sources of information could be used to measure the divergence between two Web pages. However, considering the issue of computational complexity, he have chosen a set of features that are easy to compute and that are useful in the Web Spam detection. Moreover, we have used Lucene[10] to carry out the calculus, which is a source information retrieval library. These features are described below.

- **Anchor Text - Content.** When a page links to another, this page has only a way to convince an user in order to visit this link, that is by showing relevant and summarised information of the target page. This is the function of the anchor text. Therefore, a great divergence between this piece of text and the linked page shows a clear evidence of Spam. In addition, Mishne[13] and Benczúr[3] proved that disagreement between anchor text and the target content is a very useful measure to detect Spam. In Figure 3 it is shown the distribution of KL divergence between these sources of information and, as in previous studies of this distribution, the *normal* curve is more compact than *spam* one. Anchor text alone is not a very discriminating measure, but it works better when it is combined with surrounding text and URLs terms.

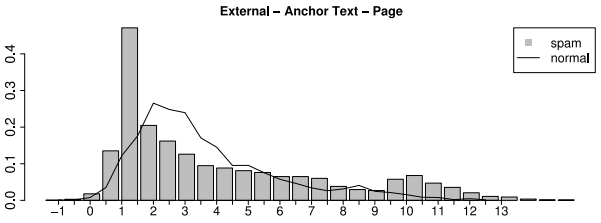

Figure 3: Histogram of KL divergence between Anchor Text and target Page Content. Reference collection is (WEBSPAM-UK2006) and external links have been used.

- **Surrounding Anchor Text - Content.** Sometimes the anchor terms provide little or no descriptive value. Let us imagine a link whose anchor text is "click here". For this reason, text surrounding the anchor can provide a information context about the linked page. Moreover, in [3] a better behaviour is observed when the anchor text is extended with neighboring words. In our experiments, we use several words around the anchor text (7 per side) to extend it. The result is a source of information much richer and very useful to detect *Spam*. Figure 4 shows how the *spam* curve is displaced towards higher values of divergence and *normal* values are concentrated near $KL \approx 2.5$.

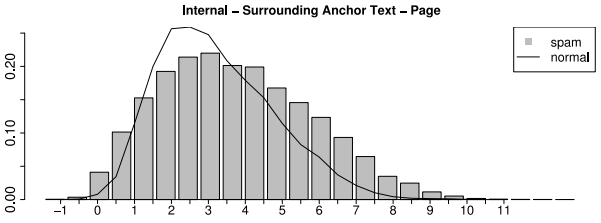

Figure 4: Histogram of KL divergence between Surrounding Anchor Text and target Page Content. Reference collection is (WEBSPAM-UK2006) and internal links have been used.

- **URL Terms - Content.** Besides the anchor text, the only information available of a link is its Url. An Url is mainly composed by a protocol, a domain, a path and a file. In the same way, these elements are composed by terms that can provide rich information of the target page. Moreover, because of the increasing use of search engines in last years, there exist Search Engine Optimisation (SEO) techniques that try to exploit the importance of Url terms in a request.

[1] www.dmoz.org

Thus, if we have an Url as "www.domain.com/viagra-youtube-free-download-poker-online.html", and after visit this page, it is an online music store, it could be said that this page uses Spam techniques. Therefore, we have retrieved the most relevant terms from an Url in order to calculate the divergence with the content of the target page. To extract the most relevant these terms, first of all, we have build a language model with terms from Urls in ODP public list. Afterwards, with help of this collection of Urls we have used the KL divergence in order to know the most relevant terms in a certain Url. Finally, we use 60% of these terms. This measure is illustrated in Figure 5, which shows the great difference between *spam* and *normal* histograms.

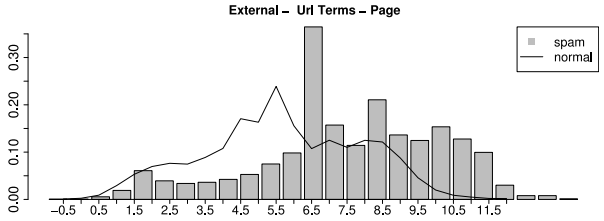

Figure 5: Histogram of KL divergence between Url Terms and target Page Content. Reference collection is (WEBSPAM-UK2007) and external links have been used.

