2018 IEEE International Conference on Software Quality, Reliability and Security Companion

Tag Recommendation for Cultural Resources

Zhiwen Lei
Institute of Automation
Chinese Academy of Science
Beijing, China
leizhiwen2016@ia.ac.cn

Yi Yang
Institute of Automation
Chinese Academy of Sciences
Beijing, China
yangyi@ia.ac.cn

Weixing Huang
Institute of Automation
Chinese Academy of Sciences
Beijing, China
weixing.huang@ia.ac.cn

Jian Wang*
Institute of Automation
Chinese Academy of Sciences
Beijing, China
jian.wang@ia.ac.cn

*Corresponding author: Jian Wang

***Abstract*—We propose a new tag recommendation method for the public digit cultural resources we collected in related program. We use LDA and Word2Vec to preprocess the resource and tag, map them to different vector space respectively. After that, we calculate the relevance of each resource and all tags by using Deep Structured Semantic Model. Lastly, range the result of calculation and treat the top N tags as the extended tags of a resource. Result show the good performance of our tag recommendation method.**

Keywords—tag recommendation; LDA; Word2Vec; DSSM

I. INTRODUCTION

A critical technology in the current Internet is how to combine the tag and the object, since the tag can bring stimulus to the user's content creation, make the creation easier. The rapid growth of the original contents and the difficulty in collecting data has brought many challenges to information retrieval, it need new methods to face these challenges. Tags or other textual features (say, title or label) usually can represent the specific content or part of it, these descriptors always provide support to content about organization, spread and retrieval service. In fact, the current shows that tags are best textual features which can be used in different information retrieval services, such as in the application of automatic object classification.By allocating descriptors to documents and other information resources, the annotation process can categorize and classify the data, these descriptors, or "tags" are generally relatively short text, can provide a simple description of the classification, information retrieval and browsing method. The keyword-based annotation of a text is only according to the text itself, this method cannot generate new information. Therefore, method that can generate new information become very popular in recent years, the main differences of these new methods and traditional keyword-based method are they always have open dictionary and nonlinear structure. The truth is, tags are always made by user or author of the information, rather than professional annotator, so there are no standards and limitations.

In this paper, for the purpose of solving the problem of the lack of available tags, we use tag recommendation method to extend the number of tags. The data we used is collected by the public digital cultural platform designed by the support project, metadata is some culture related words or short text, the kind of resource data including cultural institutions related data, encyclopedic information related data and network media related data, these three kinds of data are always free and easily accepted. So we negotiated with the site manager, acquire the permission and collect the data. Meanwhile, we also use web crawler to acquire some free and open data, finally all these data became cultural related data. The main contents of these data are: news related data, video related data, social related data, public cultural institutions related data, and other public culture related data. In actual operation, we selected some articles under the cultural categories which are in some well-known networks and some collections of famous museum online website. In the process of data collection, we also acquire some tags of these cultural objects. On the basis of these initial tags, we can organize the correspondence of resources and tags, after organization, we use the following two method to process resources and tags respectively:

For the resource, according to the distribution of the tag, we use LDA model to generate frequency co-occurrence matrix, each row of this matrix is topic vector and represents the corresponding resource.

For the tag, we use the word2vec model to preprocess it. The training data of word2vec comes from all the entries of

978-1-5386-7839-8/18/$31.00 ©2018 IEEE
DOI 10.1109/QRS-C.2018.00100

Baidupedia, by training all these collected entries, the final word2vec model can map as much as possible of all the Chinese words to a same vector space, and generate vector of these entries, then we find the vector of tags used in this paper and use it represent the corresponding tag.

By using the above method to preprocess the resources and tags, resources and tags were mapped to different vector space. Then we construct the resource-tag pairs with the representation of vector according to the corresponding relationship between resource and tag. In the next step, we build a DSSM and train this model by using these resource-tag pairs, after training, we treat each resource vector and all of tags vector as model input, calculate the similarity of resource and each tag, then we sort the result of calculation, the ranked top N tags of result was treated as extended tags about this resource. We successfully extend the number of tag of all the resources and, performance of extension is considerable good when we observed by using some indexes.

The rest of this paper is organized as following, first we reviewed the related works in section II, and then we talked about the model and algorithm in section III. After that, we presented our experimental evaluation and result in section IV and section V respectively. Section VI is the conclusion and future work.

