LDA-Based Topic Modeling in Labeling Blog Posts with Wikipedia Entries

Daisuke Yokomoto[1], Kensaku Makita[1], Hiroko Suzuki[1], Daichi Koike[1], Takehito Utsuro[1], Yasuhide Kawada[2], and Tomohiro Fukuhara[3]

[1] University of Tsukuba, Tsukuba, 305-8573, Japan
[2] Navix Co., Ltd., Tokyo, 141-0031, Japan
[3] National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan

Abstract. Given a search query, most existing search engines simply return a ranked list of search results. However, it is often the case that those search result documents consist of a mixture of documents that are closely related to various contents. In order to address the issue of quickly overviewing the distribution of contents, this paper proposes a framework of labeling blog posts with Wikipedia entries through LDA (latent Dirichlet allocation) based topic modeling. More specifically, this paper applies an LDA-based document model to the task of labelling blog posts with Wikipedia entries. One of the most important advantages of this LDA-based document model is that the collected Wikipedia entries and their LDA parameters heavily depend on the distribution of keywords across all the search result of blog posts. This tendency actually contributes to quickly overviewing the search result of blog posts through the LDA-based topic distribution. In the evaluation of the paper, we also show that the LDA-based document retrieval scheme outperforms our previous approach.

Keywords: Blog, Wikipedia, Topic Model, LDA, Topic Analysis.

1 Introduction

As blog services and blog tools are becoming more and more popular, people have been able to express one's own interests as well as opinions on the Web. Search engines are then used for accessing various information that can be found in the blogosphere, where, given a search query, a ranked list of blog posts is provided as a search result. However, such a search result in the form of a ranked list is usually not helpful for a user to quickly identify blog posts that satisfy his/her information need. This is especially true when, given a search query, the search result is a mixture of blog posts that focus on various contents.

In such a situation, the framework of *faceted search* [1], which has been well studied in the information retrieval community, can be a solution. Based on this observation, [2] proposed a framework of categorizing Japanese blog posts according to their contents, where, given a search query, those blog posts are

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 114–124, 2012.

collected from the Japanese blogosphere. In this framework, the content of each blog post is regarded as a facet of a query keyword, and a facet is automatically assigned to each blog post. This procedure of assigning a facet to a blog post is realized by utilizing Wikipedia entries as a knowledge source and each Wikipedia entry title is considered as a facet label. In its Japanese version, about 770,000 entries are included (checked at November, 2011). Given a query keyword, Wikipedia entries are collected from Wikipedia as candidates of its facets. Then, for each Wikipedia entry, its body text is analyzed as fundamental knowledge source, and terms strongly related to the entry are extracted. Those terms are then used for labelling a blog post with this entry.

One drawback of the framework of [2] is that, when labelling a blog post with Wikipedia entries, it does not exploit the distribution of keywords across all the search result of blog posts, but labels a blog post with Wikipedia entries independently of other blog posts in the whole search result. This is especially disadvantageous considering the purpose of the research, i.e., to quickly overview the search result of blog posts in terms of their contents. In order to overcome this disadvantage of [2], this paper proposes a framework of labeling blog posts with Wikipedia entries through LDA (latent Dirichlet allocation [3]) based topic modeling. More specifically, this paper applies an LDA-based document model [4] to the task of labelling blog posts with Wikipedia entries. Figure 1 illustrates the overall framework of labelling blog posts with Wikipedia entries based on an LDA-based document model (with a query keyword *"global warming"*). In this framework, first, given a query keyword, blog posts that are related to the query keyword are collected. Then, from the collected blog posts, Wikipedia entry titles are extracted. Next, the LDA parameters are estimated with the extracted Wikipedia entries, where the topics that are closely related to the collected blog posts are generated. Those LDA parameters for generated topics are also incorporated into the LDA-based document retrieval scheme [4]. When applying an LDA-based document model to the task of labelling blog posts with Wikipedia entries, we regard each blog post as a query, and the collected Wikipedia entries as the document collection from which one or more documents are retrieved.

