Blogger-Link-Topic Model for Blog Mining

Flora S. Tsai

Singapore University of Technology and Design,
Singapore 138682
fst1@columbia.edu

Abstract. Blog mining is an important area of behavior informatics because produces effective techniques for analyzing and understanding human behaviors from social media. In this paper, we propose the blogger-link-topic model for blog mining based on the multiple attributes of blog content, bloggers, and links. In addition, we present a unique blog classification framework that computes the normalized document-topic matrix, which is applied our model to retrieve the classification results. After comparing the results for blog classification on real-world blog data, we find that our blogger-link-topic model outperforms the other techniques in terms of overall precision and recall. This demonstrates that additional information contained in blog-specific attributes can help improve blog classification and retrieval results.

Keywords: Blog, blogger-link, classification, blog mining, author-topic, Latent Dirichlet Allocation.

1 Introduction

A blog is an online journal website where entries are made in reverse chronological order. The blogosphere is the collective term encompassing all blogs as a community or social network [19], which is an important area in behavior informatics [2]. Because of the huge volume of existing blog posts (documents), the information in the blogosphere is rather random and chaotic [17].

Previous studies on blog mining [7,13,15,18] use existing text mining techniques without consideration of the additional dimensions present in blogs. Because of this, the techniques are only able to analyze one or two dimensions of the blog data. On the other hand, general dimensionality reduction techniques [16,14] may not work as well in preserving the information present in blogs. In this paper, we propose unsupervised probabilistic models for mining the multiple dimensions present in blogs. The models are used in our novel blog classification framework, which categorizes blogs according to their most likely topic.

The paper is organized as follows. Section 2 describes related models and techniques, proposes new models for blog mining, and introduces a novel blog classification framework. Section 3 presents experimental results on real-world blog data, and Section 4 concludes the paper.

L. Cao et al. (Eds.): PAKDD 2011 Workshops, LNAI 7104, pp. 28–39, 2012.

2 Models for Blog Mining

In this section, we propose and apply probabilistic models for analyzing the multiple dimensions present in blog data. The models can easily be extended for different categories of multidimensional data, such as other types of social media.

Latent Dirichlet Allocation (LDA) [1] models text documents as mixtures of latent topics, where topics correspond to key concepts presented in the corpus. An extension of LDA to probabilistic Author-Topic (AT) modeling [11,12] is proposed for blog mining. The AT model generates a distribution over document topics that is a mixture of the distributions associated with the authors [11] . An author is chosen at random for each individual word in the document. This author chooses a topic from his or her multinomial distribution over topics, and then samples a word from the multinomial distribution over words associated with that topic. This process is repeated for all words in the document [12]. Other related work include a joint probabilistic document model (PHITS) [4] which modeled the contents and inter-connectivity of document collections. A mixed-membership model [5] was developed in which PLSA was replaced by LDA as the generative model. The Topic-Link LDA model [8] quantified the effect of topic similarity and community similarity to the formation of a link.

We have extended the AT model for blog links and dates. For the Link-Topic (LT) model, each link is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic. In the LT model, a document has a distribution over topics that is a mixture of the distributions associated with the links. When generating a document, a link is chosen at random for each individual word in the document. This link chooses a topic from his or her multinomial distribution over topics, and then samples a word from the multinomial distribution over words associated with that topic. This process is repeated for all words in the document.

For the LT model, the probability of generating a blog is given by:

$$\prod_{i=1}^{N_b} \frac{1}{L_b} \sum_{l} \sum_{t=1}^{K} \phi_{w_i t} \theta_{tl} \tag{1}$$

where there are N_b blogs, with each blog b having L_b links. The probability is then integrated over ϕ and θ and their Dirichlet distributions and sampled using Markov Chain Monte Carlo methods.

Likewise, for the Date-Topic (DT) model, each date is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic, and the representation of dates is performed the same way as the AT and LT model for authors and links. The probability of generating a blog is given by:

$$\prod_{i=1}^{N_b} \frac{1}{D_b} \sum_{d} \sum_{t=1}^{K} \phi_{w_i t} \theta_{td} \tag{2}$$

where there are N_b blogs, with each blog b having D_b dates. The probability is then integrated over ϕ and θ and their Dirichlet distributions and sampled using Markov Chain Monte Carlo methods.

