Latent Dirichlet Allocation in Web Spam Filtering *

István Bíró Jácint Szabó András A. Benczúr
Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{ibiro, jacint, benczur}@ilab.sztaki.hu

ABSTRACT

Latent Dirichlet allocation (LDA) (Blei, Ng, Jordan 2003) is a fully generative statistical language model on the content and topics of a corpus of documents. In this paper we apply a modification of LDA, the novel *multi-corpus LDA* technique for web spam classification. We create a bag-of-words document for every Web site and run LDA both on the corpus of sites labeled as spam and as non-spam. In this way collections of spam and non-spam topics are created in the training phase. In the test phase we take the union of these collections, and an unseen site is deemed spam if its total spam topic probability is above a threshold. As far as we know, this is the first web retrieval application of LDA. We test this method on the UK2007-WEBSPAM corpus, and reach a relative improvement of 11% in F-measure by a logistic regression based combination with strong link and content baseline classifiers.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*

General Terms

Text Analysis, Feature Selection, Document Classification, Information Retrieval

Keywords

Web content spam, latent Dirichlet allocation

1. INTRODUCTION

Identifying and preventing spam is cited as one of the top challenges in web search engines in [14, 20]. As all major search engines incorporate anchor text and link analysis algorithms into their ranking schemes, web spam appears in sophisticated forms that manipulate content as well as linkage [12].

In this paper we demonstrate the applicability of topic based natural language models for Web spam filtering. Several such models have been developed in the field of information retrieval. Hofmann [15] introduced probabilistic latent semantic indexing (PLSI), which is a generative, graphical model enhancing latent semantic analysis by a sounder probabilistic model. Although PLSI had promising results, it suffers from two limitations: the number of parameters is linear in the number of documents, and it is not possible to make inference for unseen data.

These issues are addressed by latent Dirichlet allocation developed by Blei, Ng and Jordan [4]. LDA is a fully generative graphical model for describing the latent topics of documents. LDA models every topic as a distribution over the words of the vocabulary, and every document as a distribution over the topics. These distributions are sampled from Dirichlet distributions. The words of the documents are drawn from the word distribution of a topic which was just drawn for this word from the topic distribution of the document. There are several methods developed for making inference in LDA such as variational expectation maximization [4], expectation propagation [17], and Gibbs sampling [11]. LDA is an intensively studied model, and the experiments are really impressive compared to other known information retrieval techniques.

LDA has several applications including in entity resolution [3], fraud detection in telecommunication systems [24], and image processing [8, 22], in addition to the large number of applications in the field of text retrieval. To our best knowledge our experiments provide the first application of LDA in web spam filtering, and even in Web retrieval.

In this paper we introduce and apply a slight modification of LDA, called *multi-corpus LDA* as follows. Assume we have a text classification task with m classes. We run LDA separately for each class of the training set, then take the union of the resulting topic collections and make inference w.r.t. this aggregated collection of topics for every unseen document d. The total probability of class i topics in the topic distribution of d may serve as a measure to what extent d belongs to class i. For a more detailed description, see Subsection 2.1.

*This work was supported by the EU FP7 project LiWA – Living Web Archives and by grants OTKA NK 72845, *ASTOR* NKFP 2/004/05

In our experiments we run multi-corpus LDA with $m = 2$ classes: spam and non-spam. The inference is performed using Gibbs sampling. The total probability of spam topics in the topic distribution of an unseen document gives an *LDA prediction* of being spam or honest.

1.1 Related results

Spam hunters use a variety of content based features [5, 9, 18, 10] to detect web spam; a recent measurement of their combination appears in [6]. Perhaps the strongest SVM based content classification is described in [1]. An efficient method for combining several classifiers is the use of logistic regression, as shown by Lynam and Cormack [16].

Closest to our methods are the content based email spam detection methods applied to Web spam presented at the Web Spam Challenge 2007 [7]. They use the method of [5] that compresses spam and nonspam separately; features are defined based on how well the document in question compresses with spam and nonspam, respectively. Our method is similar in the sense that we also build separate spam and nonspam content models.

1.2 Data set, evaluation, experimental setup

We test the multi-corpus LDA method in combination with the Web Spam Challenge 2008 public features[1], SVM over pivoted tf.idf [21], and the connectivity sonar features (analysis of the breadth-first and directory levels within a host together with the internal and external linkage) of [2]. Using logistic regression to aggregate these classifiers, the multi-corpus LDA method yields an improvement of 11% in F-measure and 1.5% in ROC. For classification we used the machine learning toolkit Weka [23]. For a detailed explanation, see Section 3.