- **Anchor Text - Title.** There exist papers which show the similarity between the anchor text and the title of a Web page. In both cases, they are small pieces of text which summarise the content of a page, but an anchor text is written by a foreign person to the page, whereas title is set by the owner. Using language models, there could be problems because of the small number of terms in two text units, but in our experiments we proved that in many cases, this measure provides relevant information. In Figure 6 can be observed that anchor text alone is not a very discriminating measure, but it works better when it is combined with surrounding text and URLs terms.

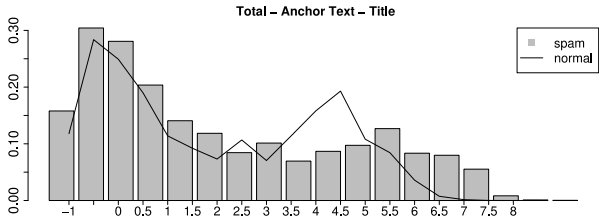

Figure 6: Histogram of KL divergence between Anchor Text and target Page Title. Reference collection is (WEBSPAM-UK2007) and both internal and external links have been used.

- **Surrounding Anchor Text - Title.** In the same way that previous measure (surrounding anchor text vs page content), we extend the anchor text with some surrounding words in order to get a better context on the link. Moreover, as in this case we have a higher number of terms in the first language model, the problems that may arise in the previous measure would be eliminated. As we can check in Figure 7, histogram is more discriminant than Figure 6. In this case, most *spam* values are located between $KL \approx 1$ and $KL \approx 3$.

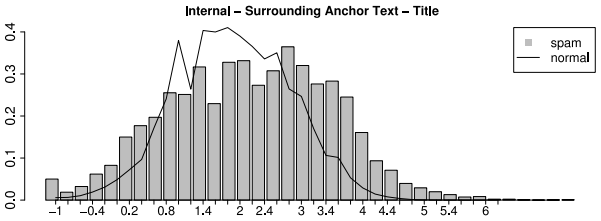

Figure 7: Histogram of KL divergence between Surrounding Anchor Text and target Page Title. Reference collection is (WEBSPAM-UK2006) and internal links have been used.

- **URL Terms - Title.** If in two previous measures (anchor text and surrounding anchor text) the first source of information from the target page was generated by a foreign person, Url or at least a part of it (file and/or path), is built by the page author. Thus, it should have a coherence between the URL terms and the title of a page. Otherwise, as it is illustrated in Figure 8 we could be analysing a Spam page. As Figure 6, *spam* curve is much more compact than *normal* one.

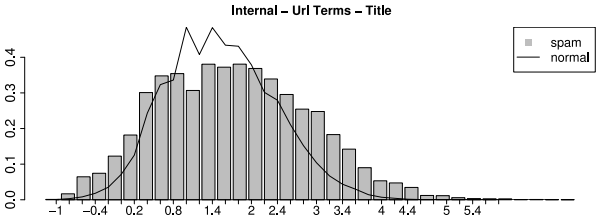

Figure 8: Histogram of KL divergence between Url Terms and target Page Title. Reference collection is (WEBSPAM-UK2006) and internal links have been used.

- **Title - Content.** It is well-known that both, terms of a URL and terms of the Web page title, have a great impact when search engines decide whether a page is relevant to a query. In other words, *spammers* perform engineering tasks in order to set key terms in these sources of information. This measure attempts to detect those cases of *Spam*, which there is not a relationship between the title and the page content at the same site. In Figure 9 it is shown the distribution of KL divergence between these sources of information.

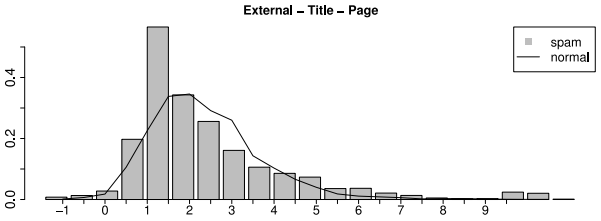

Figure 9: Histogram of KL divergence between both Title and Content from target Page. Reference collection is (WEBSPAM-UK2006) and external links have been used.

