Cross-Lingual Web Spam Classification

András Garzó Bálint Daróczy Tamás Kiss Dávid Siklósi András A. Benczúr

Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI)
Eötvös University, Budapest
{garzo, daroczyb, kisstom, sdavid, benczur}@ilab.sztaki.hu

ABSTRACT

While Web spam training data exists in English, we face an expensive human labeling procedure if we want to filter a Web domain in a different language. In this paper we overview how existing content and link based classification techniques work, how models can be "translated" from English into another language, and how language-dependent and independent methods combine. In particular we show that simple bag-of-words translation works very well and in this procedure we may also rely on mixed language Web hosts, i.e. those that contain an English translation of part of the local language text. Our experiments are conducted on the ClueWeb09 corpus as the training English collection and a large Portuguese crawl of the Portuguese Web Archive. To foster further research, we provide labels and precomputed values of term frequencies, content and link based features for both ClueWeb09 and the Portuguese data.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; I.2 [**Computing Methodologies**]: Artificial Intelligence

General Terms

Document Classification, Information Retrieval, Hyperlink Analysis

Keywords

Cross-lingual text processing, Web classification. Web spam, Content analysis, Link analysis

1. INTRODUCTION

It has already been known from the early results on text classification that "obtaining classification labels is expensive" [32]. This is especially true in multilingual collections where either separate training labels have to be produced for each language in question, or techniques of cross-lingual information retrieval [13] or machine translation [35] have to be used.

While several results focus on cross-lingual classification of general text corpora [2; 38; 43, and many more], we concentrate on the special and characteristically different problem of Web classification. Web spam filtering, the area of devising methods to identify useless Web content with the sole purpose of manipulating search engine results, has drawn much attention in the past years [41, 29, 26]. Our results on cross-lingual Web classification are motivated by the needs and opportunities of Internet archives [4].

Web classification may exploit methods of recent evaluation campaigns on Web spam filtering. Our results combine methods from two areas, cross-lingual information retrieval and Web classification. Traditional methods in cross-lingual information retrieval use dictionaries, machine translation methods, and more recently multilingual Wikipedia editions. Web classification on the other hand relies on features of content and linkage [9], some of which are language independent. However, language independence does not necessarily imply domain independence: PageRank and its variants may have different distributions for differing interconnectivity and the ratio of the "boundary": the pages not included but pointed to by some page in the domain, crawl, or language. TrustRank and query popularity based features depend on the availability of a trusted seed set, typically hosts listed in the Open Directory Project (`http://dmoz.org`), and the coverage of search queries. Finally, the typical word length and text entropy may also vary language by language.

This paper experiments with a new combination of learning methods and cross-lingual features for web classification. Our task is different from standard methods of cross-lingual text classification (see [43] and references therein) in the following aspects:

- We classify hosts not individual pages as this is the standard task for Web spam [7].
- Even if we consider a national domain, the actual language used in a host can be mixed, especially for spam pages automatically generated from chunks (see Fig. 1 as an example).
- We may exploit multilingualism by classifying a host based on its part written in English.

We note that host level classification is preferred for Web spam filtering due to the facts that (1) fine-grained page or even comment level classification is computationally unfeasible on the Web scale; and (2) the goal is to filter mass amounts of spam including link farms and machine generated content that can be blocked on the host level. Indeed, our set of labeled Portuguese spam hosts is the byproduct of the normal quality assessment procedure conducted within the Portuguese Web Archive. In previous results [9; 7, and many more] full host names are used as a domain and we use this definition in this paper, however we argue that a

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

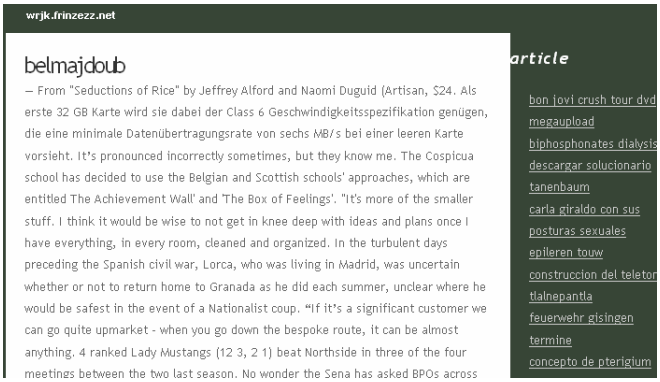


Figure 1: Portion of a mixed language machine generated spam page.

pay level domain or even IP based definition [15] would fit the problem even better. In addition, labeling a page or an entire host is almost the same effort for a human, and very frequently a single page cannot even be assessed without seeing the context, very much unlike email spam or spam in social media.

