Linked Latent Dirichlet Allocation in Web Spam Filtering [*]

István Bíró Dávid Siklósi 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, sdavid, 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 an extension of LDA for web spam classification. Our linked LDA technique takes also linkage into account: topics are propagated along links in such a way that the linked document directly influences the words in the linking document. The inferred LDA model can be applied for classification as dimensionality reduction similarly to latent semantic indexing. We test linked LDA on the WEBSPAM-UK2007 corpus. By using BayesNet classifier, in terms of the AUC of classification, we achieve 3% improvement over plain LDA with BayesNet, and 8% over the public link features with C4.5. The addition of this method to a log-odds based combination of strong link and content baseline classifiers results in a 3% improvement in AUC. Our method even slightly improves over the best Web Spam Challenge 2008 result.

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 [16, 22]. 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 [14].

In this paper we demonstrate the applicability of topic based natural language models for Web spam filtering. Several such generative models [8, 17, 4] have been developed in the field of information retrieval. One of the most successful generative topic models is latent Dirichlet allocation (LDA) developed by Blei, Ng and Jordan [4], which is a fully generative graphical model with astonishing performance in various tasks. Several LDA extensions are known with wide range of applications in the fields of language processing, text mining and information retrieval, including categorization, keyword extraction, similarity search and statistical language modeling.

Recently several models extend LDA to exploit links between web documents or scientific papers [7, 10, 9, 20]. In these models the term and topic distributions may be modified along the links. All these models have the drawback that every document is thought of either citing or cited, in other words, the citation graph is bipartite, and influence flows only from cited documents to citing ones.

In this paper we apply the recently developed **linked LDA** model [3], in which each document can cite to and be cited by others and thus be influenced and influence other documents. Linked LDA is very similar to the citation influence model of Dietz, Bickel and Scheffer [9] with the main difference that in linked LDA the citation graph is not restricted to be bipartite. This fact and its consequences are the main advantage of linked LDA, namely, that the citation graph is homogeneous with no need for a citing and a cited copy of of each document, and finally, that influence may flow along paths of length more than one, a fact that gives power to learning over graphs [24, 18]. In addition, linked LDA gives a flexible model of all possible aspects including cross-topic relations and link selection. The model may also distinguish between sites with strong, weak or even no influence from its neighbors. The linked LDA model is described in full detail in Section 2.1.

We demonstrate the applicability of linked LDA for Web spam filtering. The inferred topic distributions of documents are used as features. To assess the prediction power of these features, we test the linked LDA method in combination with the WEBSPAM-UK2007 public features[1] and

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

AIRWeb '09, April 21, 2009 Madrid, Spain.
Copyright 2009 ACM 978-1-60558-438-6 ...$5.00.

[1] `http://www.yr-bcn.es/webspam/datasets/uk2007/features/`

SVM over tf.idf. Using a log-odds based random forest to aggregate these classifiers, the inclusion of linked LDA into the public and tf.idf features yields an improvement of 3% in AUC. For a detailed explanation, see Section 3.

1.1 Related results

Spam hunters use a variety of content based features to detect web spam [11, 12, 5, 21]; 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 log-odds averaging [19]. In this paper we apply a modification of this method for combination: a random forest over the log-odds of the classifiers.

The first probabilistic models that jointly model text and link as well as the influence of topics along links is PHITS [7] and the mixed membership model [10]. Later, several similar link based LDA models were introduced, including the copycat and the citation influence models [9] and the link-PLSA-LDA and pairwise-link-LDA models [20]. These two results extend LDA over a bipartition of the corpus into citing and cited documents such that influence flows along links from cited to citing documents. They are shown to outperform earlier methods [9, 20]. While these models generate topical relation for hyperlinked documents, in a homogeneous corpus one has to duplicate each document and infer two models for them. This is in contrast to the linked LDA model that treats citing and cited documents identically.