- **Meta Tags.** Meta Tags provide structured metadata about a Web page and they are used in search engine optimisation. Although they are the target of *spammers* for a long time and search engines consider these data less and less, there are pages still using them because of their clear usefulness. In particular we have considered metatags with attributes "description" and "keywords" to build a virtual document with their terms. We have decided to use these data to calculate its divergence from other sources of information from the source page such as anchor text and surrounding anchor text, and from the target page such as page content and URL terms. This measure is illustrated in Figure 10, which shows a different distribution in *spam* and *normal* values.

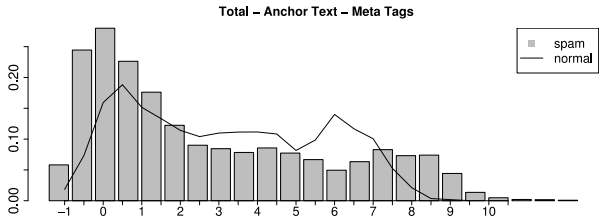

Figure 10: Histogram of KL divergence between Anchor Text and target Page Meta Tags. Reference collection is (WEBSPAM-UK2006) and both internal and external links have been used.

Combination of sources of information

In addition to these features, we have combined several sources of information from the source page with the goal in mind of creating virtual documents which provide richer information. As we have seen above, we have used Anchor Text (A), Surrounding Anchor Text (S) and Url terms (U) as sources of information. We also propose to create two new sources of information: (i) combining Anchor Text and Url terms (AU) and (ii) combining Surrounding Anchor Text and Url terms (SU). Moreover, we have these sources of information from the target page: Content Page (P), Title (T) and Meta Tags (M). We have also ruled out the use of any combination due to the limited relationship between these sources of information. Table 1 below summarise all 14 features used in this work.

We have obtained language models much richer and descriptive as result of this combination of sources of information. In many cases we can find anchors with a small amount of text that sometimes mislead our results. However, by combining different sources of information such as Anchor text, Surrounding Anchor text and Url terms we obtain a more descriptive language. Furthermore, although single measures described in section 4.1 offer histograms with interesting divergence values between *Spam* and *Not Spam*, the best measures proposed in this work are those that combine different sources of information. As it can be seen in Figure 10, these terms combination gets a divergence very efficient between *spam* pages and those that are not *spam*. Since both distributions have Gaussian shapes, *normal* histograms are more compact and theirs means are near $KL \approx 1.2$ and $KL \approx 3.5$ respectively. On the other hand *spam* histograms are wider, and theirs means are near $KL \approx 4$ and $KL \approx 5$ respectively.

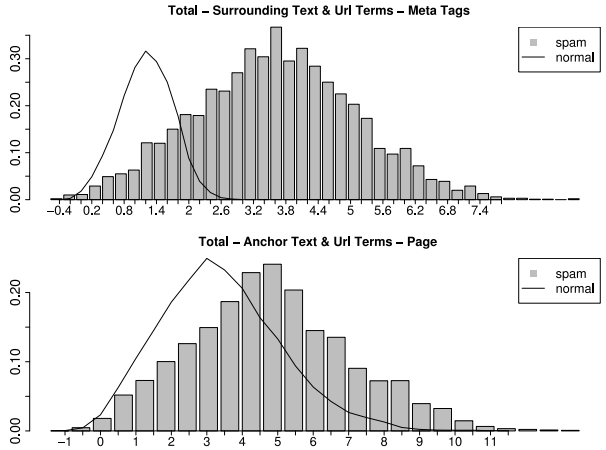

Figure 11: (Above) Histogram of KL divergence between a combination of Anchor Text, Surrounding Anchor Text and Url Terms and target Page Title and (Below) Histogram of KL divergence between a combination of Anchor Text and Url Terms and target Page Content. Reference collection is (WEBSPAM-UK2007) and both internal and external links have been used.

Internal and External Links

Web sites contain in most cases two types of links:(i) links to the same site (internal links) and (ii) links to foreign sites (external links). Internal links can provide depth and context about certain information. External links fulfill a similar function and are particularly useful to the reader by helping to find other sources cited in the informative content. Web site owners are often afraid of external links, perhaps for a commercial reason: to send traffic to other sites.