2.1 Blogger-Link-Topic (BLT) Model

Although we have extended the Author-Topic (AT) model for blog links and dates, the existing techniques based on the AT model are not able to simultaneously analyze the multiple attributes of blog documents. In order to solve the problem of analyzing the multiple attributes of blogs, we propose the Blogger-Link-Topic (BLT) model that can solve the problem of finding the most likely bloggers and links for a given set of topics. Figure 1 shows the generative model of the BLT model using plate notation.

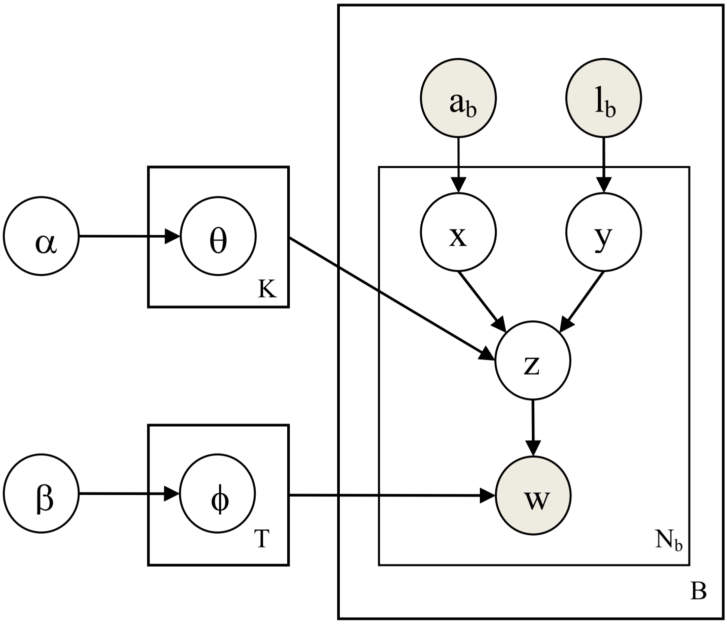

Fig. 1. The graphical model for the blogger-link-topic model using plate notation

In the Blogger-Link-Topic (BLT) model, each blogger and link is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic. For each word in the blog b, blogger x is chosen uniformly at random from a_b and link y is chosen uniformly at random from l_b. Then, a topic is chosen from a distribution over topics, θ, chosen from a symmetric Dirichlet(α) prior, and the word is generated from the chosen topic. The mixture weights corresponding to the chosen blogger are used to select a topic z, and a word is generated according to the distribution ϕ corresponding to that topic, drawn from a symmetric Dirichlet(β) prior.

We have also extended the BLT model for finding the most likely bloggers and dates for a given set of topics. For the Blogger-Date-Topic (BDT) model, each blogger and date is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic. Likewise, the model can also be extended to incorporate comments, social network of bloggers, inlinks, and tags.

For the Blogger-Comment-Topic (BCT) model, each blogger and comment is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic. For the social network of bloggers, the relationships between bloggers can be obtained either directly from the bloggers' list of friends (if available) or indirectly through inferring similar bloggers by the content they publish. Once the network is obtained, then the the Blogger-Network-Topic (BNT) model can be built, where each blogger and social network is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic. Similarly, incorporating inlinks into the model will need additional processing either through analyzing the outlinks of other blogs that point to the current blog, or by crawling through the link information in the blog network. Then, the Blogger-Inlink-Topic (BIT) model can be built, where each blogger and inlink is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic. Finally, the Blogger-Tag-Topic (BTT) model can be built from a list of either user-generated tags, subjects, or categories for the blog. In the BTT model, each blogger and tag is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic.

In order to learn the models for links and dates, we integrated the probability over ϕ and θ and their Dirichlet distributions and sampled them using Markov Chain Monte Carlo methods. Monte Carlo methods use repeated random sampling, which is necessary as our models rely on Bayesian inference to generate the results. Other ways to generate the results include variational methods and expectation propagation.

The BLT model solved the problem of finding the most likely bloggers and links for a given set of topics, a problem that is unique to blog mining. Although it is an extension of the AT model, the model is motivated by the unique structure of blog data and capitalizes on the synergy between bloggers and links. For example, if similar bloggers post similar content, or if similar contents are posted from the same links, the BLT model can model both situations and use this information to categorize the blogs into their correct topics. The BLT model is a better model than LDA and the AT model for the blog mining task because it is able to consider more attributes of blogs than just the blogger or content. The LDA model is limited because it only uses content information, while the AT model only uses blogger and content information, and does not use the additional information provided in links or dates.