In this paper we investigate how much various classes of Web content and linkage features, some requiring very high computational effort, add to the classification accuracy. As the bag of words representation turned out to describe Web hosts best for most classification tasks of the ECML/PKDD 2010 Discovery Challenge [15], we realized that new text classification methods are needed for the cross-lingual task.

Based on recent results in Web spam filtering, we also collect and handle a large number of features and test a variety of machine learning techniques, including SVM, ensemble selection, LogitBoost and Random Forest. Our key findings are summarized next.

- Hosts that contain a mix of English and national language content, likely translations, yield a very powerful resource for cross-lingual classification. Some of our methods work even without using dictionaries, not to mention without more complex tools of natural language processing.
- Similar to our previous English-only results, the bag-of-words representation together with appropriate machine learning techniques is the strongest method for Web classification.
- The "public" spam features of Castillo et al. [9], especially the content-based ones, depend heavily on the data collection and have little generalizational power. For spam classification they require cross-corpus normalization while for topical classification, the content based features do not seem to be applicable.

To assess the prediction power of the proposed features, we run experiments over the `.pt` domain [24, 23]. Our techniques are evaluated along several alternatives and yield a considerable improvement in terms of area-under-the-curve (AUC).

The rest of the paper is organized as follows. After a review of related research at the intersection of machine learning, cross-lingual information retrieval and Web mining (Section 2), we introduce the proposed learning methods and describe the classification features (Section 3). Our experimental results are presented in Section 4.

2. RELATED WORK

We base our methods on results of both cross-lingual and Web classification that we review next. In general, cross-lingual classification either works by translating documents [38, 30, 43], or terms only [2], or using an intermediate language-independent representation of concepts [44]. For general results on cross-lingual text classification we refer to [2] who propose linguistic resources such as dictionaries similar to the ones used in cross-lingual information retrieval. As a broad overview, we refer to the CLEF Ad Hoc tasks overview papers, e.g. [13] in the latter area. We also note that several results exploit Wikipedia linkage and local editions [42, 31, 22].

Several cross-language classification results, similar to ours, work over "pseudo-English" documents by translating key terms into English using dictionaries [2], or using latent semantic analysis [14, 36]. The cross-lingual classification results reported are however, unlike ours, much worse than the monolingual baselines.

Semi-supervised learning finds applications in cross-lingual classification where, similar to our methods, the unlabeled part of the data is also used for building the model. Expectation maximization is used in [38, 39] for cleansing the classifier model from translation errors; others [37] exploit document similarities over the unlabeled corpus. In [43] co-training over machine translated Chinese and English text is used for sentiment analysis.

Closest to our goals is the method of [30] for classifying Chinese Web pages using English training data, however, either because of the cultural differences between Chinese and English content or the fact that they classified on the page and not host level, they achieve accuracy metrics much weaker than for the monolingual counterpart. We also note that they are aware of the existence of multilingual content but they apparently do not exploit the full power of multilingual hosts. Finally, a recent Web page classification method described in [44] uses matrix tri-factorization for learning an auxiliary language, an approach that we find computationally unfeasible for classification in the scale of a top level domain.

Text classification is studied extensively in classical information retrieval. While traditional term-based topical classification for Web content relies on local page content only, several solutions tailored to the web use terms from linked pages as well [5, 21]. Semi-supervised learning methods (surveyed, for instance, in [47]) exploit information from both labeled and unlabeled data instances. Relational learning methods (presented, for instance, in [20]) also consider existing relationships between data instances.

Recognizing and preventing spam has been identified as one of the top challenges for web search engines [29, 41]. As all major search engines use page, anchor text, and link analysis algorithms to produce their rankings of search results, web spam appears in sophisticated forms that manipulate page contents as well as the interconnection structure [27]. Accordingly, spam hunters also rely on a variety of content [17, 34, 18] and link [28, 3, 46] based features to detect web spam; a recent evaluation of their combination is provided in [9]. In the area of the so-called Adversarial Information Retrieval, workshop series ran for five years [16],

evaluation campaigns including the Web Spam Challenges [7] were organized. The ECML/PKDD Discovery Challenge 2010 (see e.g. [15]) extended the scope by introducing labels for genre and quality by serving the needs of a fictional archive.