In addition, there are SEO techniques that consider the relationship between internal and external links in order to obtain the highest Pagerank, that is, a ratio between the number of such links. There exist also rumours about the impact of internal and external links in a page, in face of the ranking provided by a search engine. This suggests that *Spammers* may be using algorithms that take into account this information to promote their pages.

For these reasons it is also distinguished between internal links and external links in order to carry out the divergence analysis. Therefore, for each Web page we have triple-features: 14 features for internal links, 14 features for external links and 14 features for internal and external links. In Figure 11 it can be observed the difference between histograms of internal and external links. Moreover, this figure shows how the *Spam* curve is displaced towards higher values of divergence in internal links case, whereas the same curve is displaced towards lower values in other case. It can be also noticed that *Spam* distribution mean is higher ($KL \approx 3$) in external links than in internal links case ($KL \approx 4.5$). In case of *Normal* distribution, it happens the opposite.

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 $\cup$ Url Terms ($AU \rightarrow P$)
Surrounding Anchor Text $\cup$ 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 $\cup$ Url Terms ($SU \rightarrow T$)	
Meta Tags (M)	
Anchor Text ($A \rightarrow M$)	Surrounding Anchor Text ($S \rightarrow M$)
Surrounding Anchor Text $\cup$ Url Terms ($SU \rightarrow M$)	Meta Tags vs Page ($M \rightarrow P$)

Table 1: Combination of different sources of information used to calculate the KL divergence.

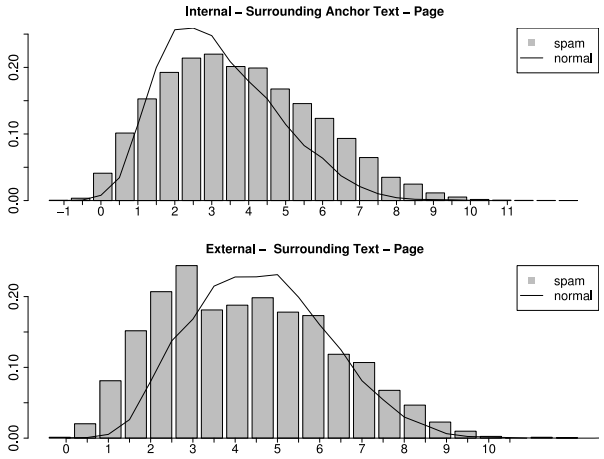

Figure 12: Histograms of KL divergence between Surrounding Anchor Text and target Page Title. Only internal links have been used in histogram above and external links in histogram below. Reference collection is (WEBSPAM-UK2007).

5. METHODOLOGY

In order to carry out the divergence analysis and considering the computational cost, we only analyse one page per host. Although it would be interesting to explore all the pages of a website in depth. Especially, we select the homepage from the source page and every page pointed by any link in the source page. Furthermore, hosts that have no outgoing links are discarded, so the final size of the dataset is reduced slightly.

First of all, links are selected for each analysed host. In particular we analyse links that have some information in the anchor text. Therefore we filter out images, links to the same page (named anchor inside a HTML document), numbers, URLs and empty strings. We also rule out links which protocol is not HTTP or links to non HTML documents.

WEBSPAM-2006 and WEBSPAM-2007 are labelled at host level [5], so we have to aggregate all language models measures at this level. We only analyse one page per host and this page represent this host, but we obtain 42 language models measures for each link in a Web page. Thus, we estimate the average of all links for each measure. A Web site is therefore represented for 42 features (average values).

6. EXPERIMENTS

We now present several experiments in identifying Web Spam on public datasets using our methodology. After filtering the pages that did not meet the requirements, such as pages not having enough text, no outgoing links, etc. WEBSPAM-UK2006 dataset used in our experiments has got 3083 hosts, 1811 of these are labelled as "normal" and 1272 as "spam". Moreover, *normal* hosts have an average of external and internal links of 12.1 and 30.6 respectively and *spam* hosts have an average of external and internal links of 7.2 and 15.3. The WEBSPAM-UK2007 dataset used has got a size of 4166, 4012 of these are labelled as "normal" and 154 as "spam". Otherwise *normal* hosts have an average of external and internal links of 3.7 and 13.4 respectively and *spam* hosts have an average of external and internal links of 9.3 and 12.06.

