Topic and Keyword Re-ranking for LDA-based Topic Modeling

Yangqiu Song[†] Shimei Pan[§] Shixia Liu[†] Michelle X. Zhou[†§] Weihong Qian[†]
[†]IBM China Research Lab, Beijing, China; [§]IBM T. J. Watson Research Center, Hawthorne, NY, USA.
[†]{yqsong,liusx,mxzhou,qianwh}@cn.ibm.com; [§]shimei@us.ibm.com

ABSTRACT

Topic-based text summaries promise to help average users quickly understand a text collection and derive insights. Recent research has shown that the Latent Dirichlet Allocation (LDA) model is one of the most effective approaches to topic analysis. However, the LDA-based results may not be ideal for human understanding and consumption. In this paper, we present several topic and keyword re-ranking approaches that can help users better understand and consume the LDA-derived topics in their text analysis. Our methods process the LDA output based on a set of criteria that model a user's information needs. Our evaluation demonstrates the usefulness of the methods in summarizing several large-scale, real world data sets.

Categories and Subject Descriptors: I.2.6 [Artificial Intelligence]: Learning H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

General Terms: Algorithms, Experimentation

Keywords: Topic Model, Topic and Keyword Re-ranking

1. INTRODUCTION

Latent topic models are effective methods for extracting latent semantic information from text corpora [4, 7, 3]. Among these models, Latent Dirichlet Allocation (LDA) [3] appears to be the most effective one. LDA models a document as a mixture of latent topics and each topic can be further represented by a set of keywords. As a result, it allows a document to belong to multiple topics. However, LDA is a general model without considering different users' information needs. For example, the LDA topics are randomly ordered, and users must manually navigate the topic list to find the ones that they are interested in. This task becomes more difficult, if there is a large number of topics. To better help users consume the LDA-derived topics in an interactive visual text analytic process [9], we focus on enhancing the LDA topic modeling results.

In this paper, we present a set of re-ranking techniques to enhance the topic modeling results. The work closest to ours is an approach described in [2], which uses a TFIDF-like term score to re-rank keywords. Compared to this work, ours also re-ranks the LDA-derived topics in addition to re-ranking the topic keywords. As a result, our work can select the most salient topics that meet different user interests.

2. LDA TOPIC MODELING

We denote a text corpus as $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ where d_m is a document and M is the number of documents in the corpus. Each document consists of a sequence of words $\mathcal{W}_m = \{w_{m,1}, w_{m,2}, \ldots, w_{m,N_m}\}$, where N_m is the number of words in document d_m. We also define a dictionary $\mathcal{V} = \{v_1, v_2, \ldots, v_V\}$, where V is the size of the dictionary. Moreover, z is a latent variable representing the latent topic associated with each observed word. We denote $\mathcal{Z}_m = \{z_{m,1}, z_{m,2}, \ldots, z_{m,N_m}\}$ as the topic sequence associated with the word sequence $\mathcal{W}_m$. The generative procedure of LDA can be formally defined as:

1. For all the topics $k \in 1, \ldots, K$:
   Choose a word distribution $\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\boldsymbol{\varphi}|\boldsymbol{\beta})$.
2. For all the documents d_m where $m \in 1, \ldots, M$:
   2.1. Choose $N_m \sim \mathrm{Poisson}(\xi)$.
   2.2. Choose a topic distribution $\boldsymbol{\theta}_m \sim \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$.
   2.3. For all the words $w_{m,n}$ where $n \in 1, \ldots, N_m$:
      Choose a topic index $z_{m,n} \sim \mathrm{Mult}(z|\boldsymbol{\theta}_m)$.
      Choose a word $w_{m,n} \sim \mathrm{Mult}(w|\boldsymbol{\varphi}_{z_{m,n}})$.

We denote $\boldsymbol{\varphi}_k = (\varphi_{k,1}, \varphi_{k,2}, \ldots, \varphi_{k,V})^T \in \mathbb{R}^V$ where $\varphi_{k,i} = p(w = v_i|z = k)$. Thus, the parameters for the topic-word distribution can be represented as $\boldsymbol{\Phi} = (\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_K)^T \in \mathbb{R}^{K\times V}$, where K is the topic number. Moreover, we denote $\boldsymbol{\theta}_m = (\theta_{m,1}, \theta_{m,2}, \ldots, \theta_{m,K})^T \in \mathbb{R}^K$ where $\theta_{m,k} = p(z = k|d_m)$. Then the parameters for document-topic distribution is $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)^T \in \mathbb{R}^{M\times K}$.

Given a set of training documents, inferencing a topic model involves estimating the document-topic distribution $\boldsymbol{\Theta}$ and the topic-word distribution $\boldsymbol{\Phi}$ [3, 5]. In our experiments, we use Gibbs sampling [5].

3. LDA MODEL ENHANCEMENT

In this section, we introduce our topic and keyword re-ranking methods to enhance the LDA output for more effective human consumption.