One of the most important advantages of this LDA-based document model is that the collected Wikipedia entries and their LDA parameters heavily depend on the distribution of keywords across all the search result of blog posts. This tendency actually contributes to quickly overviewing the search result of blog posts through the LDA-based topic distribution. In the evaluation of the paper, we also show that the LDA-based document retrieval scheme outperforms our previous approach of [2].

2 Related Works

In TREC 2009 blog track [5], faceted blog distillation task was studied, where three facets, namely, *opinionated/personal/in-depth* are introduced and participants are required to assign facets to blog feeds. [6] invented a multi-faceted

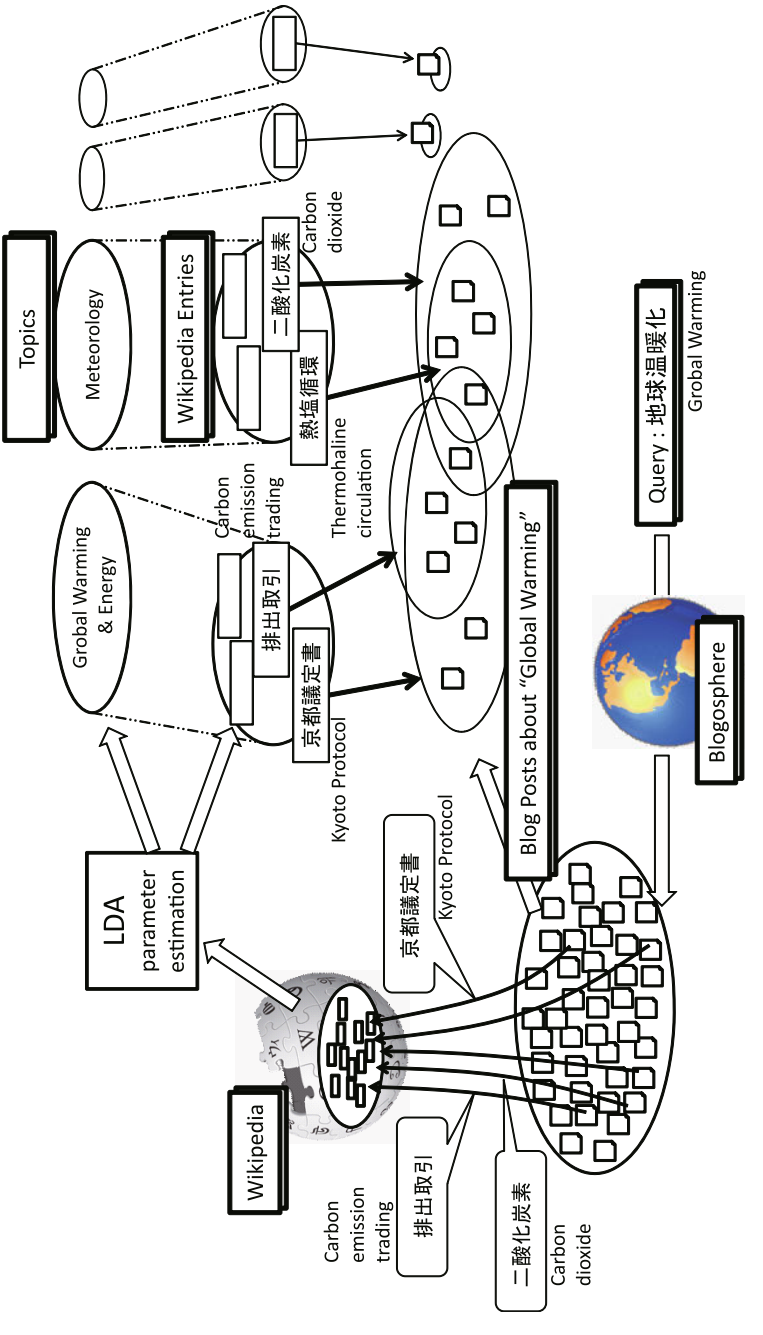

Fig. 1. Framework of Labeling Blog Posts with Wikipedia Entries through LDA-based Topic Modeling