2. METHOD

First we describe latent Dirichlet allocation [4]. For a detailed elaboration, we refer to Heinrich [13]. We have a vocabulary V consisting of words, a set T of k topics and n documents of arbitrary length. For every topic z a distribution φ_z on V is sampled from $\mathrm{Dir}(\beta)$, where $\beta \in \mathbb{R}_+^V$ is a smoothing parameter. Similarly, for every document d a distribution ϑ_d on T is sampled from $\mathrm{Dir}(\alpha)$, where $\alpha \in \mathbb{R}_+^T$ is a smoothing parameter.

The words of the documents are drawn as follows: for every word position of document d a topic z is drawn from ϑ_d, and then a word is drawn from φ_z and filled into the position.

LDA can be thought of as a Bayesian network, see Figure 1.

One method for making inference for LDA is Gibbs sampling [11]. Gibbs sampling is a Markov chain Monte Carlo algorithm for sampling from a joint distribution $p(x)$, $x \in \mathbb{R}^n$, if all conditional distributions $p(x_i|x_{-i})$ are known ($x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$). The k^{th} transition $x^{(k)} \to x^{(k+1)}$ of the Markov chain is generated as follows. Choose an index $1 \le i \le n$ (usually $i = k \bmod n$), and let $x^{(k+1)} = x^{(k)}$ everywhere except at index i where $x_i^{(k+1)}$ is sampled from $p(x_i|x_{-i}^{(k)})$.

[1]Downloaded from http://www.yr-bcn.es/webspam/datasets/uk2007/features/ in March 31, 2008.

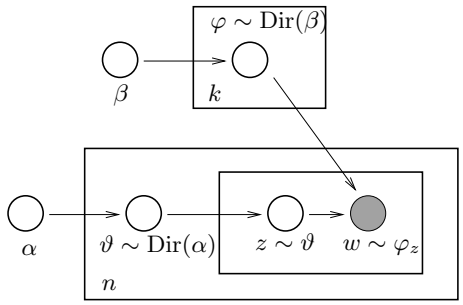

Figure 1: LDA as a Bayesian network

In LDA the goal is to estimate the distribution $p(z|w)$ for $z \in T^P$, $w \in V^P$ where P denotes the set of word positions in the documents. Thus for Gibbs sampling one has to calculate $p(z_i|z_{-i}, w)$ for $i \in P$. This has an efficiently computable closed form (for a deduction, see [13])

$$p(z_i|z_{-i}, w) = \frac{n_{z_i}^{t_i} - 1 + \beta_{t_i}}{n_{z_i} - 1 + \sum_t \beta_t} \cdot \frac{n_d^{z_i} - 1 + \alpha_{z_i}}{n_d - 1 + \sum_z \alpha_z}. \quad (1)$$

Here d is the document of position i, t_i is the actual word in position i, $n_{z_i}^{t_i}$ is the number of positions with topic z_i and word t_i, n_{z_i} is the number of positions with topic z_i, $n_d^{z_i}$ is the number of topics z_i in document d, and n_d is the length of document d. After a sufficient number of iterations we arrive at a topic assignment sample z. Knowing z, we can estimate φ and ϑ as

$$\varphi_{z,t} = \frac{n_z^t + \beta_t}{n_z + \sum_t \beta_t} \quad (2)$$

and

$$\vartheta_{d,z} = \frac{n_d^z + \alpha_z}{n_d + \sum_z \alpha_z}. \quad (3)$$

For an unseen document d the ϑ topic distribution can be estimated exactly as in (3) once we have a sample from its word–topic assignment z. Sampling z can be performed with a similar method as before, but now only for the positions i in d:

$$p(z_i|z_{-i}, w) = \frac{\tilde{n}_{z_i}^{t_i} - 1 + \beta_{t_i}}{\tilde{n}_{z_i} - 1 + \sum_t \beta_t} \cdot \frac{n_d^{z_i} - 1 + \alpha_{z_i}}{n_d - 1 + \sum_z \alpha_z}. \quad (4)$$

The notation $\tilde{n}$ refers to the union of the whole corpus and document d.

In the next subsection we will make use of the observation that the first factor in product (4) is approximately equal to φ_{z_i,t_i} by (2).