3.1 Keyword Re-ranking

We first describe how we re-rank LDA-derived topic keywords since the order of these keywords directly affects the

CIKM'09, November 2–6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-512-3/09/11 ...$10.00.

semantics and thus the importance of a topic. The native order of topic keywords produced by LDA may not be ideal for users to understand the semantics of a topic. For example, when LDA is applied to a financial news corpus, common words such as Dow, Jones, Wall, Street etc., are normally ranked high in many topics because they are relevant to all of the topics. These words however are not useful in helping users identify interesting topics since all of them are about finance. To better help people identify salient information, we thus re-rank the LDA-derived topic keywords to refine the topic definitions.

Inspired by the term re-weighting techniques used in information retrieval (IR) [13, 8], we have experimented with two LDA-versions of TFIDF-like scores:

$$KR_1 = \frac{\hat{\varphi}_{k,i}}{\sum_{k=1}^{K} \hat{\varphi}_{k,i}} \tag{1}$$

and

$$KR_2 = \hat{\varphi}_{k,i} \cdot \log \frac{\hat{\varphi}_{k,i}}{(\prod_{k=1}^{K} \hat{\varphi}_{k,i})^{\frac{1}{K}}} \tag{2}$$

where the native weight $\hat{\varphi}_{k,i}$ generated by LDA corresponds to the term frequency. The topic proportion sum and topic proportion product are used respectively in KR_1 and KR_2 to simulate the inverse document frequency to re-weight the native weights. In fact, KR_2 is the same as the re-weighting technique used in [2].

3.2 Topic Re-ranking

The LDA-derived topics are randomly ordered which may not be equally important to a user. It is thus useful to order the topics so that the most important ones can be shown first. In general, the definition of *importance* may vary from one user to another. For example, a user may prefer to see the most talked topics, i.e. topics that cover more documents. In this case, the rank of a topic would be higher, if it covers more document content in the corpus. In contrast, a user may be interested in a set of distinct topics that have the least amount of content overlap with one another. In this case, we will rank topics based on their content uniqueness. Next, we describe a few application-independent topic re-ranking methods that computes the topic ranks based on different ranking criteria.

3.2.1 Weighted Topic Coverage and Variation

The first topic re-ranking method assumes that topics that cover a significant portion of the corpus content are more important than those covering little content. However, we consider topics that appear in all the documents (e.g. a topic on disclaims derived from a legal collection) to be too generic to be interesting, although they have significant content coverage. Thus we rank such topics lower. As a result, our first topic ranking metric is a combination of both content coverage and topic variance. More precisely, we define:

$$\mu(z_k) = \sum_{m=1}^{M} N_m \cdot \hat{\theta}_{m,k} \Bigg/ \sum_{m=1}^{M} N_m \tag{3}$$

and

$$\sigma(z_k) = \sqrt{\sum_{m=1}^{M} N_m \cdot \left(\hat{\theta}_{m,k} - \mu(z_k)\right)^2 \Bigg/ \sum_{m=1}^{M} N_m}\,, \tag{4}$$

where the weight N_m is the document length.

Then the rank of a topic is defined as:

$$TR_k^{c.v.} \triangleq (\mu(z_k))^{\lambda_1} \cdot (\sigma(z_k))^{\lambda_2}, \tag{5}$$

where λ_1 and λ_2 are the control parameters. Specifically, if $\lambda_1 = 1$ and $\lambda_2 = 0$, the ranking is determined purely by topic coverage [14]. In contrast, if $\lambda_1 = 0$ and $\lambda_2 = 1$, the rank is simply determined by topic variance, a criterion that is similar to principle component analysis [1].

3.2.2 Laplacian Score

While topic variance used in the first method reflects a topic's representative power, the Laplacian score of a topic represents its power in discriminating documents from different classes [6]. Our second method on topic re-ranking is motivated by the observation that two similar documents are probably related to the same topic while documents that are dissimilar probably belong to different topics. Since the Laplacian score of a topic reflects its power in discriminating documents from different classes while preserving the local structure of a document collection, we develop a Laplacian score-based topic ranking method so that it assigns high ranks to those topics with high discriminating power. It consists of five main steps:

1. Represent each document d_m as a node in a graph. Its features are represented by $\hat{\boldsymbol{\theta}}_m$.

2. Construct the T-nearest neighbor graph based on a similarity matrix $\mathbf{S}$ where $\mathbf{S}_{ij} = \exp\left\{-d_{ij}^2/2\sigma^2\right\}$. Here, d_{ij} can be either Euclidian distance or Hellinger distance [2].

3. Compute graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{S}$ where $\mathbf{D}$ is a diagonal matrix and $\mathbf{D}_{ii} = \sum_{j=1}^{M} \mathbf{S}_{ij}$ is the degree of the *ith* vertex.

4. For each topic $\mathbf{t}_k = (\hat{\theta}_{1,k}, \hat{\theta}_{2,k}, \ldots, \hat{\theta}_{M,k})^T \in \mathbb{R}^M$, let $\tilde{\mathbf{t}}_k = \mathbf{t}_k - \frac{\mathbf{t}_k^T \mathbf{D1}}{\mathbf{1}_k^T \mathbf{D1}} \mathbf{1}$ where $\mathbf{1}$=$[1, 1, ...1]^T$.