blog search framework, where various facets are introduced in terms of topics, bloggers, links, and sentiments. [7] also proposed a framework of generating a faceted search interface for Wikipedia. Compared to those previous works [7,6,5], the proposed method is innovative in that it realizes a novel technique of automatically generating sub-topic oriented facets for blog posts collected from the blogosphere. Our work is related to [7] in that both techniques collect facet candidates from Wikipedia. In [7], it is also presented how to rank facet hierarchies, where the cost of navigation through Wikipedia facet hierarchies is modeled and is utilized in facet hierarchy ranking. However, compared to our technique, that of [7] modeled the cost of navigation only within the Wikipedia facet hierarchy, where the target of navigation is also Wikipedia articles. Our technique is different from that of [7] in that, in our technique, given the set of blog posts collected with an initial query as the target of navigation, facet candidates that are not frequently observed in the collected blog posts are removed. As a future work, it is also possible to introduce the formalization of the navigation cost of [7] into our task.

Another related works include techniques of clustering and summarizing search results [8], or those of clustering search results and assigning cluster labels [9,10,11]. Compared with those techniques on search results clustering, the proposed technique is advantageous in that it is capable of assigning facets to even quite a small number of blog posts, simply because it utilizes Wikipedia as a knowledge source for assigning facets to blog posts.

The technique presented in this paper is also related to previous works on assigning Wikipedia concepts to document clusters (e.g., [12]) and those combining Wikipedia concepts as well as important terms extracted from the cluster content in cluster labeling (e.g., [13]). However, those previous works mostly concentrate on clustering standard document sets such as those of newspaper articles with broad range of topics. In this paper, on the other hand, we focus on extracting facets from Wikipedia, given the set of blog posts collected with an initial query, where the collected blog posts cover much narrower range of contents. Compared with the tasks studied in those previous works, the task of facet categorization of related blog posts studied in this paper is relatively difficult to tackle.

With respect to related works of the LDA-based document retrieval scheme [4], [4] argued that the LDA-based document retrieval scheme overcomes the disadvantages of the pLSI model [14] as well as the cluster-based language model [15].

3 Retrieving Blog Posts with a Query Keyword

Given a query keyword t_0, this section describes how to retrieve blog posts with t_0 as a search query. With this procedure, we intend to collect candidates of blog posts that are closely related to t_0.

First, we use an existing Web search engine API, which returns a ranked list of blog posts, given a topic keyword. For the evaluation in section 6, during the period from July to September, 2010, we used the Japanese search engine

"Yahoo! Japan" API[1] for Japanese. Blog hosts are limited to major 8 hosts[2] for Japanese. For each query, this search engine API returns a ranked list of at most 1,000 blog posts. A list of blog feeds is then generated from the returned ranked list of blog posts by simply removing duplicated feeds. From the retrieved blog feeds, we next collect blog posts that include the initial topic keyword t_0 in the body text into the set $BP(t_0)$ of blog posts as candidates for those closely related to t_0.