Our baseline classification procedures are collected by analyzing the results of the Web Spam Challenges and the ECML/PKDD Discovery Challenge 2010. Best results either used bag of words vectors or the so-called "public" feature sets of [8]. The Discovery Challenge 2010 best results [25, 1, 33] and our analysis [15, 40] show that the bag of words representation variants proved to be very strong for the English collection. For classification techniques, a wide selection including decision trees, random forest, SVM, class-feature-centroid, boosting, bagging and oversampling in addition to feature selection (Fisher, Wilcoxon, Information Gain) were used. In our previous work [40], we improved over the best results of the Challenge participants by the combination of SVM and biclustering over the bag of words representation of the hosts. These experiments indicate little use of link and content based features. A possible reason is that the DC2010 training and test sets were constructed in a way that no IP and domain was allowed to be split between training and testing. The rationale is that once a domain or IP is found to consist of spam, its subdomains or other hosts on the same server are much more likely spam and their classification becomes straightforward. This simple consideration was not implemented in early datasets: the Web Spam Challenge data sets were labeled by uniform random sampling. For this reason, we have to reconsider the applicability of propagation [46] and graph stacking [9].

3. METHOD

Our Web host classification applies a classifier ensemble consisting of features based on content and linkage as well as various English, translated, and semi-supervised Portuguese bag of words models. The following subsections describe the core ingredients. The standard content and link-based features[1] and the necessary transformations from the English to the Portuguese collection are described in Sections 3.1 and 3.2, respectively. In Section 3.3 we describe our bag-of-words translation method and SVM based classifiers, followed by a semi-supervised algorithm that relies on multilingual host content to first give prediction using a pure English model and then apply the results to train a Portuguese model. Finally the ensemble method ingredients are found in Section 3.5.

3.1 Features: Content

Among the early content spam papers, Fetterly et al. [17] demonstrated that a sizable portion of machine generated spam pages can be identified through statistical analysis. Ntoulas et al. [34] introduce a number of content based spam features including number of words in the page, title, anchor, as well as the fraction of page drawn from popular words, and the fraction of most popular words that appear in the page. Spam hunters use a variety of additional content based features [6, 18] to detect web spam; a recent measurement of their combination appears in [9] who also provide these methods as a *public feature set* for the Web Spam Challenges.

We use the public feature set [9] that includes the following values computed for the home page of the domain, the page with the maximum PageRank, and the average over the entire host:

1. Number of words in the page, title;
2. Average word length, average word trigram likelihood;
3. Compression rate, entropy;
4. Fraction of anchor text, visible text;
5. Corpus and query precision and recall.

Here feature classes 1–4 can be normalized by using the average and standard deviation values over the two collections, while class 4 is likely domain and language independent.

Corpus precision and recall are defined over the k most frequent words in the dataset, excluding stopwords. Corpus precision is the fraction of words in a page that appear in the set of popular terms while corpus recall is the fraction of popular terms that appear in the page. This class of features is language independent but rely on different lists of most frequent terms for the two data sets.

Query precision and recall is based on frequencies from query logs that have to be either compiled separately for each language or domain (questions from Portugal likely have different distribution than from Brazil), or the English query list has to be translated. Since we had no access to a query log in Portuguese, we selected the second approach.

3.2 Features: Linkage

Recently several results have appeared that apply rank propagation to extend initial judgments over a small set of seed pages or sites to the entire web, such as trust [28, 46] or distrust. These ideas were distilled into the public link based feature set [9] and include the following values with averages, standard deviation, and several functions computed from them:

- Assortativity, reciprocity;
- In and out-degree;
- Host and page neighborhood size at various distances;
- PageRank and truncated variants.

One of the strongest features is TrustRank [28], PageRank personalized on known honest hosts. TrustRank however needs a trusted seed set. Typically hosts that appear in the Open Directory Project (ODP) are used as seed. Unfortunately, ODP acts as our negative sample set as well, hence in this paper we have to omit TrustRank, one of the strongest link-based features in our discussion.

3.3 Features: Bag-of-Words

Spam can be classified purely based on the terms used. Based on our recent result, we use libSVM [10] with several kernels and apply late fusion as described in [40]. The bag of words representation of a Web host consists of the top 10,000 most frequent terms after stop word removal.

In order to classify hosts in Portuguese, we translate the Portuguese terms to construct an English bag of words representation of the host. The procedure is described in Algorithm 1 with the following considerations:

- Short terms are not translated as they typically cause noise and often coincide between the languages.