II. Related Work

The tag is always very helpful for the resource, it can be used to classify and organize the data. Automatic tagging technique is widely discussed by many researchers. [4] build a kind of Tag Suggestion System which can utilize tags that related to the result of queries to expand the content of queries. [3] design a tool named AutoTag which recommend tags to blogs in using Collaborative filtering, work with the quality control process provided by bloggers to simplify the process of tagging and improve the quality of it. Blogs play a role of users, the tag assigned to blog is just like the item interested by user. Chirita et. al. propose a technique named P-TAG in [12] to generate personalized tags for web pages, when people browse web pages, P-TAG provide keywords that related to page textual content and desktop data of browsers, show personalized perspective in this way. Belém et. al. propose a novel method in [2] to recommend tag to the target object, in consideration of the following three elements: the word relevance of initial tag about target object, the word extracted from Multiple text feature and different measures of tag similarity. [2] also use heuristic method that can add new measurement to ready-made strategy, trying to some candidate word to describe the accuracy of the content about target object. Huang et. al. set up a novel deep structured semantic model that can map queries and results to a low dimensional space where the relevance of query and result is defined by the distance of them [9]. The model is trained by maximizing the conditional likelihood of the result documents which has been clicked while giving the query while given keyword is extracted from Clickthrough Data set of website. [6] propose a hierarchical Auto-Tagging system, TagHats, to strengthen the information sharing between users. TagHats can assign three kinds of tags: category, topic and keyword, different category tags can classify documents according to different viewpoints. Different topic tags and keyword tags can identify the content described by document more exactly. Furthermore, these hierarchical tags are very helpful for the all documents and the support to users, because different user have different demand in terms of different tag. [20] propose a LDA-based tag recommendation method, resources are often marked by many users, so each resource have a steady and complete tag set, this tag set can be used to generate potential topics, topics are often described by symbols and tags. Based on this, it map new resources to potential topics according to content, and recommend most relevant tags from these topics. Song et. al. find existing study on the tag recommendation is focused on improving the accuracy or process of automation, ignoring the problem of efficiency in [14]. So they propose a high level automatic model. This model can real-time recommend tags to resources. Resources which has been tagged is regarded as a triple (word, document and tag) and represented as a bipartite graph. By using Spectral Recursive Embedding (SRE), resources have been divided into different clusters. All tags of each cluster have been sorted by a new algorithm. [14] also propose a Bidirectional Poisson Mixture Model (PMM) that can mix component of each cluster and aggregate words to word cluster. Model the document distribution by using this way. After modeling, new document can be classified based on prior probability. Tags can be recommended to document based on the tag ranking. [8] propose an Auto Tagging method aimed at sparse short text resources, this method can create a special resource related corpus and generate potential topic to resource and corpus by using LDA. Then achieve the most relevant tags and give them to resource automatically. [11] find some shortcomings of Interactive personalized tag recommendation system in Flickr (when a user input or select a new tag to a specific picture, if this or another user has input some tags, system will recommend these tags to initial user. the tag which was recommended will update dynamically based on each selection or input) and propose a method named Hybrid to improve these shortcomings. Compared with original method, Hybrid has a better performance. Hybrid has only 1 tunable parameter and the Robustness of Hybrid is very high. moreover, [11] also provides a simple method of conservative performance analysis methodology, shows how typical classification algorithm is applied to this problem, introduce a new measurement to capture the effect of whole process of tagging, specify when the pure local program that use history of a user can be replaced by global program that use history of all users. [13] test different kinds of tag ranking strategies, build tag clouds to represent the set of object which has been tagged. Si et. al. propose a scalable real-time tag recommendation method in [19], model the documents, words and tags by using LDA. Through this model, they can real-time derived likelihood of recommending some tag to a document and use it to select recommended tags. Heymann et. al. focus on the prediction of tags in social network in [15]. Giving an object set and a tag set which all tags in it are used by user, they can predicted if a tag can be used to a specific object. They crawled resources form del.icio.us and the date of these resources to research this problem. For the URLs of del.icio.us, they predicted the tags based on the texts on pages, final texts and hosts around of it. They use a Entropy-based measurement to capture the Generality of tags and analysis the performance

of prediction. [15] also propose an Association rules based on tagging. This rule can supply the prediction with high accuracy. [18] build a system named TagAssist, it can utilize content of existing tags to assist to recommend tags to blogs, improving the quality of tags by compressing content of it. Moreover, this system can analysis the quality of recommended tags by using a series of criterion. With the help of tags manually add by users, this system can give resources the most suitable tags and increase the utilization rate of the system retrieval and browsing.