In order to check that language models features improve the precision of spam detection, we decided to use pre-computed features available for public datasets[2]. In particular we have used the content-based features and the transformed link-based features. In addition, we have combined different feature sets in order to obtain a classifier which was able to detect both content-spam and link-spam cases. Finally we combine content, link and language models features by achieving a more accurate classifier.

For the classification tasks, we have used the Metacost algorithm (cost-sensitive decision tree with bagging) implemented in Weka. We chose this classifier because we think that errors for misclassifying *normal* pages as *spam* do not have the same impact than misclassify a *spam* page as *normal*. Thus, as in Castillo et al[5], we have imposed a zero cost to right predictions, and we have set to *spam* pages misclassified as *normal* a cost R times higher than *normal* pages misclassified as *spam*. Furthermore, as the aim of this work is to maximize the *F-measure*, we have look for the value of R which maximize this measure. In Figure 13 is ilustrated the evolution of *F-measure* obtained by applying different costs to R. According to these results we have set $R = 4$ in WEBSPAM-UK2006 dataset and $R = 14$ in WEBSPAM-UK2007.

6.1 Results

The results of our experiments are shown in Table 2 and Table 3. As it can be checked, if we only use the pre-computed features from datasets, we obtain the best results combining content and link based features. For this reason

[2]http://webspam.lip6.fr

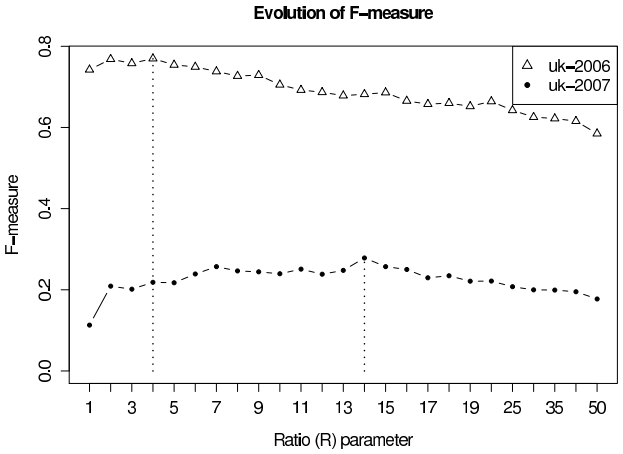

Figure 13: Evolution of *F-measure* obtained by applying different costs to R in the confusion matrix. Content and Links based features are used on `WEBSPAM-UK2006` and `WEBSPAM-UK2007`.

we have chosen the union of these two sets of features as a baseline for our experiments.

Table 2 illustrates that language models features are not as efficient for themselves as content or link features, but we should take into account that they are a smaller number of features than content or link sets of features. Moreover, in Table 3 is shown that language models features works better than link based features. Furthermore, with language models features, we can detect a lot of Spam pages which use techniques based on content, although some instances of link spam pages can be also detected.

Concerning `WEBSPAM-UK2006` dataset and continuing this reasoning, since content-based features worsening F-measure when they are combined with language models features (C ∪ LM), link-based features (L ∪ LM) get an improvement of 3%. Although the most important, if we consider the combination of content and links features (C ∪ L) as baseline, the classifier improve 6% the *F-measure* from 0.75 to 0.81 using language models features (C ∪ L ∪ LM).

On the other hand, as it is shown in Table 3 the detection rate in `WEBSPAM-UK2007` dataset is lower than previous dataset. In this case, the collection is less balanced having 4593 hosts labeled as "normal" and 182 as "spam". For this reason, detection of spam is more difficult now. As a result of the former, the best *F-measure* in our experiments was obtained when we used the combination of content, links and language models features (C ∪ L ∪ LM), by improving 2% the baseline.

7. CONCLUSIONS

Every day, spammers are making progress on new techniques used to mislead search engines. Otherwise, users are gradually more and more demanding and they require more precise results. In addition, they do not want to find spam pages in their hits. In the last years there have been a lot of research works in this area and in this work we have tried to study in depth techniques for detecting content-based Web Spam.