[1] `http://barcelona.research.yahoo.net/webspam/datasets/uk2007/features/`

Algorithm 1 Algorithm for translating Portuguese term counts for evaluation by an English model

```
for all Top 10,000 most frequent English terms en do
    count[en] = count of term en in host h
for all Top 10,000 most frequent Portuguese terms pt of
at least four letters do
    count_pt[pt] = count of term pt in host h
    variants = number of single-term English
               translations of pt
    if variants > 0 then
        for all en: single-term Portuguese translations of pt
        do
            count[en]+ = count_pt[pt]/variants
Classify h using term counts count[en]
```

- Multiple translation alternatives exist. We consider all translations, but we split the term frequency value between them in order not to overweight terms with many translations. A smarter but more complex weighting method is described in [39].
- Multi-word translation, such as *Monday* through *Friday* translated into *Segunda* through *Sexta feira*, cannot be handled based on single term frequencies. Since counting expressions (multi-word sequences) would complicate the process, we omitted this step in our experiments.
- Portuguese terms may coincide with English ones and counted in the first **for** loop. And they may have no translation, in which case the term is omitted.

We use the BM25 term weighting scheme. Let there be H hosts consisting of an average $\overline{\ell}$ terms. Given a term t of frequency f over a given host that contains ℓ terms, the weight of t in the host becomes

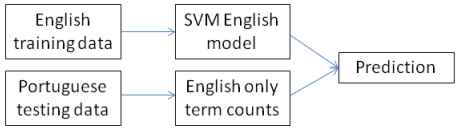
$$\log \frac{H - h + 0.5}{h + 0.5} \cdot \frac{f(k+1)}{f + k(1 - b + b \cdot \frac{\ell}{\overline{\ell}})}. \qquad (1)$$

This expression turned out to perform best in our earlier results [15]. As optimal parameters, an exceptionally low value $k = 1$ and a large $b = 0.5$ turned out to perform best in preliminary experiments. Low k means very quick saturation of the term frequency function while large b downweights content from very large Web hosts. We do not show extensive experiments on these parameters.

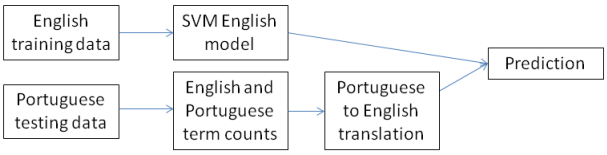
3.4 Semi-supervised cross-lingual learning based on multilingual Web sites

A large portion of national language Web content appears on the same host in English version as well, as seen in Fig. 3. This figure shows the proportion of the total frequency of the 10,000 most frequent Portuguese terms within the sum of the Portuguese and English top 10,000 frequencies. This fact gives rise to several options of English, Portuguese and mixed language text classification. As summarized in Fig. 2, the simplest solution is to ignore non-English content and simply use term frequencies of the most frequent English terms as measured over the English part of ClueWeb09. Another solution, as described in Section 3.3, is to translate the whole content term by term into English and use the model trained over ClueWeb09 again.

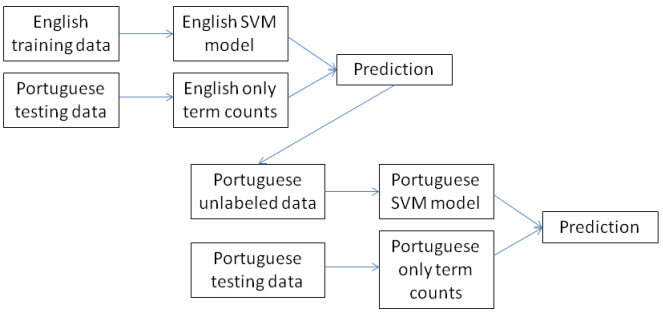
We may however rely on mixed language hosts to classify without using a dictionary in a semi-supervised proce-



(a) Prediction by using the English terms only.



(b) Terms in the English model translated into Portuguese to classify in the target language.



(c) After applying the method of Fig. 2(a), strongest positive and negative predictions are used for training a model in the target language.

Figure 2: Three methods for classifying mixed language content based on a monolingual training set.