For the algorithms and dataset used in tag recommendation, [9] use a kind of method named word hashing in order to make the model can be used for large scale web search engine. Word hashing can expand the scale of the model and make the vocabulary which have large amount words can be easily utilized. System in [6] contain three algorithms, first is hierarchical classification, for assigning category tags and topic tags. The second is keyword extraction, for constructing document structure. The third is the method that select candidate topics from each category. The experiment in Oshietegoo show that the system can assign appropriate tag for each document. [8] use BibSonomy data set to offline analysis algorithms and result show the good performance of it. Scientific document data in CiteULike and web pages in del.icio.us is used to experiment, result show the model in [14] can recommend tags effectively and accurately. Algorithm in [13] based on random walk in the graph, diversification and level aggregation, this algorithm is proved effective in the experiment by using labelled data set in Flickr. The result of [15] give an inspiration to researches which regard tagging system as a potential IR tool. [19] use distributed training process in order to handle large-scale data in web. Training the model in some computers. For the algorithms used in tag recommendation. The data used in [19] is blogs data, including 386012 documents, result show the performance of this model is much better than Collaborative filtering based on search algorithms. [3] use information retrieval algorithms to analysis the relevance between blogs, specifically, using IR engine to supply the indexes to all blogs, input the queries generated from initial blog to IR engine, use specific retrieval model to retrieve the blogs which have the highest score are most relevant. And then use the simply heuristic algorithms to recommend tags to blog, combining the tags which represent the blogs that have highest score according to weight, the weight depended on the Frequency of occurrence in these blogs. The algorithm used in [2] are RankSVM and Genetic Programming, they are used to generate sorting function to combine different measurements, accurately analysis the relevance of a tag and a giving object. For the algorithm which used this technique, empirical analysis shows good performance of it, so the technique proposed in [12] can provide a large scale of Automatic Generation of Personalized Annotation TAGs for the web, and use it as the key step of recognizing Semantic Web.

III. Overview of the Tag Recommendation

In this section, we first described the data used in this paper in detail, and then discussed the preprocessing of the data. This process includes two steps. For the resources, we use the LDA to generate the topic distribution of each resource, treat it as the representation. For the tags, we use the word2vec to generate the representative vector of each tag. After that, we calculate the relevance of each resource and all tags by using Deep Structured Semantic Model. Lastly, range the result of calculation and treat the top N tags as the extended tags of a resource.

A. Data preprocessing

The data we used in this paper is extracted from big data platform of related program. The original data are public digital cultural resources and initial tags of these resources. The cultural resources include the description text of cultural videos, the introduction of collections in museum, books related to culture. The number of resources and tags are show in Table I.

Table I. The number of resources and tags

Resources	Tags	Resource-tag pairs
6448	13693	64101

For the public digital cultural resources, we use latent Dirichlet allocation (LDA) to generate the vector to represent them. LDA is a topic model and can extract the topic of every document in the document sets according to the given probability distribution [16][17][21]. **Fig. 1** shows the plate graph representation of the LDA model.

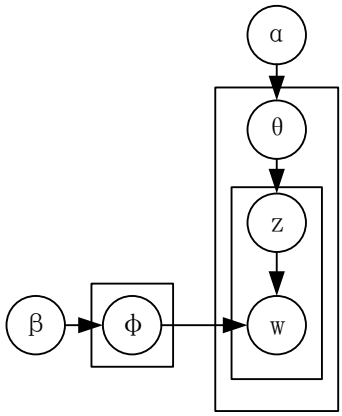

Fig. 1. Plate graph presentation of LDA model

In this figure, LDA model can be divided into two processes. In the process of $\alpha \rightarrow \theta \rightarrow z$, the doc-topic distribution generate the number z of each topic. In the process of $\beta \rightarrow \phi \rightarrow w$, the word of document(resource) can be generated by the distribution of ϕ. In this paper, we treat the LDA model by using our public digit cultural resource, and obtain the probability distribution of topic in each document. Then we can compute $p(topic|doc)$ of all topics to each document. We treat this probabilities as the elements of resource vector.