4 Collecting Wikipedia Entry Titles from Blog Posts

From the set $BP(t_0)$ of collected blog posts, we collect Wikipedia entry titles and construct the set $\mathbb{E}(BP(t_0))$ of Wikipedia entries. Here, we require that the entry e to be collected satisfy that the document frequency $\mathrm{df}(BP(t_0), t(E))$ of the title $t(E)$ of e over the set of collected blog posts $BP(t_0)$ is more than or equal to 10[3].

$$\mathbb{E}(BP(t_0)) = \left\{ E \;\middle|\; \mathrm{df}(\ BP(t_0),\ t(E)\ )\ \geq 10 \right\}$$

5 LDA-Based Document Model

5.1 Latent Dirichlet Allocation

LDA [3] can be used to model and discover underlying topic structures of discrete data such as text. The formalization of LDA below follows the notation of quantities below:

- M: the total number of documents
- K: the number of topics
- V: vocabulary size
- $\boldsymbol{\alpha}, \boldsymbol{\beta}$: Dirichlet parameters
- $\boldsymbol{\vartheta}_m$: topic distribution for document m
- $\boldsymbol{\Theta} = \{\boldsymbol{\vartheta}_m\}_{m=1}^{M}$: a $M \times K$ matrix
- $\boldsymbol{\varphi}_k$: word distribution for topic k
- $\boldsymbol{\Phi} = \{\boldsymbol{\varphi_k}\}_{k=1}^{K}$: a $K \times V$ matrix
- N_m: the length of document m
- $z_{m,n}$: topic index of n-th word in document m
- $w_{m,n}$: a particular word for word placeholder $[m, n]$

Rough description of the generation process for LDA is as follows:

1. For each topic $k \in [1, K]$, pick a multinomial distribution $\boldsymbol{\varphi}_k$ from a Dirichlet distribution with parameter $\boldsymbol{\beta}$;

[1] `http://www.yahoo.co.jp/` (in Japanese).

[2] `fc2.com, yahoo.co.jp, yaplog.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa. net, hatena.ne.jp`

[3] We empirically chose this lower bound through preliminary evaluation.

2. For each document $\boldsymbol{w}_m = \{w_{m,n}\}_{n=1}^{N_m}$, pick a multinomial distribution $\boldsymbol{\vartheta}_m$ from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$,
3. For each word placeholder $[m, n]$, pick a topic $z_{m,n}$ from the multinomial distribution $\boldsymbol{\vartheta}_m$,
4. Pick word $w_{m,n}$ for the word placeholder $[m, n]$ from the multinomial distribution $\boldsymbol{\varphi}_{z_{m,n}}$.

Based on the above description, we can write the joint distribution of all known and hidden variables given the Dirichlet parameters as follows:

$$P(\boldsymbol{w_m}, \boldsymbol{z_m}, \boldsymbol{\vartheta_m}, \Phi \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\boldsymbol{\Phi} \mid \boldsymbol{\beta}) \prod_{n=1}^{N_m} P(w_{m,n} \mid \boldsymbol{\varphi_{z_{m,n}}}) P(z_{m,n} \mid \boldsymbol{\vartheta_m}) P(\boldsymbol{\vartheta_m} \mid \boldsymbol{\alpha})$$

And the likelihood of a document $\boldsymbol{w_m}$ is obtained by integrating over $\boldsymbol{\vartheta_m}$, $\boldsymbol{\Phi}$ and summing over $\boldsymbol{z_m}$ as follows:

$$P(\boldsymbol{w_m} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \int P(\boldsymbol{\vartheta_m} \mid \boldsymbol{\alpha}) P(\boldsymbol{\Phi} \mid \boldsymbol{\beta}) \cdot \prod_{n=1}^{N_m} P(w_{m,n} \mid \boldsymbol{\vartheta_m}, \boldsymbol{\Phi}) \, d\boldsymbol{\Phi} d\boldsymbol{\vartheta}_m$$

Finally, the likelihood of the whole data collection $\mathcal{W} = \{\boldsymbol{w}_m\}_{m=1}^{M}$ is product of the likelihoods of all documents:

$$P(\mathcal{W} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{m=1}^{M} P(\boldsymbol{w}_m \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$$

In the evaluation of section 6, we used GibbsLDA++ [16] for LDA parameter estimation, where for the number of topics $K = 50$, $\alpha = 50/K$, and $\beta = 0.1$.

