Modeling Website Topic Cohesion at Scale to Improve Webpage Classification

Dhivya Eswaran[*]
CS Dept., IIT Madras
Chennai, India
dhivyae@cse.iitm.ac.in

Paul N. Bennett
Microsoft Research
Redmond, WA USA
pauben@microsoft.com

Joseph J. Pfeiffer III
CS Dept., Purdue University
West Lafayette, IN USA
jpfeiffer@purdue.edu

ABSTRACT

Considerable work in web page classification has focused on incorporating the topical structure of the web (e.g., the hyperlink graph) to improve prediction accuracy. However, the majority of work has primarily focused on relational or graph-based methods that are impractical to run at scale or in an online environment. This raises the question of whether it is possible to leverage the topical structure of the web while incurring nearly no additional prediction-time cost. To this end, we introduce an approach which adjusts a page content-only classification from that obtained with a global prior to the posterior obtained by incorporating a prior which reflects the topic cohesion of the site. Using ODP data, we empirically demonstrate that our approach yields significant performance increases over a range of topics.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Text classification, large-scale, context-dependent priors

1. INTRODUCTION

Topical webpage classification has proven useful for a variety of information retrieval applications. For example, one of its most common applications is in building profiles for personalization [5, 7] including building short- and long-term topical profiles [1]. Additionally, topical classification has also been used for improving time-to-find information by organizing search results into a categorical display [3], detecting temporal shifts in query intent using changes in clicked URL and query topic distribution [9], and improving search relevance by matching query intent to topic intent [2]. Given the wide variety of applications for text classification in information retrieval, there remains interest in novel methods to improve text classification accuracy.

To this end, there has been considerable effort on incorporating the topical structure of the web (e.g., the hyperlink graph) to improve prediction accuracy [4, 6, 8]. Given both the size of the web and the frequently changing nature of web pages, production environments require efficient classifiers that can be updated with every re-crawl of a page while meeting very stringent performance demands. However, the majority of work has focused on relational or graph-based methods that are impractical to run at scale or in an online environment [4, 6, 8]. This raises the question of whether it is possible to leverage the topical structure of the web while incurring nearly no additional prediction-time cost.

Our key insight is to leverage the topical structure of the web. Namely, collections of webpages are typically organized into websites whose contents are topically related. For example, `www.espn.com` is a high-level web domain containing a collection of webpages centered around the *Sports* topic. More generally, each website tends to have a different topical focus that we view as a website-dependent prior over topics. In this paper, we demonstrate that these website-dependent priors can improve the accuracy of a classifier trained globally across all websites. Using this intuition, we develop a wrapper approach to incorporate website-dependent priors that can be used in conjunction with any probabilistic classifier.

Importantly, our method scales at a cost similar to the content-only classification of the webpage. That is, at training time we need only aggregate and estimate the priors for each website. For our purposes, a URL's corresponding 'website' is its top-level domain. Given this identification into website, it is straightforward to construct a mapping of website to distribution over topics using a map-reduce approach. Then, at prediction-time – the most computationally restrictive time in a production environment – incorporating website topical structure is a simple website look-up and a constant-time reweighting of the topical classification.

2. RELATED WORK

Here, we highlight previous work that has discussed adjusting posterior predictions when either the priors do not match [10] or attribute distributions vary [14]. Previous work discussing the priors mismatch assumes the corresponding correct priors cannot be directly estimated, leading to an iterative expectation-maximization solution. Similarly, modeling the distribution of covariates [14] for high dimensional text data would be difficult given the small sample

[*] Work performed while at Microsoft.

SIGIR '15, August 09-13, 2015, Santiago, Chile.
© 2015 ACM. ISBN 978-1-4503-3621-5/15/08... $ 15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767834