For the initial tags of these resources, we use Word2Vec to generate the vector to represent them. Word2Vec is a tool used

to extract the vector of a word. It contains Continuous Bag of Words (CBOW) and Skip-gram models [5][10]. In this paper, we used the CBOW model to map all initial tags to a vector space and we can achieve the representative vector of each tag. The CBOW model is shown in **Fig. 2**

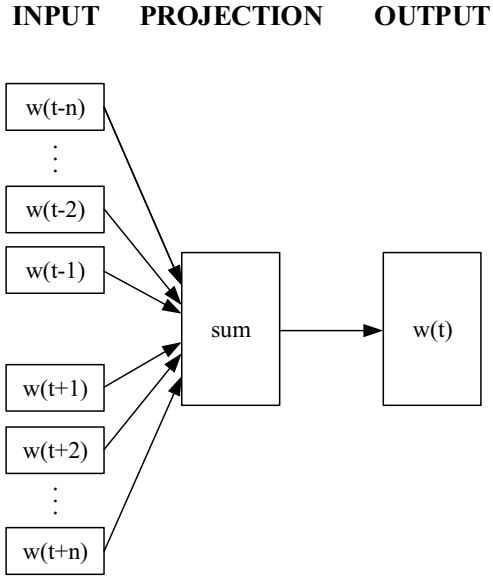

Fig. 2. CBOW model of word2Vec

The CBOW model contain three layers: input layer, projection layer and output layer. They predict w(t) according to w(t-n), …, w(t-1), w(t+1), …, w(t+n). For the purpose of generating the vector of all initial tags, we train the CBOW model by using the data that comes from all the entries of Baidupedia, including 864705 words. After training, all these words was mapped into a vector space and have a vector representation, we search words and their vector representation from the result based on our initial tags. 11766 of all 13693 tags can be found in the result.

After we get the vector of each resource and tag, we put the resource-tag pair to train Deep Structured Semantic Model (DSSM). The DSSM structure used in this paper is shown in **Fig. 3**. In this figure, I indicates the input vector and O indicates output vector, for the intermediate hidden layer l_1, l_2, … l_n, W_1, W_2, … W_n indicate the corresponding weight matrix and b_1, b_2, … b_n indicate the bias term.

For the layer of this network, we have

$$l_1 = W_1 I$$

$$l_i = f(W_i l_{i-1} + b_i), i = 2, \dots, N-1 \quad (1)$$

$$O = f(W_N l_{N-1} + b_N)$$

where the activation function is the $\tan h$ function:

$$f(x) = \frac{1-e^{-2x}}{1+e^{-2x}} \quad (2)$$

For the cosine similarity of user and resource, we can compute it by using

$$cos(O_R, O_T) = \frac{O_R^T O_T}{\|O_R\| \cdot \|O_T\|} \quad (3)$$

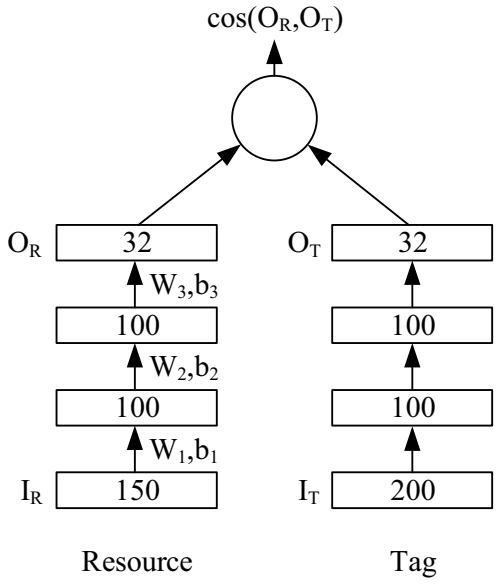

Fig. 3. The structure of DSSM

where the O_R and O_T are the semantic vectors of resource and tag. This cosine similarity is also relevance score of resource and tag. We can rank the tags by this relevance score and list them when the model have been trained and treat the top N tags as the extended tags of a resource.