In comparison to other models, BLT is different than Link-PLSA-LDA [10] and the mixed-membership model [5] because it does not use PLSA as the generative building block. In the joint probabilistic document model (PHITS) [4] and citation-topic (CT) model [6], the "links" were defined as the relations among documents rather than the actual URL, permalinks, or outlinks. Similarly, in the Topic-Link LDA model [8] defined the formation of a "link" as the effect of topic similarity and community similarity. Thus, our usage of links and the adaptation to blogs are some of the distinguishing characteristics of our model that sets it apart from the ones previously proposed.

The BLT is a general model can excel in blog mining tasks such as classification and topic distillation. The advantage of BLT over simple text classification techniques is that the simple methods will not be able to simultaneously leverage the multiple attributes of blogs. Other blog mining tasks include opinion retrieval, which involved locating blog posts that express an opinion about a given target. For the BLT model to be used in opinion retrieval, the model needs to be first trained on a set of positive, negative, and neutral blogs. Then BLT can be used to detect the most likely blogs which fit the labeled training set. Splog detection is another challenging task that can benefit from the BLT model. To adapt the BLT model to spam posts, trainings needs to be performed on a set of spam and non-spam blogs. After training, the BLT model can be used to leverage the information on bloggers and links to detect the most likely spam and non-spam blogs. Topic distillation, or feed search, searches for a blog feed with a principle, recurring interest in a topic. A subset of this task will be to search for a blog post in a feed related to a topic that the user wishes to subscribe. The BLT model can thus be used to determine the most likely bloggers and links for a given blog post of a given topic.

2.2 Blog Classification Framework

To demonstrate the usefulness of the BLT model, we propose a novel blog classification framework that is able to classify blogs into different topics. In this framework, blog documents are processed using stopword removal, stemming, normalization, resulting in the generation of the term-document matrix. Next, different techniques are implemented to create the document-topic matrix. From the document-topic matrix, the classification into different topics can be performed and evaluated against available relevance judgements.

The heuristics used to develop the blog classification framework can be described as follows:

1. Given the output of the BLT model as a term-topic matrix, a blogger-topic matrix, and a link-topic matrix, the first step is to convert the matrices to a document-topic matrix.
2. Since the document-topic matrix is not normalized, the second step is to normalize the matrices.

3. From the normalized document-topic matrix, the most likely topic for each document (blog) is compared against the ground truth.
4. The corresponding precision and recall can then be calculated.

In order to calculate the document-topic matrices for the various techniques, we propose the following technique. From the results of the Blogger-Link-Topic model, we can obtain a set of document-topic matrices, which can be used to predict the topic given a document. The document-topic matrix (DT_w) based on terms (words) is given by:

$$DT_w = WD' \times WT; \tag{3}$$

where WD is the term-document matrix, and WT is the term-topic matrix. The justification for the equation above is to convert the output of the BLT model WT, which lists the most probably terms for each topic, into a list of the most probable documents per topic, DT_w. Once we have a list of the most likely documents per topic, we can compare them with the ground truth to judge the precision and recall performance.

The document-topic matrix (DT_b) based on bloggers is given by:

$$DT_b = BD' \times BT; \tag{4}$$

where BD is the blogger-document matrix, and BT is the blogger-topic matrix. The justification for the equation above is to convert the output of the BLT model BT, which lists the most probably bloggers for each topic, into a list of the most probable documents per topic, DT_b.

Likewise, the document-topic matrix based on links (DT_l) is given by:

$$DT_l = LD' \times LT; \tag{5}$$

where LD is the link-document matrix, and LT is the link-topic matrix. The justification for the equation above is to convert the output of the BLT model LT, which lists the most probably links for each topic, into a list of the most probable documents per topic, DT_l.

The three document-topic matrices were used to create the normalized document-topic matrix (DT_{norm}) given by:

$$DT_{norm} = \frac{DT_w + DT_b + DT_l}{||DT_w||_{max} + ||DT_b||_{max} + ||DT_l||_{max}} \tag{6}$$

From DT_{norm}, we then predicted the corresponding blog category for each document, and compared the predicted results with the actual categories.