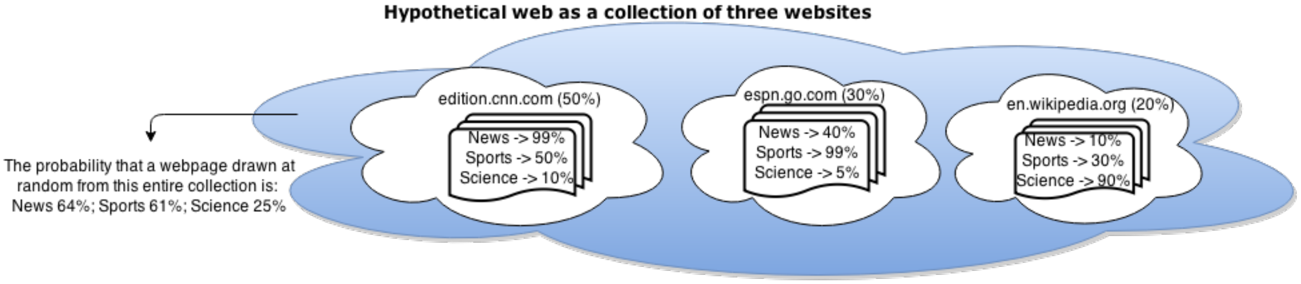

Figure 1: Example collection of three websites to demonstrate that the overall distribution of topics may vary greatly from that within website due to topical cohesion. Note that topic probabilities (over the web or within a website) need not sum to one, as we do not restrict a webpage to have only one category label.

sizes for the website-specific instances. In this work, we utilize the immediately available URL to group the items, allowing for a simpler approach to modeling the website-specific priors that is more suitable for online and time-constrained situations.

3. PROBLEM APPROACH

We hypothesize that the universe of webpages is not one homogeneous entity with a uniform distribution of labels, but is composed of many communities (websites[1]) each of which has its own prior distribution of class labels that can significantly differ from the global prior distribution of labels (Figure 1).

Ideally then, when a new webpage $\mathbf{x}$ is presented to the classifier, the model should make the class prediction c using both its content and its web domain w according to $P(c|\mathbf{x}, w)$. One way to approach this problem is to build a separate model for each website w; but this is not feasible due to limited labeled data. Alternatively, we can let all the websites share the same set of classifiers and incorporate the website specific class-distributional information in the priors alone. This is the key insight underlying our method.

3.1 Using a Website-specific Prior from a Generative Viewpoint

For explanatory purposes, it is useful to first approach how to incorporate website-specific priors from the viewpoint of a generative model. Consider Bayes' rule applied to the global distribution of the class (without using any web domain information) labels:

$$P_g(c|\mathbf{x}) \propto P(\mathbf{x}|c)P_g(c) \tag{1}$$

Here, we use a subscript g for the global prior, $P_g(c)$, but omit one on the class-conditional language model, $P(\mathbf{x}|c)$. This corresponds to the following hypothesized generative model for webpages: first, a class is picked from the global class distribution $P_g(c)$. Then, conditioned on the class, words are generated according to the language model, $P(\mathbf{x}|c)$. Assuming that the words are generated according to the same class-conditional language model, $P(\mathbf{x}|c) = P(\mathbf{x}|c, w)$, we can write:

$$\begin{aligned} P(c|\mathbf{x}, w) \propto& P(\mathbf{x}|c, w)P_w(c) \\ =& P(\mathbf{x}|c)P_w(c) \end{aligned} \tag{2}$$

While this assumption that the same class-dependent language model is used to generate content across websites may not hold in practice, it is a useful first-pass approximation. In the next section, we discuss how a similar approach can be taken with a discriminative classifier.

3.2 Discriminative Website-Specific Classifiers

Here, we wish to utilize a global discriminative classifier, $P_g(c|\mathbf{x})$, to make website-specific predictions of a webpage (i.e., $P(c|\mathbf{x}, w)$). We highlight that the conditional $P(\mathbf{x}|c)$ appears under our generative model for both the global class conditional and domain-specific conditional (i.e., Equations 1 and 2). Solving Equation 1 for $P(\mathbf{x}|c)$ and incorporating the normalization $P_g(\mathbf{x})$, we recover:

$$P(\mathbf{x}|c) = \frac{P_g(c|\mathbf{x})P_g(\mathbf{x})}{P_g(c)}$$

Inserting the above into Equation 2, we derive:

$$\begin{aligned} P(c|\mathbf{x}, w) &= \frac{P(\mathbf{x}|c)P_w(c)}{P_w(\mathbf{x})} \\ &= \frac{\frac{P_g(c|\mathbf{x})P_g(\mathbf{x})}{P_g(c)}P_w(c)}{P_w(\mathbf{x})} \\ &\propto P_g(c|\mathbf{x})\frac{P_w(c)}{P_g(c)} \end{aligned} \tag{3}$$

where in the last step we have simply dropped the variables that do not depend on the class c. The normalization of this last line can be marginalized over classes to write:

$$P(c|\mathbf{x}, w) = \frac{P_g(c|\mathbf{x})\frac{P_w(c)}{P_g(c)}}{\sum_c P_g(c|\mathbf{x})\frac{P_w(c)}{P_g(c)}} \tag{4}$$