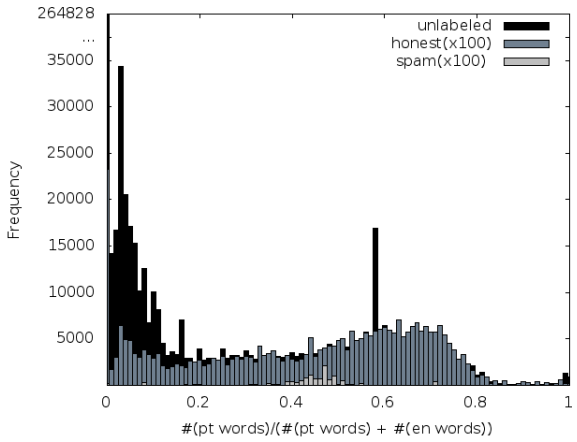


Figure 3: Statistics for the language distribution of most frequent terms in Web hosts over the .pt domain, with the 257,000 English-only hosts removed, separate for spam, ODP and unlabeled hosts. A very large fraction of the unlabeled hosts is English only, shown with a break in the horizontal scale.

dure using these (unlabeled) hosts. In Algorithm 2 we give a two-step stacked classification procedure summarized in Fig. 2(c). First we select hosts that contain an appropriate mix of English and Portuguese terms, the middle range in Fig. 3 between threshold_low = 0.4 and threshold_high = 0.6. Based on the English term frequencies of these hosts, we give prediction using a model trained over the English part of ClueWeb09. Now we turn to Portuguese term count based modeling. Even in the case when no labeled Portuguese training data exists, we may now use the outcome of the English model as training labels. More precisely, if a host has predicted value less than pred_low = 0.1, then we use the host as a negative, and if more than pred_high = 0.9, then as a positive training instance.

Algorithm 2 Stacked classification of mixed-language hosts based on an English model

```
for all hosts h do
  ratio[h] = total frequency of top 10,000 Portuguese
  terms divided by total frequency of top 10,000 Por-
  tuguese and English terms
  if threshold_low < ratio[h] < threshold_high then
    pred[h] = prediction for h based on the English model
    if pred[h] < pred_low then
      Add h to negative training instances
    if pred[h] > pred_high then
      Add h to positive training instances
Train a model based on Portuguese term counts using the
positive and negative instances h
Classify all testing h using the Portuguese only model
```

We select the intermediate training set efficiently by first running a MapReduce job only to count the dictionary term distribution, and then compute features for the selected hosts but not for the others.

We also note that the procedure summarized by the scheme in Fig. 2(c) can be used with any classifier and feature set. In addition to training using Portuguese term frequencies, we also compute the public content based features and compare models trained on ClueWeb09 vs. the semi-supervised "training" set.

3.5 Classification Framework

In our classifier ensemble we split features into related sets as described in Sections 3.1–3.3 and for each set we use a collection of classifiers that fit the data type and scale. These classifiers are then combined by ensemble selection. We used the classifier implementations of the machine learning toolkit Weka [45]. We use a procedure similar to [15] that we summarize here.

In the context of combining classifiers for Web classification, to our best knowledge, ensemble selection was only used by our previous result [15]. Before that, only simple methods that combine the predictions of SVM or decision tree classifiers through logistic regression or random forest have been used [11]. We believe that the ability to combine a large number of classifiers while preventing overfitting makes ensemble selection an ideal candidate for Web classification, since it allows us to use a large number of features and learn different aspects of the training data at the same time. Instead of tuning various parameters of different classifiers, we can concentrate on finding powerful features and selecting the main classifier models which we believe to be able to capture the differences between the classes to be distinguished.

We used Weka ensemble selection [45] for performing the experiments. We allow Weka to use all available models in the library for greedy sort initialization and use 5-fold embedded cross-validation during ensemble training and building. We set AUC as the target metric to optimize for and run 100 iterations of the hillclimbing algorithm.

We use the following model types for building the model library for ensemble selection: bagged and boosted decision trees, logistic regression, LogitBoost, naive Bayes and random forests. For most classes of features we use all classifiers and allow selection to choose the best ones. The exception is static term vector based features, where, due to the very large number of features, we use SVM as described in Sections 3.3–3.4.

4. EXPERIMENTS

We evaluate the performance of the proposed classification approach on a 2009 crawl of the Portuguese Web Archive of more than 600,000 domains and 70M pages. For training our English language models, we used the English part of ClueWeb09 of approximately 20M domains and 500M pages. Web spam labels were provided by the Portuguese Web Archive and the Waterloo Spam Rankings [12], respectively. While the Waterloo Spam Rankings contain negative training instances as well, for the Portuguese data we used pages from the Open Directory Project (ODP) for this purpose. The distribution of labels and the number of pages in labeled and all hosts is seen in Fig. 4. In our results we use the ClueWeb09 labels for training and the Portuguese Web Archive data for testing only, thus measuring the case when training only over English language labeled data.