Our results improve over our related multicorpus LDA model [2] for Web spam detection. Multicorpus LDA separately builds LDA models for the collection of spam and normal sites, then take the union of the resulting topic collections and make inference with respect to this aggregated collection of topics for every unseen document d. The total probability of spam topics in the topic distribution of d may serve as a spamicity-measure. In this paper we do not make experiments on multicorpus LDA, but make comparison to the measurements presented in [2], see Subsection 3.2.

2. METHOD

In the classical latent Dirichlet allocation model [4] we have a vocabulary V consisting of terms, a set T of k topics and a set D of m documents of arbitrary length. For every topic $z \in T$ a distribution φ_z on V is sampled from $\mathrm{Dir}(\beta)$, where $\beta \in \mathbb{R}_+^V$ is a positive 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 positive 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 term is drawn from φ_z and filled into that position. The notation is summarized in the widely used Bayesian network representation of LDA in Figure 1. Inference is mostly done by Gibbs sampling or variational inference.

2.1 Linked LDA

The linked LDA model extends LDA to model the effect of a hyperlink between two documents on the topic and term distributions. The key idea, summarized as a Bayes net in Figure 2, is to modify the topic distribution of a position on the word plate based on a link from the current document on the document plate. Linked LDA relies on the LDA distributions φ_z and ϑ_d, but involves an additional distribution χ_d on the set $S_d = \{d \text{ and its outneighbors}\}$ for every document d, sampled from $\mathrm{Dir}(\gamma_d)$, where γ_d is a positive smoothing vector on S_d.

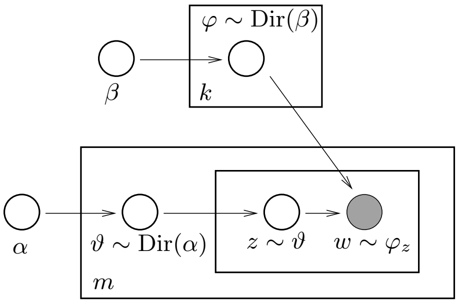
$\varphi \sim \mathrm{Dir}(\beta)$
β
k
α
$\vartheta \sim \mathrm{Dir}(\alpha)$
$z \sim \vartheta$
$w \sim \varphi_z$
m

Figure 1: LDA as a Bayesian network

As also seen in the Bayes net of Figure 2, in the linked LDA model the words of the documents are drawn as follows. For every word position i of document d, we

- draw an **influencing document** $r \in S_d$ from χ_d,
- draw a topic z from ϑ_r (instead of ϑ_d as in LDA),
- draw a term from φ_z and fill into the position.

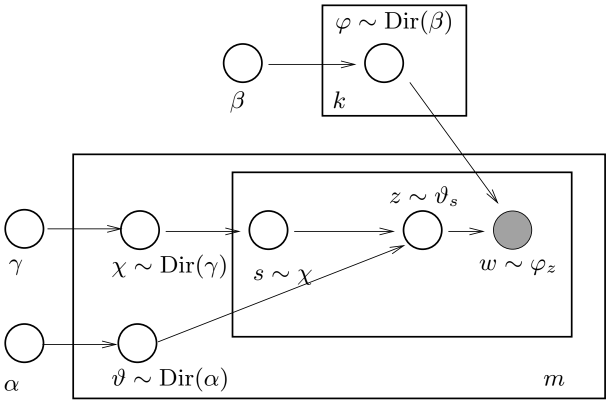
$\varphi \sim \mathrm{Dir}(\beta)$
β
k
$z \sim \vartheta_s$
γ
$\chi \sim \mathrm{Dir}(\gamma)$
$s \sim \chi$
$w \sim \varphi_z$
α
$\vartheta \sim \mathrm{Dir}(\alpha)$
m

Figure 2: Linked LDA as a Bayesian network

Note that for sake of a unified treatment, d itself can be an influencing document of itself. Note that the citation influence model [9] has a nonuniform solution such that for every word a Bernoulli trial is used to decide whether the influencing document is d itself or one of its outneighbors.