Then we can estimate the probability of the resource given a tag. This posterior probability of a resource given a tag can be computed through a softmax function

$$P(R|U) = \frac{e^{\gamma cos(O_R, O_T)}}{\sum_{T' \in T} e^{\gamma cos(O_R, O_{T'})}} \quad (4)$$

where

- γ is a smoothing factor, which is usually set empirically;
- R indicates the set of candidate resource to be extended, it contain all the resources;
- T indicates the tag.

For each resource and tag pair, we use (R, T^+) to replace (R, T). T^+ is the initial tag, and we use T^+ and N randomly unvisited resource to approximate T.

In training, the model parameters are estimated to maximize the likelihood of the visited resources given the tag across the training set.

$$L(\Lambda) = -\log \prod_{(U,R^+)} P(R^+|U) \quad (5)$$

IV. Experiment

In this section, we elaborate the experiment process. In the experiment, we used the deep learning tools TensorFlow. TensorFlow is an open source software library for numerical computation using data flow graphs. The graph nodes represent mathematical operations, while the graph edges rep-resent the multidimensional data arrays (tensors) that flow between them. This flexible architecture lets you deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile device without rewriting code. TensorFlow was originally developed by researchers and engineers for the purposes of conducting machine learning and deep neural networks research. The system is general enough to be applicable in a wide variety of other domains as well. We use the off-the-shelf modules and functions to build the DSSM.

A. Experiment setup

In experiment, we first use LDA and Word2Vec to generate the vector resource and tag, and then training the DSSM according to the initial resource-tag pair. The dimension of resource and tag are set to be 200 and 150 respectively. The After that, we calculate the relevance of each resource and all tags, range the result of calculation and treat the top N as the new tags of this resource. **Fig. 4** shows this process

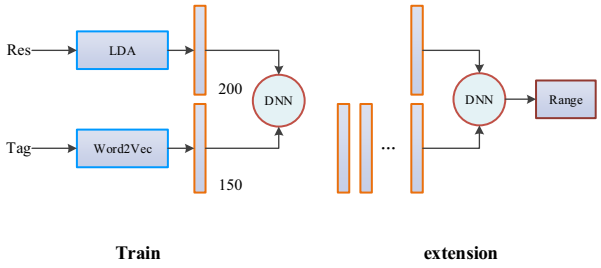

Fig. 4. The process of train and prediction

Algorithm 1 is the process of training. We use the gradient descent to approach the optimal solution in the iteration. After N times of iteration, both W_R and W_T will become the approximate optimal solution.

Algorithm 1 Training the DSSM

1: **Input:** N = the number of training iterations
R_A = resource view architecture, T_A = tag view architecture
R_D = resource input files, T_D = tag input files
$\mathbf{W_R}$ = resource view weight matrix, $\mathbf{W_T}$ = tag view weight matrix
2: **Initialization**
3: Initialize $\mathbf{W_R}$ and $\mathbf{W_T}$ using R_A and T_A
4: **For** n = 1 to N
5: $N_R \leftarrow R_D$
6: $N_T \leftarrow T_D$
7: train $\mathbf{W_R}$ and $\mathbf{W_T}$ using N_R and N_T
8: **End**
9: **Output:** $\mathbf{W_R}$ = final resource weight matrix, $\mathbf{W_T}$ = final tag weight matrix

B. Metrics

In order to evaluate the performance of algorithm, we use two kind of metrics:

- Mean Reciprocal Rank (MRR), which computes the inverse of the rank of correct tag among other tags and average the score across the whole testing data.
- Precision, which computes the percentage of how many initial tags are in the extended tags.

C. Result and discussion

We set up different kinds of extended amount N in the experiment, the results are as Table II shown. From this table, we can find that the Precision trends to increasing along with the extended amount N while the MRR decreased in general. The reason is with the increase of extended, the number of initial tag which is contained in extended tags are become larger, while the probability of extended tags situated appropriately decreased.

Table II. The results of extended amount N

N	MRR	Precision
20	42.34%	61.24%
30	41.26%	65.36%
40	43.13%	71.29%
50	38.96%	73.68%
60	38.27%	78.63%

V. Conclusion

In this paper, we discuss the availability of using Deep Structured Semantic Model-based ranking method in tag recommendation. After demonstration and comparison, for the cultural resource related data used in this paper, MRR and Precision of DSSM are better than that of other methods. The data used in this paper are all collected by big data platform of related program. It dose also illustrate the superiority of DSSM. In actual situation, the tag extended by DSSM is proved very helpful in the later process of these cultural related resources.