5.2 LDA-Based Retrieval of Wikipedia Entries

In our LDA-based framework of retrieving Wikipedia entries given a blog post as a query, let us denote a blog post used as a query as $B \in BP(t_0)$. Also, we denote a Wikipedia entry to be retrieved and considered as a document model as E, where each Wikipedia entry E is taken from the set $\mathbb{E}(BP(t_0))$ of Wikipedia entries collected from the set $BP(t_0)$ of blog posts as described in section 4.

The basic approach for using language models for IR is the query likelihood model where each document is scored by the likelihood of its model generating a query B,

$$P(B \mid E) = \prod_{w \in B} P(w \mid E)$$

where E is a document model, B is the query and w is a query term in B[4]. $P(B \mid E)$ is the likelihood of the document model generating the query terms under

[4] More specifically, we require a query term w in B to be the title of a Wikipedia entry, and also require that the term frequency $freq(B, t(E))$ of the title $t(E)$ of E within B is more than or equal to 3.

the bog-of-words assumption that terms are independent given the documents. Following [4], we specify $P(w \mid E)$ by combining the LDA model $P_{lda}(w \mid E)$ with the traditional linear interpolation of the maximum likelihood estimate $P_{ML}(w \mid E)$ of word w in the document E and the maximum likelihood estimate $P_{ML}(w \mid \mathbb{E}(BP(t_0)))$ of word w in the entire collection $\mathbb{E}(BP(t_0))$ as below:[5]

$$P(w \mid E) = \lambda\Big\{\mu P_{ML}(w \mid E) + (1-\mu)P_{ML}(w \mid \mathbb{E}(BP(t_0)))\Big\} + (1-\lambda)P_{lda}(w \mid E)$$

where the LDA model $P_{lda}(w \mid E)$ is given as:

$$P_{lda}(w \mid E) = \sum_{k=1}^{K} P(w \mid \boldsymbol{\varphi_k})P(z_k \mid E)$$

6 Evaluation

6.1 Evaluation of Wikipedia Entry Ranking

For evaluation, we pick up the 9 query keywords listed in Table 1. For each query keyword t_0, Table 1 also shows the number of collected blog posts $|BP(t_0)|$. For each query keyword, we select two or three blog posts that have a high similarity value with the query keyword as the title of a Wikipedia entry, where the similarity is measured as introduced in [2]. In total, we select 20 blog posts for evaluation.

Table 1. 9 Query Keywords and the Size of Blog Posts for Evaluation

Query Keyword t_0	# of Blog Posts $\lvert BP(t_0)\rvert$
smoking	9,926
organ transplantation	1,502
global warming	8,687
medical error	1,914
population aging	2,205
Toyota Prius	5,099
smartphone	10,039
alcoholism	2,060
restructuring	5,007

For each of the 20 blog posts for evaluation, we compare the following three methods for ranking Wikipedia entries in terms of labeling the blog post:

[5] When combining $P_{ML}(w \mid E)$ and $P_{ML}(w \mid \mathbb{E}(BP(t_0)))$, we compared the Dirichlet smoothing as employed in [4] with the linear linear interpolation shown here, where the best performance with $\lambda = \mu = 0.7$ outperformed that of the Dirichlet smoothing employed in [4].