We use the area under the ROC curve (AUC) [19] as used at Web Spam Challenge 2008 [7] to evaluate our classifiers. We do not give results in terms of precision, recall, F-measure or any other measure that depends on the selection of a threshold, as these measures are sensitive to the threshold and do not give a stable comparison of two results. These measures, to our best knowledge, were not used in Web classification evaluation campaigns since after Web Spam Challenge 2007.

4.1 Feature distributions

As seen by the language distribution in Fig. 3, our Portuguese testing data set consists of hosts with English to Portuguese ratio uniformly spread between mostly English to fully Portuguese, with the exception of a large number of English only hosts. These latter hosts are, however, underrepresented in the labeled set that we use for testing our cross-lingual method, hence we take no specific action to classify them.

Since most often web sites are topically classified based on the strong signals derived from terms that appear on their pages, our first and often most powerful classifier is SVM over tf.idf, averaged over all pages of the host. After stop word removal, we use the most frequent 10,000 terms both in English and in Portuguese.

The distribution of content features differs significantly between ClueWeb09 and the Portuguese crawl. As an example, the relative behavior of spam compared to normal hosts also significantly differs between ClueWeb09 and the

category	.pt count	ClueWeb count
spam	124	439
honest	3375	8421
hosts	686443	19228332
pages	71656081	502368557

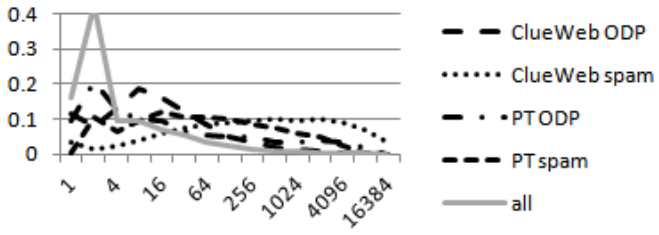


Figure 4: The number of positive and negative labeled host instances and the host and page count for the two data sets. The labeled ClueWeb data is identical to that of [12]. The chart on the right shows the fraction of labeled and all hosts with a given number of pages, with an exponential binning.

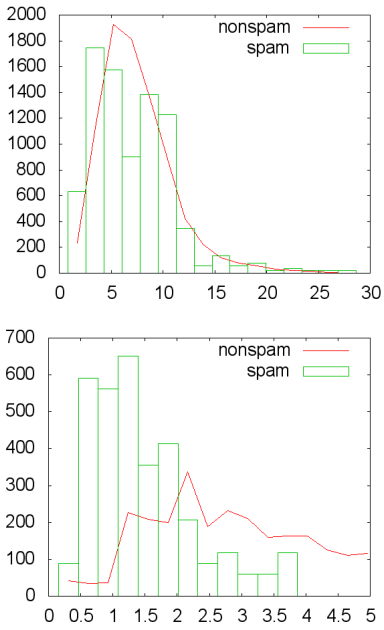


Figure 5: Distribution of the title length of the home page over the ClueWeb09 (top) and the Portuguese data (bottom), separate for spam and normal hosts.

Content ensemble	0.719
Content LogitBoost	0.751
Link	0.921
English	0.752
Translated	0.861
Stacked	0.894
Translated+Stacked avg	0.895
English+Stacked avg	0.899
English+Translated avg	0.952
English+Translated+Stacked avg	0.952
English+Link avg	0.898
Translated+Link avg	0.950
Translated+Stacked+Link avg	0.953
Stacked+Link avg	0.964
English+Translated+Link avg	0.967
English+Stacked+Link avg	0.976
English+Translated+Stacked+Link	0.976

Table 1: AUC of the main classification methods over the Portuguese test data. In the two variants of the content based features, we give results of the ensemble selection in the first and a single LogitBoost in the second column.

Portuguese data as seen in Fig. 5. Hence we may not expect content based features to work well across models.

4.2 Results

We show our results in terms of the AUC measure over the Portuguese Web test data set trained over the ClueWeb09 labels in Table 1. First, we give results obtained by using the public content and link based features [9]. These features work relative well for spam. Improved results are obtained by using LogitBoost only instead of the full classifier ensemble, as seen by comparing the first and second rows of Table 1. Link features (row 3) perform surprisingly well despite of the lack of TrustRank features.

The relative power of content and link based features over the training corpus is apparently similar. In our crossvalidation experiment over ClueWeb09, the training set, we obtain an AUC of 0.806 for content and 0.804 for linkage. For the Portuguese data, the link features trained over the ClueWeb09 corpus perform much better (0.921) than cross-validated over the training data itself. This may be due to the fact that labeled spam comes from a relative small number of link farms and hence have a very characteristic link structure.