For inference a Gibbs sampling procedure can be constructed (described in [3]), along the same lines as for LDA [15].

3. EXPERIMENTS

We make experiments on the WEBSPAM-UK2007 corpus[2]. In order to define features for hosts, we aggregate the words appearing in all HTML pages of the host to form one document per host in a bag of words model. We keep only alphanumeric characters and the hyphen but remove all words containing a hyphen not between two alphabetical

[2] `http://barcelona.research.yahoo.net/webspam/-datasets/uk2007/`

words. After stemming by TreeTagger[3] and removing stop words by the Onix list[4], the most frequent 100,000 words form the vocabulary.

Our experiments include the 6000 hosts having a WEBSPAM-UK2007 spam or normal label along with an additional 39,000 hosts linked by one of these. We weight directed links between hosts by their multiplicity, and for every site we keep only at most 10 outlinks with largest weight. After linked LDA inference is run, we have the ϑ topic distribution vectors as features to the classification.

In our experiments we perform two-class spam classification. We use the linear kernel SVM, C4.5 decision tree and Bayes net implementations of the machine learning toolkit Weka [23].

As the simplest baseline we use the public features[5] with C4.5 decision tree and the tf.idf vectors with SVM. For tf.idf, only terms appearing in at least half of the sites are kept. Another baseline is formed by the ϑ topic distributions of the original LDA model as features for classification by BayesNet. Both the baseline and the linked LDA based classifiers are trained on the WEBSPAM-UK2007 training labels (3900 sites) and are evaluated on the WEBSPAM-UK2007 test labels (2027 sites) for direct comparability with the Web Spam Challenge 2008 results.

Our combination of the classifiers is inspired by the log-odds averaging by Lynam and Cormack [19]. We first make a 10-fold cross validation on the WEBSPAM-UK2007 training labels to score every host by every classifier. Then for every classifier we calculate the log-odds as a feature, which is the logarithm of the fraction of the number of spams with lower score over the number of normal sites with higher score. Finally, we train a random forest over this (quite small) feature set and give predictions for the WEBSPAM-UK2007 test labels.

3.1 LDA parameters

For LDA inference the following parameter settings are used. The number of topics is chosen to be $k = 30$ and $k = 90$. The Dirichlet parameter vector β is constant $200/|V|$, and α is constant $50/k$. For a document d, the smoothing parameter vector γ_d is chosen in such a way that

$$\gamma_d(c) \propto w(d \to c) \quad \text{for all } c \in S_d, c \neq d \text{ and}$$

$$\gamma_d(d) \propto 1 + \sum_{c \in S_d, c \neq d} w(d \to c)$$

such that $\sum_{c \in S_d} \gamma_d(c) = |d|/p$, where $|d|$ is the number of word positions in d (the document length), $w(d \to c)$ denotes the multiplicity of the $d \to c$ link in the corpus, and p is a normalization parameter. We tried three values $p = 1, 4, 10$.

We have developed an own C++-code for LDA and linked LDA, which is publicly available[6]. The computations were run on Linux machines with 50GB RAM and multicore 64-bit 3.2 GHz Xeon processors with 2MB cache.

We performed 50 iterations for Gibbs sampling as several measurements indicate that both the AUC value on classifying over the topic distributions and the likelihood stabilizes after 50 iterations [13, 3].

3.2 Results

The results of the classification can be seen in Tables 1-3, the evaluation metric is AUC. For linked LDA only the parameter choice $p = 4, k = 30$ was included in the combination, as it gave the best result.