Note the intuitive form of this conditional in Eq. 4. It only requires adjusting a classifier's prediction by the ratio of the website-specific to global prior. This process is summarized in Figure 2. Now, we turn to estimating the priors.

4. ESTIMATING PRIORS

We assume that we have a set of labeled data to both estimate the global priors and website-specific priors. While we leave the more general case of estimating website-specific priors from unlabeled data as future work, we do consider

[1] In this paper, we reserve the term "webpage" to refer to the content associated with a single URL and website to refer to a collection of webpages. For simplicity, we represent a website as the top-level domain of the URL although future work contemplates that improving identification of website-breaking of the URL may further improve our approach.

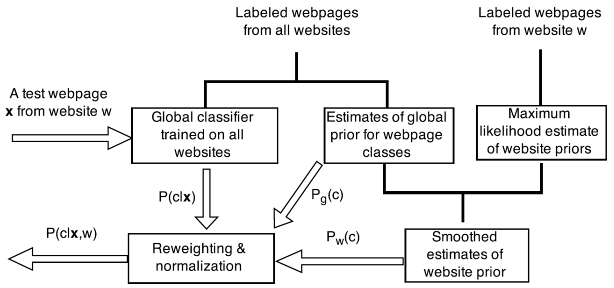
Labeled webpages from all websites
Labeled webpages from website w
A test webpage **x** from website w
Global classifier trained on all websites
Estimates of global prior for webpage classes
Maximum likelihood estimate of website priors
P(c|**x**)
P$_g$(c)
P(c|**x**,w)
P$_w$(c)
Reweighting & normalization
Smoothed estimates of website prior

Figure 2: Overview of the proposed methodology.

how to smooth estimates to deal with the realistic conditions that labeled data for any particular website may be very sparse.

4.1 Estimating the Global Prior

The global prior can be estimated from the training set T or more generally any labeled dataset L drawn from the same distribution as T. $P_g(c)$ is obtained using the maximum likelihood estimate, i.e., the proportion of webpages in the labeled data that actually belong to the class. If $\mathcal{L}(\mathbf{x})$ denotes the set of classes (or labels) that a webpage $\mathbf{x}$ belongs to and $\mathbb{1}(\cdot)$ is the indicator function[2], the estimate for the global prior of a class c is the usual:

$$P_g^L(c) \quad = \quad \frac{\sum_{\mathbf{x}\in L} \mathbb{1}(c \in \mathcal{L}(\mathbf{x}))}{|L|}. \tag{5}$$

4.2 Estimating Website-Specific Priors

Using a website-conditioned counterpart of Eq. 5 directly to estimate $P_w(c)$ may result in overfitting the website-specific priors to the labeled set. To overcome this problem, we smooth the website-specific prior using a Bayesian m-estimate with the global prior. That is:

$$P_w^L(c) = \frac{\sum_{\mathbf{x}\in w} \mathbb{1}(c \in \mathcal{L}(\mathbf{x})) + m_{c,w} P_g^L(c)}{|w| + m_{c,w}}, \text{ where } m_{c,w} > 0 \tag{6}$$

Here, $m_{c,w}$ can be regarded as a virtual count of examples observed from the global distribution for this particular website and class. We focus on the case where there is a single value of m across all classes and websites, *i.e.,* $m = m_{c,w}$ for all c, w. By tying the parameter value across classes and websites in this way, the method is less prone to overfitting.