For the author-topic (AT) model, the normalized document-topic matrix DT_{norm} is given by:

$$DT_{norm} = \frac{DT_w + DT_b}{||DT_w||_{max} + ||DT_b||_{max}} \tag{7}$$

and for Latent Dirichlet Allocation (LDA), the normalized document-topic matrix DT_{norm} is just the normalized document-topic matrix (DT_w) based on terms (words):

$$DT_{norm} = \frac{DT_w}{||DT_w||_{max}} \tag{8}$$

The document-topic matrix can be efficiently generated by assuming the sparseness levels of the term-document matrix and the term-topic matrix. Fast sparse matrix multiplication can be performed by moving down both the row and column lists in tandem, searching for elements in the row list that have the same index as elements in the column list. Since the lists are kept in order by index, this can be performed in one scan through the lists. Each iteration through the loop moves forward at least one position in one of the lists, so the loop terminates after at most Z iterations (where Z is number of elements in the row list added to the number of elements in the column list).

Once we obtain the document-topic matrix, the category corresponding to the highest score for each blog document is used as the predicted category, and compared to the actual blog category. Given the normalized document-topic matrix DT and the ground truth topic vector G for each document, the predicted category vector C is calculated by finding the topic corresponding to the maximum value for each row vector $d_i = (w_{i1}, ..., w_{iX})$. Thus, the average precision and average recall can be calculated by averaging the results across all the topics.

3 Experiments and Results

Experiments were conducted on two datasets: BizBlogs07 [3], a data corpus of business blogs, and Blogs06 [9], the dataset used for TREC Blog Tracks from 2006–2008. BizBlogs07 contains 1269 business blog entries from various CEOs's blog sites and business blog sites. There are a total of 86 companies represented in the blog entries, and the blogs were classified into four (mutually exclusive) categories based on the contents or the main description of the blog: Product, Company, Marketing, and Finance [3]. Blogs in the Product category describe specific company products, such as reviews, descriptions, and other product-related news. Blogs in the Company category describe news or other information specific to corporations, organizations, or businesses. The Marketing category deals with marketing, sales, and advertising strategies for companies. Finally, blogs in the Finance category relates to financing, loans, credit information [3].

The Blogs06 collection [9] is a large dataset which is around 24.7 GB after being compressed for distribution. The number of feeds is over 100k blogs while the number of permalink documents is over 3.2 million. For the purpose of our experiments, we extracted a set of topics from the 2007 TREC Blog topic distillation task. The subset of topics were taken from the MySQL database created in [13], which indexed the feeds component. The 6 topics we extracted were 955 (mobile phone), 962 (baseball), 970 (hurricane Katrina), 971 (wine), 978 (music), and 983 (photography). The original number of RSS (Really Simple Syndication) blog items (which correspond to a document) for the 6 topics was 9738; however, after extracting those items with non-empty content, blogger, link, and date information, the total number of items reduced to 2363 documents.

3.1 Blogger-Link-Topic Results

The most likely terms and corresponding bloggers and links from the topics of marketing and finance of the BizBlogs07 collection are listed in Tables 1-2.

Table 1. BLT Topic 3: Marketing

Term	**Probability**
market	0.03141
compani	0.02313
busi	0.02057
time	0.01027
make	0.00831
Blogger	**Probability**
Chris Mercer	0.12553
FMF	0.12475
Larry Bodine	0.10897
john dodds	0.07022
Manoj Ranaweera	0.04233
Link	**Probability**
merceronvalue.com	0.13012
www.freemoneyfinance.com	0.12931
pm.typepad.com	0.11578
makemarketinghistory.blogspot.com	0.07267
jkontherun.blogs.com	0.05117

As can be seen from the tables, the topics correspond to the four categories of Product, Company, Marketing, and Finance, based on the list of most likely terms. The list of top bloggers correspond to the bloggers that are most likely to post in a particular topic. Usually a blogger with a high probability for a given topic may also post many blogs for that topic. The list of links correspond to the top links for each topic, based on the most likely terms and top bloggers that post to the topic.

3.2 Blog Classification Results

The results of the Blogger-Link-Topic (BLT) model for blog classification is compared with three other techniques: Blogger-Date-Topic (BDT), Author-Topic (AT), and Latent Dirichlet Allocation (LDA). The BizBlogs07 classification results for precision and recall were tabulated in Tables 3 and 4. The Blogs06 classification results are shown in Tables 5 and 6.