2.1 Multi-corpus LDA

As outlined in the introduction, in the multi-corpus setting we run two distinct LDA's: one in the collection of labeled spam sites with $k^{(s)}$ topics, called *spam topics*, and one in the collection of labeled non-spam sites with $k^{(n)}$ topics, called *non-spam topics*. The vocabulary is the same for both LDA's. After both inferences have been done, we have word distributions for all $k^{(s)} + k^{(n)}$ topics.

From now on we think of the obtained word distributions of the unified collection of spam and non-spam topics as

if they were estimated from only one presumed corpus. To make inference for an unseen document d, we perform Gibbs sampling on this presumed unique distribution using (4). Observe that the $\tilde{n}$ terms in the first factor of the product are not known, as also the topic assignments of the presumed corpus are not known. Thus we approximate this first factor by φ_{z_i,t_i}, and $p(z_i|z_{-i},w)$ by

$$p(z_i|z_{-i},w) \approx \varphi_{z_i,t_i} \cdot \frac{n_d^{z_i} - 1 + \alpha_{z_i}}{n_d - 1 + \sum_z \alpha_z}, \qquad (5)$$

which is a closed form expression that can be computed in $O(P \cdot k)$ steps where P is the number of word occurrences in the corpus and k is the number of topics. To distinguish this method from the original Gibbs sampling inference developed in [11], we call it the *multi-corpus inference*. This is applied only to unseen documents.

After a sufficient number of iterations we calculate ϑ_d as in (3), and define the *LDA prediction* to be $\sum\{\vartheta_{d,z} : z \text{ is a spam topic}\}$. As a simplest solution we may classify d as spam if its LDA prediction is above a certain threshold.

3. EXPERIMENTS

The data set we used is the UK2007-WEBSPAM corpus. We kept only the labeled sites with 203 labeled as spam and 3589 as non-spam. We aggregated the words and meta keywords appearing in all pages of the sites to form one document per site in a bag of words model (only multiplicity and no order information used). We kept only alphanumeric characters and the hyphen but removed all words containing a hyphen not between two alphabetical words. We deleted all stop words enumerated in the list of http://www.lextek.com/manuals/onix/stopwords1.html, and used the TreeTagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ for stemming. After this procedure the most frequent 22,000 words formed the vocabulary.

We used Phan's GibbsLDA++ C++ code [19] to run LDA and a modified version of it to run the multi-corpus inference for unseen documents in multi-corpus LDA. We applied 5-fold cross validation. Two LDA's were run on the training spam and non-spam corpora, and then multi-corpus inference were made to the test documents by Gibbs sampling as in (5).

The Dirichlet parameter β was chosen to be constant 0.1 throughout, while $\alpha^{(s)} = 50/k^{(s)}$, $\alpha^{(n)} = 50/k^{(n)}$, and during multi-corpus inference α was constant $50/(k^{(s)} + k^{(n)})$ (these are the default values in GibbsLDA++).

We stopped Gibbs sampling after 2000 steps for inference on the training data, and after 1000 steps for the multi-corpus inference for an unseen document.

The number of topics were chosen to be $k^{(s)} = 2, 5, 10, 20$ and $k^{(n)} = 10, 20, 50$. Consequently, we performed altogether 12 runs, and thus obtained 12 one-dimensional LDA predictions as features. By observing the F-measure curves as shown in Figure 2 we selected the three best performing parameter choices. F-measures and ROC values are shown in Table 1. The best result corresponds to the choice $k^{(s)} = 10$ and $k^{(n)} = 50$ with an F-measure of 0.46.

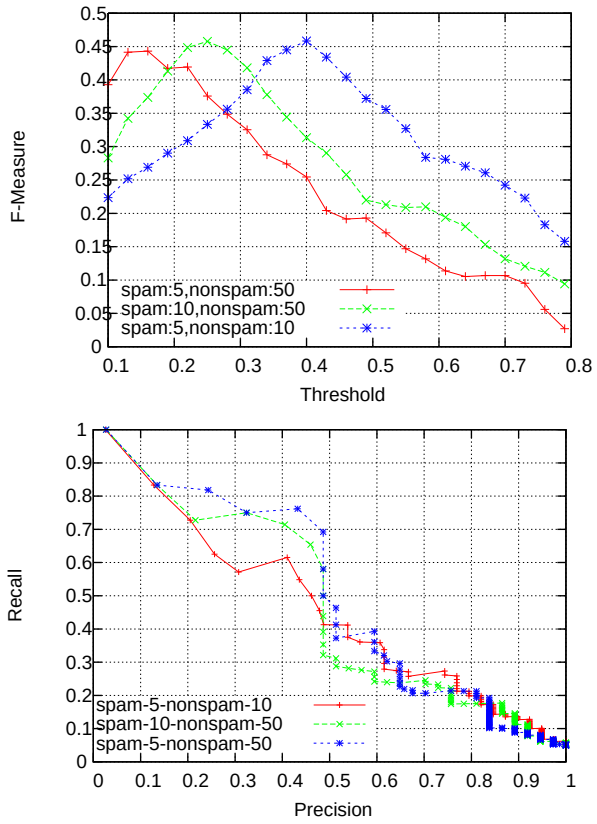

Figure 2: F-measure curves with varying thresholds (horizontal axis), and precision-recall curves of the three best LDA features.