5. EMPIRICAL METHODOLOGY

As it is commonly used in many topic classification studies, we used the Open Directory Project (ODP, `dmoz.org`) ontology of the web to study the empirical effectiveness of our proposed approach. We crawled these web pages and randomly split them into 70% train, 15% validation, and 15% test. We chose to restrict the class set to focus on those categories with more than 1K positive training examples. This resulted in the 219 most frequent categories at the top two levels of the hierarchy. While still at a high-level of granularity, this is a similar number of categories shown to improve downstream applications such as personalization [1, 11]. Furthermore, like those works we choose to flatten the hierarchy and deal with the issue that a document can belong to multiple categories by training a binary classifier per class. We focus on the impact of predictive performance on the top category and the top three categories as use of the top three categories tend to be most common in the literature in downstream applications such as personalization. As standard in text classification, we examine the micro and macro averages of precision, recall, and F1 [13, 12].

Baseline
To create our baseline global classifier, we train a binary logistic regression classifier with L2 regularization on each of the classes. We use a tfidf representation where each feature vector is normalized to the unit sphere to deal with length variation. When optimized by a score cut threshold over the validation set, the micro-average F1 of the baseline is 0.60 which is a reasonably strong baseline whose performance is on a par with similar classifiers used in downstream applications [1, 11].

Cohesion
The *Cohesion* approach reported uses the labeled training data to estimate the website-specific priors according to Eq. 6 and incorporate them into the website specific posterior as in Eq. 4. Treating each binary classifier separately, we performed binary normalization in the denominator using its positive and negative examples. The final posteriors are used to rank the classes. A single value of the smoothing parameter m is chosen by doing a sweep over the validation data over $m \in \{0,\ 0.01,\ 0.025,\ 0.05,\ 0.075,\ 0.1,\ 0.25,\ 0.5,\ 0.75,\ 1,\ 2.5,\ 5,\ 7.5,\ 10,\ 25,\ 50,\ 75,\ 100,\ 250,\ 500,\ 750,\ 1000\}$ and selecting the value of m that yielded the best reduction in residual macro- and micro-F1 (equally weighting both).

6. RESULTS AND DISCUSSION

Table 1 presents the summary of each method when predicting the top category and top 3 categories for the value of m selected on the validation set. For the rank 3 cutoff, Figure 3 (left) shows the variation of residue with m while Figure 3 (right) shows a histogram of number of classes impacted versus the per-class change in F1 performance. First, we note in the table that the Cohesion method improves both macro and micro F1 at both rank 1 and rank 3 cutoffs. In fact, the Cohesion method improves every measure except for macro-precision at rank 1. Taken together with the improvement in micro-precision at rank 1, we see the Cohesion method hurts precision across a majority of classes although it improves for the more common classes (which have a larger impact on the micro-average). As the rank cutoff is increased to rank=3, we see that all measures improve – indicating that those categories that suffered at rank=1 still remain quite high in the ranks. At rank=3, Cohesion improves the precision/recall/F1 on 62%/77%/75% of classes, respectively. Thus, the gains are very robust over a large number of classes. Figure 3 (right) shows further detail on the distribution of gains/losses for F1 for rank=3.

To explain the trend observed in Figure 3 (left), as described earlier m indicates the pseudo count of webpages from the global distribution. Low values of m overfit the website-specific prior to the labeled set and a high residual F1 (error) is observed. As m is increased, the contribution from the global prior increases till m is just enough to offset the overfitting to training data. If m is increased beyond this value, the performance is again hurt as the topic cohesion of the websites are increasingly ignored.

[2] For a boolean variable y, $\mathbb{1}(y) = 1$ if y is true; else $\mathbb{1}(y) = 0$.