We have presented a methodology that makes the most of the power of statistical models and natural language processing (NLP). Specifically, we have used language models to represent a Web document and then we try to calculate disagreement between two Web pages. To build language models we used different sources of information from each page. In particular we obtained three sources of information from the source page: (i) Anchor Text, (ii) Surrounding Anchor Text and (iii) Url terms. We also got three sources of information from the target page: (i) Title, (ii) Content Page and (iii) Meta Tags. Moreover we combined some of these sources of information in order to obtain richer language models. Thus, for every analysed page and for every one of its links we got 14 measures. These measures are obtained by applying Kullback-Liebler divergence on different language models from source and target pages.

To represent a host, we used only one of its pages and we also calculated the average of each one of these divergence presented measures. So we got 14 language-models-based features for each host. Given the different nature of internal and external links, in our study we also distinguished these types of links, by finally representing each host with 42 features: 14 features for internal links, 14 features for external links and 14 features for internal and external links.

For classification tasks we have used the public `WEBSPAM-UK2006` and `WEBSPAM-UK2007` datasets and we used their precomputed features (linked ant content based features) in a separate way and together with our language-models-based features in order to evaluate the efficiency of our proposed features. As baseline we have used the combination of linked-and-content-based features because they are the features set which we have obtained the best results. Still, by adding the

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

Table 2: Features, True Positive rate (TP), False Positive rate (TP), F-measure (F) and Area Under Roc Curve (AUC) for Web Spam classifiers using different feature sets on UK-2006.

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

Table 3: Features, True Positive rate (TP), False Positive rate (TP), F-measure (F) and Area Under Roc Curve (AUC) for Web Spam classifiers using different feature sets on UK-2007.

language-models-based features to the previous set (content and link) it is obtained an improvement of the *F-measure* near 6%, in `WEBSPAM-UK2006` collection case, and an improvement of 2% in the `WEBSPAM-UK2007` collection. Therefore, thanks to these experiments we have proved that, the application of language models on a combination of sources of information extracted from two Web pages, improve the detection Web spam tasks.

In future works we would like to analyze the relationship between a page and those that point to it, and to measure disagreement between different data sources. In this work we have not analyzed all pages in a Web site because of its computational cost, but in future we think that it could be interesting to solve this problem. Finally, we would want to find a way to optimize the used language models and then smoothing techniques could be applied.

8. ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Science and Innovation within the project QEAVis-Catiex (TIN2007-67581-C02-01) and the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267).

9. REFERENCES

[1] J. Abernethy, O. Chapelle, and C. Castillo. Webspam identification through content and hyperlinks. In *Proceedings of the fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of web spam. In *AIRWeb'06: Proceedings of the 2th international workshop on Adversarial information retrieval on the web*, 2006.

[3] A. A. Benczúr, I. Bíró, K. Csalogány, and M. Uher. Detecting nepotistic links by language model disagreement. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 939–940, New York, NY, USA, 2006. ACM.

[4] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. Spamrank - fully automatic link spam detection. In *In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[5] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, 2006.

[6] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, New York, NY, USA, 2007. ACM.

[7] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[8] B. Davison. Recognizing nepotistic links on the web, 2000.

[9] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In *WebDB '04: Proceedings of the 7th International Workshop on the Web and Databases*, pages 1–6, New York, NY, USA, 2004. ACM.

[10] O. Gospodnetic and E. Hatcher. *Lucene in Action*. Manning, 2004.

[11] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of the first International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[12] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*, pages 576–587. VLDB Endowment, 2004.

[13] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[14] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 83–92, New York, NY, USA, 2006. ACM.

[15] J. Piskorski, M. Sydow, and D. Weiss. Exploring linguistic features for web spam detection: a preliminary study. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 25–28, New York, NY, USA, 2008. ACM.

[16] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM.

[17] X. Qi, L. Nie, and B. D. Davison. Measuring similarity to detect qualified links. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 49–56, New York, NY, USA, 2007. ACM.

[18] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2 edition, 2005.

[19] D. Zhou, C. J. C. Burges, and T. Tao. Transductive link spam detection. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 21–28, New York, NY, USA, 2007. ACM.