Next, we give our results based on the bag of words representation for training in English and using labels of the Portuguese collection only for testing. Considering the Portuguese corpus as it was written in English (row "English") is clearly a bad idea, still its performance matches that of the content features. The translation model (row "Translated") works much better than the fully English one and is further improved by the stacked framework of Section 3.4 (row "Stacked"). Finally, we combine subsets of the classifiers by averaging their predicted spamicity values. The first block contains all four combinations of the three bag of words methods (English, Translated and Stacked); and the second block in addition combines with the LogitBoost classifier output over the link features. The combination of all models except the Translated one is the overall best method (last two rows). Here we observe that the combination of the English and translated classifiers can only be beaten by using the linkage features. On the other hand the Stacked model combines very well with linkage and the final best result consists of their combination with the English classifier.

Conclusion

In the paper we have demonstrated the applicability of cross-lingual Web host level spam filtering. Our experiments were tested over more than 600,000 hosts of the `.pt` domain by using the near 20M host English part of the ClueWeb09 data sets. Our results open the possibility for Web classification practice in national Internet archives who are mainly concerned about their resources, require fast reacting methods, and have very limited budget for human assessment.

By our experiments it has turned out that the strongest resources for cross-lingual classification are linkage as well as multilingual Web sites that discuss the same topic in both English and the local language. Note that these Web sites cannot be considered parallel corpora: we have no guarantee of exact translations, however, as our experiments also indicate, their content in different languages are topically identical. The use of dictionaries to translate a bag of words based model also works and combine well with other methods. The normalization of the "public" Web spam content based features [9] across languages however seems to fail; also these features perform weak for topical classification. Link based features can however be used for language-independent Web spam classification, regardless of their weakness identified in our previous result [15].

To foster further research, we provide labels and precomputed values of term frequencies, content and link based features for both ClueWeb09 and the Portuguese data available at `http://datamining.sztaki.hu/en/crosslingual/`.

Acknowledgments

To Daniel Gomes and colleagues from the Portuguese Web Archive `http://arquivo.pt/` for providing us with the Portuguese labeled data set. To Julien Masanès, Philippe Rigaux and colleagues from the Internet Memory Foundation for drawing our attention to the relevance of the problem and testing preliminary versions of our method on their data.

Work conducted at the Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI) was supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956), by the EU FP7 Project LAWA—Longitudinal Analytics of Web Archive Data, and by the grant OTKA NK 105645.

The work of Tamás Kiss reported in the paper has been developed in the framework of the project "Talent care and cultivation in the scientific workshops of BME" project. This project is supported by the grant TÁMOP - 4.2.2.B-10/1–2010-0009.

Work conducted at the Eötvös University, Budapest was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013). The research was carried out as part of the EITKIC_12-1-2012-0001 project, which is supported by the Hungarian Government, managed by the National Development Agency, financed by the Research and Technology Innovation Fund and was performed in cooperation with the EIT ICT Labs Budapest Associate Partner Group. (`www.ictlabs.elte.hu`)

5. REFERENCES

[1] L. D. Artem Sokolov, Tanguy Urvoy and O. Ricard. Madspam consortium at the ECML/PKDD discovery challenge 2010. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[2] N. Bel, C. Koster, and M. Villegas. Cross-lingual text categorization. *Research and Advanced Technology for Digital Libraries*, pages 126–139, 2003.

[3] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. SpamRank – Fully automatic link spam detection. In *Proc. 1st AIRWeb, held in conjunction with WWW2005*, 2005.

[4] A. A. Benczúr, M. Erdélyi, J. Masanés, and D. Siklósi. Web spam challenge proposal for filtering in archives. In *Proc. 5th AIRWeb, held in conjunction with WWW2009*. ACM Press, 2009.

[5] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[6] A. Bratko, B. Filipič, G. Cormack, T. Lynam, and B. Zupan. Spam Filtering Using Statistical Data Compression Models. *The Journal of Machine Learning Research*, 7:2673–2698, 2006.

[7] C. Castillo, K. Chellapilla, and L. Denoyer. Web spam challenge 2008. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[8] 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, December 2006.

[9] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. *Proc. 30th ACM SIGIR*, pages 423–430, 2007.

[10] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[11] G. Cormack. Content-based Web Spam Detection. In *Proc. 3rd AIRWeb*, 2007.

[12] G. Cormack, M. Smucker, and C. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465, 2011.

[13] G. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. Clef 2007: Ad hoc track overview. *Advances in Multilingual and Multimodal Information Retrieval*, pages 13–32, 2008.