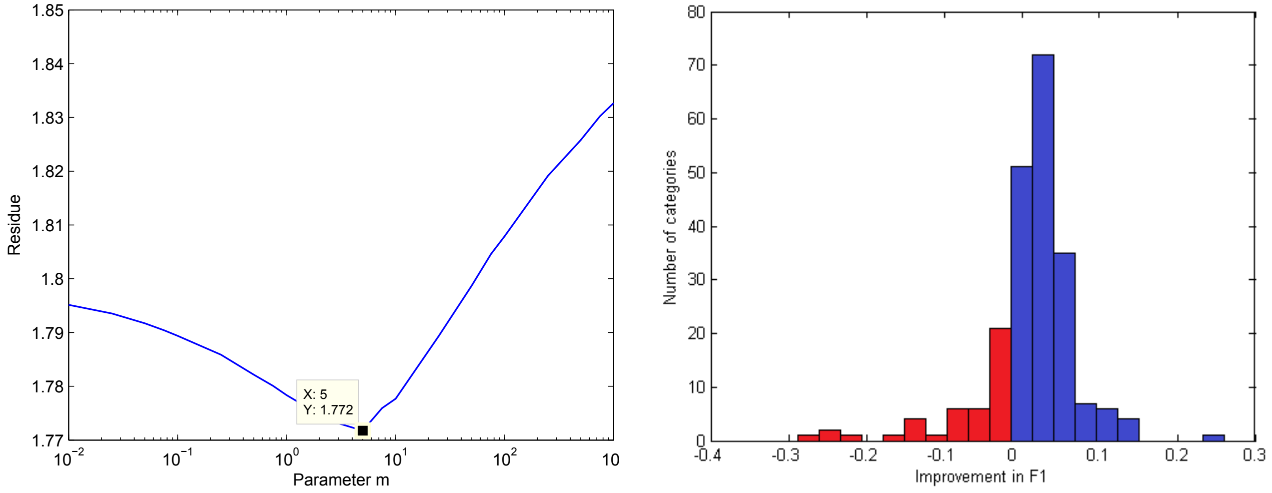

Figure 3: Predictions on the top three categories: variation of residual micro- and macro- F1 (left) and the distribution of absolute (right) improvement of the Cohesion model of the 165 improved categories (blue) and 54 hurt categories (red) (no categories tied).

	Method	Micro-Precision	Micro-Recall	Micro-F1	Macro-Precision	Macro-Recall	Macro-F1
Rank=1	Baseline	0.6509	0.3423	0.4487	**0.7683**	0.1623	0.2494
	Cohesion	**0.6923**	**0.3641**	**0.4772**	0.7328	**0.2006**	**0.2969**
Rank=3	Baseline	0.4180	0.6595	0.5117	0.4329	0.4423	0.4277
	Cohesion	**0.4353**	**0.6868**	**0.5328**	**0.4428**	**0.4718**	**0.4410**

Table 1: Predicting the top and top 3 categories for webpages in the test set. The best performer for each rank cutoff in each column is shown in bold. Underline indicates significant differences ($p \leq 0.05$) in macro-averages according to a two-sided macro-sign test [13].

7. CONCLUSION AND FUTURE WORK

In summary, we demonstrated how the topic cohesion of websites could be leveraged at scale to improve webpage classification through website-specific priors. Though our current approach is limited to the availability of labeled data for each domain, the significant overall gains we obtained in our experiments with a large majority of classes (75%) improving over the baseline as well demonstrate the effectiveness and robustness of our method. Future work in this area could investigate the amount of the labeled training data required for our approach and examine cautious techniques to learn website-specific priors from unlabeled data to generalize across websites where labeled data may be completely unavailable. In addition, we are also interested in methods that go beyond the top-level domain representation to model cohesion as a hierarchical relationship.

8. REFERENCES

[1] P. Bennett et al. Modeling the impact of short- and long-term behavior on search personalization. In *SIGIR '12*, 2012.

[2] P. N. Bennett, K. Svore, and S. T. Dumais. Classification-enhanced ranking. In *WWW '10*, 2010.

[3] S. Dumais, E. Cutrell, and H. Chen. Bringing order to the web: Optimizing search by showing results in context. In *CHI'01*, 2001.

[4] N. Friedman et al. Learning probabilistic relational models. In *In IJCAI '99*, 1999.

[5] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based user profiles for search and browsing. *WIAS*, 1(3-4):219–234, 2004.

[6] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[7] Z. Ma, G. Pant, and O. R. L. Sheng. Interest-based personalized search. *TOIS*, 25(1), Feb. 2007.

[8] X. Qi and B. D. D. 0001. Classifiers without borders: incorporating fielded text from neighboring web pages. In *SIGIR '08*, 2008.

[9] K. Radinsky et al. Modeling and predicting behavioral dynamics on the web. In *WWW '12*, 2012.

[10] M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

[11] R. W. White, P. N. Bennett, and S. T. Dumais. Predicting short-term interests using activity-based search context. In *CIKM '10*, 2010.

[12] Y. Yang. An evaluation of statistical approaches to text categorization. *Inf. Retr.*, 1(1-2):69–90, May 1999.

[13] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR '99*, 1999.

[14] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML '04*, 2004.