Table 2. BLT Topic 4: Finance

Term	**Probability**
monei	0.01754
year	0.01639
save	0.01576
make	0.01137
time	0.00895
Blogger	**Probability**
FMF	0.49493
Jeffrey Strain	0.06784
Tricia	0.05400
Chris Mercer	0.04487
Michael	0.04177
Link	**Probability**
www.freemoneyfinance.com	0.50812
www.pfadvice.com	0.06948
www.bloggingawaydebt.com	0.05526
merceronvalue.com	0.04588
jkontherun.blogs.com	0.04498

Table 3. Blog Precision Results for BizBlogs07

	Product	**Company**	**Marketing**	**Finance**	**Average**
BLT	*0.8760*	*0.4415*	*0.8996*	*0.9569*	**0.7935**
BDT	*0.6718*	*0.2151*	*0.4349*	*0.9540*	**0.5690**
AT	*0.8682*	*0.3434*	*0.8922*	*0.9626*	**0.7666**
LDA	*0.8760*	*0.4566*	*0.7026*	*0.8534*	**0.7222**

Table 4. Blog Recall Results for BizBlogs07

	Product	**Company**	**Marketing**	**Finance**	**Average**
BLT	*0.7829*	*0.6842*	*0.7576*	*0.8397*	**0.7661**
BDT	*0.8301*	*0.6129*	*0.5691*	*0.4510*	**0.6158**
AT	*0.8000*	*0.6741*	*0.7533*	*0.7512*	**0.7447**
LDA	*0.7655*	*0.5789*	*0.5555*	*0.8652*	**0.6913**

Table 5. Blog Precision Results for Blogs06

	Mobile	**Baseball**	**Katrina**	**Wine**	**Music**	**Photo**	**Average**
BLT	*0.8549*	*0.9159*	*0.7083*	*0.8418*	*0.4961*	*0.5512*	**0.7281**
BDT	*0.9604*	*0.8097*	*0.8095*	*0.6461*	*0.5741*	*0.0271*	**0.6378**
AT	*0.8264*	*0.9867*	*0.3125*	*0.8338*	*0.7566*	*0.4995*	**0.6861**
LDA	*0.8484*	*0.9425*	*0.7262*	*0.6997*	*0.5959*	*0.1175*	**0.6550**

The BLT model obtained better overall precision and recall results for both BizBlogs07 and Blogs06 data. For Blogs06, the worst performing category was Photography, due to the general nature of the topic. The Baseball category had the highest overall precision, while the Wine category had the highest overall

Table 6. Blog Recall Results for Blogs06

	Mobile	**Baseball**	**Katrina**	**Wine**	**Music**	**Photo**	**Average**
BLT	*0.6419*	*0.8449*	*0.6247*	*0.9874*	*0.9034*	*0.3970*	**0.7332**
BDT	*0.4771*	*0.7821*	*0.5540*	*0.9060*	*0.9583*	*0.1250*	**0.6337**
AT	*0.7054*	*0.6778*	*0.5526*	*0.9094*	*0.6315*	*0.6617*	**0.6897**
LDA	*0.5554*	*0.4863*	*0.7011*	*0.8502*	*0.8527*	*0.3071*	**0.6255**

recall. The reasons could also be due to the specialized nature of both categories. For both of the datasets, the best to worst performing models were the same, with BLT achieving the best results overall, followed by AT, LDA, and BDT. This confirms our theory that blogger and link information can help to improve results for the blog mining classification task.

3.3 Results on Co-occurrence of BLT and AT Models

As BLT is based on the AT model, the two models will co-occur if all the links are identical. That is, if all the blog data come from the same link, which we define as the root domain. This could conceivably occur if all the blogs were obtained from the same source. In the case of identical links, the output of the BLT model will be identical to the AT model. We confirm this in the following set of experiments for the 2007 TREC Blog data, where all the links were changed to the same name: "www.goobile.com". The output of the first two topics in the BLT model is shown in Tables 7–8. In the tables, the topic numbers correspond to the numbers in the 2007 TREC Blog Track. The output of the AT model is exactly the same as the BLT model, except for the link information which is not included in the AT model. However, the case shown is an exception rather than rule, as it is highly unlikely that all blogs will have the same source. Therefore, the BLT model is still useful to analyze the additional link information not provided in the AT model.