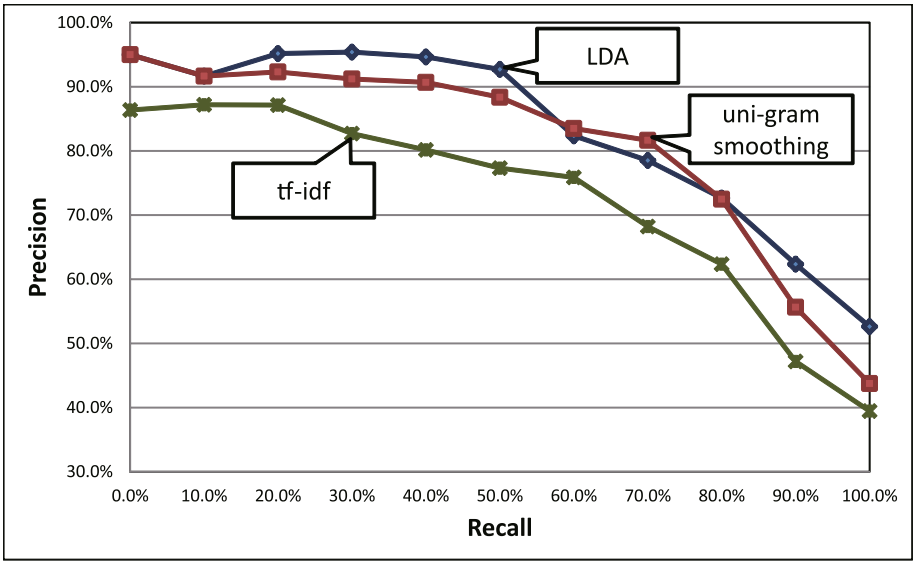

Fig. 2. Evaluation Results of Wikipedia Entry Ranking

1. the LDA-based retrieval model presented in section 5.2 (denoted as "LDA").
2. the same as above, except that we specify $P(w \mid E)$ as the linear interpolation of the maximum likelihood estimate $P_{ML}(w \mid E)$ of word w in the Wikipedia entry E and the maximum likelihood estimate $P_{ML}(w \mid \mathbb{E}(BP(t_0)))$ of word w in the entire collection $\mathbb{E}(BP(t_0))$:

$$P(w \mid E) \;=\; \mu P_{ML}(w \mid E) \;+\; (1-\mu) P_{ML}(w \mid \mathbb{E}(BP(t_0)))$$

   where $\mu = 0.7$ (denoted as "uni-gram smoothing").
3. Wikipedia entries are ranked according to the similarity value with the blog post as a query, where, as introduced in [2], the similarity is measured as the inner product of the term frequency vector of the blog post as a query and the inverse document frequency (within the whole Japanese version of Wikipedia) vector of a Wikipedia entry (denoted as "tf-idf").

For those 20 blog posts for evaluation, 60.1 Wikipedia entries are ranked on the average, out of which 12.1 entries are manually judged as correct labels for the query blog post. For the three methods "LDA", "uni-gram smoothing", and "tf-idf", Figure 2 plots the recall-precision curves that are averaged over the 20 blog posts for evaluation[6]. As can be clearly seen from this result, "LDA" constantly outperforms "tf-idf", where their differences are statistically significant at more than half of the 11 recall points at a level of 0.03 in terms of micro average. Considering the result that "LDA" and "uni-gram smoothing" are mostly

[6] For each query blog post, precision at every 11 recall point 0%, 10%, ..., 90%, 100% is measured and averaged over the 20 blog posts.

Table 2. An Example of LDA-based Topic Modeling of Wikipedia Entries

Topic ID		*T1*: global warming and energy	*T2*: meteorology	*T3*: astronomy	*T4*: politics
blog post	labeled Wikipedia entries	—	carbon dioxide, thermohaline circulation	sun, sunspot	—
B1	summary	—	Carbon dioxide does not cause global warming.	Not global warming but global cooling due to sunspot.	—
blog post	labeled Wikipedia entries	fossil fuel, alternative energy	—	—	—
B2	summary	Raise the price of fossil fuel to stop global warming.	—	—	—
blog post	labeled Wikipedia entries	fuel and light expenses, Photovoltaic power generation	—	—	Democratic Party of Japan, manifesto
B3	summary	Fuel and light expenses will greatly increase by introducing Photovoltaic power generation.	—	—	It is not beneficial for Japan to keep the manifesto of the Democratic Party of Japan.

comparative in their performance, the difference between "LDA" and "tf-idf" is mainly due to whether or not considering the distribution of keywords across all the search result of blog posts.