pair of topic numbers	F	ROC
5/50	0.451	0.855
10/50	0.458	0.861
5/10	0.458	0.868

Table 1: F-measures and ROC values of the three best performing LDA predictions.

Figure 2 indicates that the multi-corpus method is robust to the parameter of topic numbers, as the performance does not really change by changing the topic numbers. As one can expect, the maximum of such an F-measure curve is approximately $k^{(s)}/k^{(n)}$.

We combined the single best performing $k^{(s)} = 10$, $k^{(n)} = 50$ LDA prediction with an SVM classifier over the tf.idf features,a C4.5 classifier over the public and the sonar features with logistic regression. All methods were performed by the machine learning toolkit Weka [23]. The F-measures and ROC values are shown in Table 2. LDA improved a relative 11% over the F-measure and 1.5% over the ROC of the remaining combined features.

We also performed single-feature classification for the three best performing LDA predictions over the UK2006-WEBSPAM corpus. Here we have 2125 sites labeled as spam, and 8082

feature set	F	ROC
text (SVM)	0.554	0.864
public & text & sonar (log)	0.601	0.954
public & text & sonar & lda (log)	0.667	0.969

Table 2: F/ROC values

pair of topic numbers	F	ROC
5/50	0.704	0.881
10/50	0.735	0.902
5/10	0.723	0.906

Table 3: F-measures and ROC values for UK2006-WEBSPAM

labeled as non-spam. The parameters and the setup was the same as above. The results can be seen at Table 3.

Conclusion and future work

We presented a novel multi-corpus LDA technique that resulted in a relative improvement of about 10% over a strong content and link feature baseline. Although apparently the UK2007-WEBSPAM data is much more sensitive to content than to link features, we reached improvement over the UK2006-WEBSPAM data as well. We believe that similar to the success of email spam filtering methods [7] semantic analysis to spam filtering is a promising direction. In future work we plan to implement the email content filtering methods of [7] and test its combination with LDA.

4. REFERENCES

[1] J. Abernethy, O. Chapelle, and C. Castillo. WITCH: A New Approach to Web Spam Detection. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[2] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The Connectivity Sonar: Detecting site functionality by structural patterns. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HT)*, pages 38–47, Nottingham, United Kingdom, 2003.

[3] I. Bhattacharya and L. Getoor. A latent dirichlet model for unsupervised entity resolution. *SIAM International Conference on Data Mining*, 2006.

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(5):993–1022, 2003.

[5] 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.

[6] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. Technical report, DELIS – Dynamically Evolving, Large-Scale Information Systems, 2006.

[7] G. Cormack. Content-based Web Spam Detection. In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2007.

[8] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. *Proc. CVPR*, 5, 2005.

[9] 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.

[10] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide web. In *Proceedings of the 28th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil, 2005.

[11] T. Griffiths. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235, 2004.

[12] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan, 2005.

[13] G. Heinrich. Parameter estimation for text analysis. Technical report, Technical Report, 2004.

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

[15] T. Hofmann. Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[16] T. Lynam, G. Cormack, and D. Cheriton. On-line spam filter fusion. *Proc. of the 29th international ACM SIGIR conference on Research and development in information retrieval*, pages 123–130, 2006.

[17] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. *Uncertainty in Artificial Intelligence (UAI)*, 2002.

[18] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, pages 83–92, Edinburgh, Scotland, 2006.

[19] X.-H. Phan. `http://gibbslda.sourceforge.net/`.

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

[21] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Information Processing and Management*, 32(5):619–633, 1996.

[22] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering Objects and their Localization in Images. *Computer Vision, ICCV 2005. Tenth IEEE International Conference on*, 1, 2005.

[23] 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.

[24] D. Xing and M. Girolami. Employing Latent Dirichlet Allocation for fraud detection in telecommunications. *Pattern Recognition Letters*, 28(13):1727–1734, 2007.