Table 7. BLT Topic 955: Mobile Phone

Term	**Probability**
mobil	0.02008
phone	0.01816
servic	0.00865
camera	0.00828
imag	0.00790
Blogger	**Probability**
administrator	0.07436
alanreiter	0.06120
dennis	0.05517
admin	0.05174
data	0.05124
Link	**Probability**
www.goobile.com	1.0000

Table 8. BLT Topic 962: Baseball

Term	**Probability**
year	0.02003
game	0.01229
time	0.01169
plai	0.01069
basebal	0.01066
Blogger	**Probability**
bill	0.19567
scott	0.07846
yardwork	0.04825
dave	0.04246
scottr	0.03421
Link	**Probability**
www.goobile.com	1.0000

4 Conclusion

In this paper, we have proposed a probabilistic blogger-link-topic (BLT) model based on the author-topic model to solve the problem of finding the most likely bloggers and links for a given set of topics. The BLT model is useful for behavior informatics, which can help extract discriminative behavior patterns from high-dimensional blog data. The BLT results for blog classification were compared to other techniques using blogger-date-topic (BDT), author-topic, and Latent Dirichlet Allocation, with BLT obtaining the highest average precision and recall.

The BLT model can easily be extended to other areas of behavior informatics, such as to analyze customer demographic and transactional data, human behavior patterns and impacts on businesses. In addition, the BLT model can help in the analysis of behavior social networks handling convergence and divergence of behavior, and the evolution and emergence of hidden groups and communities.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
2. Cao, L.: In-depth behavior understanding and use: the behavior informatics approach. Information Science 180, 3067–3085 (2010)
3. Chen, Y., Tsai, F.S., Chan, K.L.: Machine learning techniques for business blog search and mining. Expert Syst. Appl. 35(3), 581–590 (2008)
4. Cohn, D., Hofmann, T.: The missing link – a probabilistic model of document content and hypertext connectivity. In: Advances in Neural Information Processing Systems, vol. 13, pp. 430–436 (2001)
5. Erosheva, E., Fienberg, S., Lafferty, J.: Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5220–5227 (2004)

6. Guo, Z., Zhu, S., Chi, Y., Zhang, Z., Gong, Y.: A latent topic model for linked documents. In: SIGIR 2009: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 720–721. ACM, New York (2009)
7. Liang, H., Tsai, F.S., Kwee, A.T.: Detecting novel business blogs. In: ICICS 2009: Proceedings of the 7th International Conference on Information, Communications and Signal Processing (2009)
8. Liu, Y., Niculescu-Mizil, A., Gryc, W.: Topic-link lda: joint models of topic and author community. In: ICML 2009: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 665–672. ACM, New York (2009)
9. Macdonald, C., Ounis, I.: The TREC Blogs06 collection: Creating and analysing a blog test collection. Tech. rep., Dept of Computing Science, University of Glasgow (2006)
10. Nallapati, R., Cohen, W.: Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM). Association for the Advancement of Artificial Intelligence (2008)
11. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: AUAI 2004: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 487–494. AUAI Press, Arlington (2004)
12. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: KDD 2004: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 306–315. ACM, New York (2004)
13. Tsai, F.S.: A data-centric approach to feed search in blogs. International Journal of Web Engineering and Technology (2012)
14. Tsai, F.S.: Dimensionality reduction techniques for blog visualization. Expert Systems With Applications 38(3), 2766–2773 (2011)
15. Tsai, F.S., Chan, K.L.: Detecting Cyber Security Threats in Weblogs using Probabilistic Models. In: Yang, C.C., Zeng, D., Chau, M., Chang, K., Yang, Q., Cheng, X., Wang, J., Wang, F.-Y., Chen, H. (eds.) PAISI 2007. LNCS, vol. 4430, pp. 46–57. Springer, Heidelberg (2007)
16. Tsai, F.S., Chan, K.L.: Dimensionality reduction techniques for data exploration. In: 2007 6th International Conference on Information, Communications and Signal Processing, ICICS, pp. 1568–1572 (2007)
17. Tsai, F.S., Chan, K.L.: Redundancy and novelty mining in the business blogosphere. The Learning Organization 17(6), 490–499 (2010)
18. Tsai, F.S., Chen, Y., Chan, K.L.: Probabilistic Techniques for Corporate Blog Mining. In: Washio, T., Zhou, Z.-H., Huang, J.Z., Hu, X., Li, J., Xie, C., He, J., Zou, D., Li, K.-C., Freire, M.M. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4819, pp. 35–44. Springer, Heidelberg (2007)
19. Tsai, F.S., Han, W., Xu, J., Chua, H.C.: Design and Development of a Mobile Peer-to-Peer Social Networking Application. Expert Syst. Appl. 36(8), 11077–11087 (2009)