In the future, we will focus on the following issues. First of all, the amount and dimension of data used in this paper are not enough. Secondly, the data used in this paper are all cultural resource related paper, not universal for other types of data. Lastly, the amount of compared algorithms in the experiment is too little.

ACKNOWLEDGMENT

We would like to thank all colleagues and students who helped for our work. We thank the National Philosophy and Social Science Fund Project (15BXW061) “Collective Memory and National Identity in Internet Age”.

REFERENCES

[1] Elkahky, Ali Mamdouh, Yang Song, and Xiaodong He. "A multi-view deep learning approach for cross domain user modeling in recommendation systems." In Proceedings of the 24th International Conference on World Wide Web, pp. 278-288. International World Wide Web Conferences Steering Committee, 2015.

[2] Belém, Fabiano, Eder Martins, Tatiana Pontes, Jussara Almeida, and Marcos Gonçalves. "Associative tag recommendation exploiting multiple textual features." In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1033-1042. ACM, 2011.

[3] Mishne, Gilad. "Autotag: a collaborative approach to automated tag assignment for weblog posts." In Proceedings of the 15th international conference on World Wide Web, pp. 953-954. ACM, 2006.

[4] Wang, Jian, and Brian D. Davison. "Explorations in tag suggestion and query expansion." In Proceedings of the 2008 ACM workshop on Search in social media, pp. 43-50. ACM, 2008.

[5] Le Q V, Mikolov T. Distributed Representations of Sentences and Documents[J]. 2014, 4:II-1188.

[6] Nishida, Kyosuke, and Ko Fujimura. "Hierarchical auto-tagging: organizing Q&A knowledge for everyone." In Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1657-1660. ACM, 2010.

[7] Krestel, Ralf, Peter Fankhauser, and Wolfgang Nejdl. "Latent dirichlet allocation for tag recommendation." In Proceedings of the third ACM conference on Recommender systems, pp. 61-68. ACM, 2009.

[8] Diaz-Aviles, Ernesto, Mihai Georgescu, Avaré Stewart, and Wolfgang Nejdl. "Lda for on-the-fly auto tagging." In Proceedings of the fourth ACM conference on Recommender systems, pp. 309-312. ACM, 2010.

[9] Huang, Po-Sen, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. "Learning deep structured semantic models for web search using clickthrough data." In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pp. 2333-2338. ACM, 2013.

[10] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[11] Garg, Nikhil, and Ingmar Weber. "Personalized, interactive tag recommendation for flickr." In Proceedings of the 2008 ACM conference on Recommender systems, pp. 67-74. ACM, 2008.

[12] Chirita, Paul-Alexandru, Stefania Costache, Wolfgang Nejdl, and Siegfried Handschuh. "P-tag: large scale automatic generation of personalized annotation tags for the web." In Proceedings of the 16th international conference on World Wide Web, pp. 845-854. ACM, 2007.

[13] Skoutas, Dimitrios, and Mohammad Alrifai. "Ranking tags in resource collections." In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1207-1208. ACM, 2011.

[14] Song, Yang, Ziming Zhuang, Huajing Li, Qiankun Zhao, Jia Li, Wang-Chien Lee, and C. Lee Giles. "Real-time automatic tag recommendation." In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 515-522. ACM, 2008.

[15] Heymann, Paul, Daniel Ramage, and Hector Garcia-Molina. "Social tag prediction." In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 531-538. ACM, 2008.

[16] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[17] Blei D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77-84.

[18] Sood, Sanjay, Sara Owsley, Kristian J. Hammond, and Larry Birnbaum. "TagAssist: Automatic Tag Suggestion for Blog Posts." In ICWSM. 2007.

[19] Si, Xiance, and Maosong Sun. "Tag-LDA for scalable real-time tag recommendation." Journal of Computational Information Systems 6, no. 1 (2009): 23-31.

[20] Krestel, Ralf, and Peter Fankhauser. "Tag recommendation using probabilistic topic models." ECML PKDD Discovery Challenge 2009 (2009): 131.

[21] Bela A. Frigyik, Amol Kapila, and Maya R. Gupta. Introduction to the Dirichlet Distribution and Related Processes. ee.washington.edu. Retrieved 14 May 2015.