6.2 An Example of LDA-Based Topic Modeling of Wikipedia Entries

For the query keyword *"global warming"*, Table 2 shows topics z_k which have the highest probability value $P(z_k \mid E)$ for at least one Wikipedia entry E which is manually judged as correct. As shown in Table 2, in this case, we have four topics, out of which the one with the topic ID = *"T1: global warming and energy"*: is

most closely related to *"global warming"*, while other topics have rather little relation to *"global warming"*. As can be clearly seen in Table 2, those topics greatly contribute to quickly understanding the contents of blog posts.

We examined three blog posts as the query, where each blog post has one or two corresponding topics as in Table 2. Those topics are then allocated with a Wikipedia entry E which has the highest probability value $P(z_k \mid E)$, where the probability value should be sufficiently large. With this result, it becomes becomes much easier to quickly overview the distribution of topics over the query blog posts.

7 Conclusion

This paper proposed a framework of labeling blog posts with Wikipedia entries through LDA-based topic modeling. More specifically, this paper applied an LDA-based document model to the task of labelling blog posts with Wikipedia entries. One of the most important advantages of this LDA-based document model is that the collected Wikipedia entries and their LDA parameters heavily depend on the distribution of keywords across all the search result of blog posts. This tendency actually contributed to quickly overviewing the search result of blog posts through the LDA-based topic distribution. In the evaluation of the paper, we also showed that the LDA-based document retrieval scheme outperformed our previous approach.

References

1. Tunkelang, D.: Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers (2009)
2. Yokomoto, D., Makita, K., Utsuro, T., Kawada, Y., Fukuhara, T.: Utilizing Wikipedia in categorizing topic related blogs into facets. In: Proc. 12th PACLING, #20 (2011)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
4. Wei, X., Croft, W.B.: LDA-Based document models for ad-hoc retrieval. In: Proc. 29th SIGIR, pp. 178–185 (2006)
5. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the TREC-2009 blog track. In: Proc. TREC 2009 (2009)
6. Fujimura, K., Toda, H., Inoue, T., Hiroshima, N., Kataoka, R., Sugizaki, M.: BLOGRANGER - a multi-faceted blog search engine. In: Proc. 3rd Ann. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2006)
7. Li, C., Yan, N., Roy, S.B., Lisham, L., Das, G.: Facetedpedia: Dynamic generation of query-dependent faceted interfaces for Wikipedia. In: Proc. 19th WWW, pp. 651–660 (2010)
8. Harashima, J., Kurohashi, S.: Summarizing search results using PLSI. In: Proc. 2nd Workshop on NLPIX, pp. 12–20 (2010)
9. Toda, H., Kataoka, R., Oku, M.: Search result clustering using informatively named entities. International Journal of Human-Computer Interaction, 3–23 (2007)

10. de Winter, W., de Rijke, M.: Identifying facets in query-biased sets of blog posts. In: Proc. ICWSM, pp. 251–254 (2007)
11. Shibata, T., Bamba, Y., Shinzato, K., Kurohashi, S.: Web information organization using keyword distillation based clustering. In: Proc. WI-IAT, pp. 325–330 (2009)
12. Hu, J., Fang, L., Cao, Y., Zeng, H.J., Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging Wikipedia semantics. In: Proc. 31st SIGIR, pp. 179–186 (2008)
13. Carmel, D., Roitman, H., Zwerdling, N.: Enhancing cluster labeling using Wikipedia. In: Proc. 32nd SIGIR, pp. 139–146 (2009)
14. Hoffman, T.: Probabilistic latent semantic indexing. In: Proc. 22nd SIGIR, pp. 50–57 (1999)
15. Liu, X., Croft, W.B.: Cluster-based retrieval using language models. In: Proc. 27th SIGIR, pp. 186–193 (2004)
16. Phan, X.H., Nguyen, C.T.: GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA) (2007)