	$p = 1$	$p = 4$	$p = 10$
$k = 30$	0.768	**0.784**	0.783
$k = 90$	0.764	0.777	0.773

Table 1: Classification accuracy measured in AUC for linked LDA with various parameters, classified by BayesNet.

features	AUC
LDA with BayesNet	0.766
tf.idf with SVM	0.795
public (link) with C4.5	0.724
public (content) with C4.5	0.782

Table 2: Classification accuracy measured in AUC for the baseline methods.

features	AUC
tf.idf & LDA	0.827
tf.idf & linked LDA	0.831
public & LDA	0.820
public & linked LDA	0.829
public & tf.idf	0.827
public & tf.idf & LDA	0.845
public & tf.idf & linked LDA	**0.854**
public & tf.idf & LDA& linked LDA	**0.854**

Table 3: Classification accuracy measured in AUC by combining the classifications of Tables 1 and 2 with a log-odds based random forest. For linked LDA the parameters are chosen to be $p = 4, k = 30$.

The tables indicate that linked LDA slightly outperforms LDA by about 3% using BayesNet, showing the predicting power of the links in the corpus. Most notably, linked LDA achieved 8% improvement over the public link features with C4.5, and it is at par with the public content features. The addition of linked LDA to the log-odds based combination of the public and tf.idf based classifiers results in a 3% improvement in AUC.

We trained the classifiers on the WEBSPAM-UK2007 training and evaluated them on the WEBSPAM-UK2007 testing labels as in the Web Spam Challenge 2008 setup. Thus we can compare these measurements to the Challenge results[7]. The present methods improve a lot over the 0.796 AUC achieved by our research group using the multicorpus LDA model. The winner, Geng et al., managed to have an AUC of 0.848, and even this value is improved by our public & tf.idf & linked LDA combination, with AUC 0.854.

[3] http://www.ims.uni-stuttgart.de/projekte/-corplex/TreeTagger/

[4] http://www.lextek.com/manuals/onix/stopwords1.html

[5] http://www.yr-bcn.es/webspam/datasets/uk2007/features/

[6] http://www.ilab.sztaki.hu/~ibiro/linkedLDA/

[7] http://webspam.lip6.fr/wiki/-pmwiki.php?n=Main.PhaseIIIResults

Conclusion and future work

In this paper we applied the newly introduced linked LDA model [3] to Web spam classification. In our experiments linked LDA outperformed LDA and other baseline classifications by about 3-8% in AUC. Combining tf.idf, the public and the linked LDA features with a log-odds based random forest we achieved an AUC of 0.854, beating the Web Spam Challenge 2008 winner (0.848). As another experiment we are currently measuring the quality of the inferred linked LDA edge weights χ, by using it in a stacked graphical classification procedure, for both the link graph and the cocitation graph.

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] I. Bíró, J. Szabó, and A. A. Benczúr. Latent Dirichlet Allocation in Web Spam Filtering. manuscript, 2008.

[3] I. Bíró, J. Szabó, and A. A. Benczúr. Very Large Scale Link Based Latent Dirichlet Allocation for Web Document Classification. manuscript, `http://www.ilab.sztaki.hu/~ibiro/linkedLDA/`, 2009.

[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. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, 2007.

[7] D. Cohn and T. Hofmann. The Missing Link-A Probabilistic Model of Document Content and Hypertext Connectivity. *Advances in Neural Information Processing Systems*, pages 430–436, 2001.

[8] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[9] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, pages 233–240. ACM Press New York, NY, USA, 2007.

[10] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications, 2004.

[11] 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.

[12] 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.

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

[14] 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.

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

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

[17] T. Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1):177–196, 2001.

[18] Z. Kou and W. W. Cohen. Stacked graphical models for efficient inference in markov random fields. In *SDM 07*, 2007.

[19] 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.

[20] R. Nallapati, A. Ahmed, E. Xing, and W. Cohen. Joint Latent Topic Models for Text and Citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM Press New York, NY, USA, 2008.

[21] 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.

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

[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] X. Zhu, J. Kandola, Z. Ghahramani, and J. Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. *Advances in Neural Information Processing Systems*, 17:1641–1648, 2005.