[14] S. Dumais, T. Letsche, M. Littman, and T. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21, 1997.

[15] M. Erdélyi, A. Garzó, and A. A. Benczúr. Web spam classification: a few features worth more. In *Joint WICOW/AIRWeb Workshop on Web Quality, in conjunction with WWW2011, Hyderabad, India.* ACM Press, 2011.

[16] D. Fetterly and Z. Gyöngyi. Fifth international workshop on adversarial information retrieval on the web (AIRWeb 2009). 2009.

[17] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics – Using statistical analysis to locate spam web pages. In *Proceedings of the 7th*

International Workshop on the Web and Databases (WebDB), pages 1–6, Paris, France, 2004.

[18] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide web. In *Proc 28th ACM SIGIR*, Salvador, Brazil, 2005.

[19] J. Fogarty, R. S. Baker, and S. E. Hudson. Case studies in the use of roc curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005*, GI '05, pages 129–136, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2005. Canadian Human-Computer Communications Society.

[20] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[21] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using Web structure for classifying and describing Web pages. In *Proc. 11th WWW*, 2002.

[22] J. Göbölös-Szabó, N. Prytkova, M. Spaniol, and G. Weikum. Cross-lingual data quality for knowledge base acceleration across wikipedia editions. In *Proc. QDB*, 2012.

[23] D. Gomes, J. Miranda, and M. Costa. A survey on web archiving initiatives. In S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, editors, *Research and Advanced Technology for Digital Libraries*, LNCS vol. 6966, pages 408–420. Springer Berlin Heidelberg, 2011.

[24] D. Gomes, A. Nogueira, J. Miranda, and M. Costa. Introducing the portuguese web archive initiative. 2009.

[25] X.-C. Z. Guang-Gang Geng, Xiao-Bo Jin and D. Zhang. Evaluating web content quality via multi-scale features. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[26] Z. Gyöngyi and H. Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, October 2005.

[27] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proc 1st AIRWeb*, Chiba, Japan, 2005.

[28] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proc. 30th VLDB*, pages 576–587, Toronto, Canada, 2004.

[29] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[30] X. Ling, G. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu. Can chinese web pages be classified with english data source? In *Proc. 17th WWW*, pages 969–978. ACM, 2008.

[31] X. Ni, J. Sun, J. Hu, and Z. Chen. Cross lingual text classification by mining multilingual topics from wikipedia. In *Proc. fourth ACM WSDM*, pages 375–384. ACM, 2011.

[32] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134, 2000.

[33] V. Nikulin. Web-mining with wilcoxon-based feature selection, ensembling and multiple binary classifiers. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[34] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. 15th WWW*, pages 83–92, Edinburgh, Scotland, 2006.

[35] J. Olive, C. Christianson, and J. McCary. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer, 2011.

[36] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *Proc. 48th ACL*, pages 1118–1127. Association for Computational Linguistics, 2010.

[37] G. Ramírez-de-la Rosa, M. Montes-y Gómez, L. Villasenor-Pineda, D. Pinto-Avendano, and T. Solorio. Using information from the target language to improve crosslingual text classification. *Advances in Natural Language Processing*, pages 305–313, 2010.

[38] L. Rigutini, M. Maggini, and B. Liu. An em based training algorithm for cross-language text categorization. In *Proc. 2005 IEEE/WIC/ACM Web Intelligence*, pages 529–535. IEEE, 2005.

[39] L. Shi, R. Mihalcea, and M. Tian. Cross language text classification by model translation and semi-supervised learning. In *Proc. EMNLP 2010*, pages 1057–1067. Association for Computational Linguistics, 2010.

[40] D. Siklósi, B. Daróczy, and A. Benczúr. Content-based trust and bias classification via biclustering. In *Proc. 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 41–47. ACM, 2012.

[41] A. Singhal. Challenges in running a commercial search engine. In *IBM Search and Collaboration Seminar 2004*. IBM Haifa Labs, 2004.

[42] P. Sorg and P. Cimiano. Enriching the crosslingual link structure of wikipedia-a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, pages 49–54, 2008.

[43] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243. Association for Computational Linguistics, 2009.

[44] H. Wang, H. Huang, F. Nie, and C. Ding. Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization. In *Proc. 34th ACM SIGIR*, pages 933–942. ACM, 2011.

[45] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.

[46] B. Wu, V. Goel, and B. D. Davison. Topical TrustRank: Using topicality to combat web spam. In *Proc. 15th WWW*, Edinburgh, Scotland, 2006.

[47] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.