5. Compute the Laplacian score of the k-th topic:

$$L_k = \frac{\tilde{\mathbf{t}}_k^T \mathbf{L} \tilde{\mathbf{t}}_k}{\tilde{\mathbf{t}}_k^T \mathbf{D} \tilde{\mathbf{t}}_k}. \tag{6}$$

Remark 1: To find the T-nearest neighbors of a topic, we keep a T-size heap. For each topic, we compute its distances to all the other topics and then check whether to insert it to the heap. Thus, the main time complexity is in graph Laplacian construction which is $O(M^2K + M^2 \log T)$.

3.2.3 Pairwise Mutual Information

We have also developed a topic re-ranking approach based on the pairwise mutual information of two topics. This metric computes the information that two topics share. It also measures how much knowing one of the topics reduces our uncertainty about the other. Using this metric, we can rank topic by measuring the amount of the pairwise mutual information between two topics. Specifically, we use the following procedure [12] to determine the rank of each topic.

1. For $\forall i, j$, first compute $MI(\mathbf{t}_i, \mathbf{t}_j)$ based on the doc-topic distributions of $\mathbf{t}_i$ and $\mathbf{t}_j$. Then construct a complete graph $\mathcal{G}$ where the weight of an edge $e_{\mathbf{t}_i,\mathbf{t}_j}$ is $MI(\mathbf{t}_i, \mathbf{t}_j)$.

2. Build the maximal spanning tree MST of the complete graph $\mathcal{G} : (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are vertex and edge sets.

3. Define the relevant topic set $\mathcal{T}^{rel} = \{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_K\}$ and the corresponding edges in MST.

4. While $|\mathcal{T}^{rel}| > 0$,

4.1. if $\exists$ a node v where $\mathbf{t}_v \in \mathcal{T}^{rel}$ that is not connected to the others in $\mathcal{T}^{rel}$, remove this topic $\mathbf{t}_v$ ($\mathcal{T}^{rel} \leftarrow \mathcal{T}^{rel} - \mathbf{t}_v$).

4.2. otherwise remove the least weighted edge in $\mathcal{T}^{rel}$.

5. Rank the topics according to the order in which they were removed. Rank the last removed topic the highest.

Remark 2: We use the Prime's algorithm to construct the MST. Thus, to compute the pairwise mutual information for topic re-ranking needs $O(K^2M)$. By using a heap to construct a priority queue, we can build an MST in $O(|\mathcal{E}|\log|\mathcal{V}|) = O(K^2 \log K)$ time.

3.2.4 Topic Similarity

The last similarity-based re-ranking method is developed to maximize topic diversity and minimize redundancy. While all the above methods use topic-document relationships, here we employ topic-word distributions to compute topic similarity. We have slightly modified the algorithm proposed in [11] to derive a topic rank:

1. For $\forall i, j$, compute the similarity s_{ij} for $\boldsymbol{\varphi}_i$ and $\boldsymbol{\varphi}_j$ based on *maximal information compression index* [11].

2. Sort the similarities for each topic.

3. Define the reduced topic set $\mathcal{T}^{red} = \{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, ..., \boldsymbol{\varphi}_K\}$.

4. While $|\mathcal{T}^{red}| > 0$, remove $\boldsymbol{\varphi}_j$ in $\mathcal{T}^{red}$ which satisfies $j = \arg\max_i \max_j s_{ij}$.

5. The rank of a topic is determined by the topic removal order. The last removed topic should be ranked the highest.

Remark 3: In this algorithm, constructing the similarity scores needs $O(K^2M)$ and sorting the scores needs $O(K^2 \log K)$.

4. EXPERIMENTS

We have tested our topic and keyword re-ranking techniques using two different data sets in a series of experiments.

The first data set is a personal email collection dated from February to December 2008 with 8326 email messages. Each email is associated with a set of meta data such as sender, receiver, time, subject, body and reply counts. Only the subject and the body of each email were used to train the topic model. We pre-processed each email to remove irrelevant information such as email signature and also did stop word removal. After pre-processing, the email collection contained 958,069 word tokens in total.

The second data set is an online document collection that contained text retrieved by a search engine. These documents came from various news, blog and forum web sites. The search engine used "AIG insurance" as the query and retrieved 34,690 documents from January 2008 to April 2009. After pre-processing, the final AIG collection contained 11,491,246 word tokens in total.

To test our methods, we ran LDA five times and we show both the average and standard derivation for the five runs. We adopted an LDA algorithm that was trained with optimized hyper-parameters. For each run of LDA, we set the maximum iteration to 1000. The initial model parameters were set to the default values in the Mallet LDA toolkit [10]. We empirically set the topic number for the email and AIG data sets to 18 and 20 respectively.

4.1 Topic Re-ranking

For each data set, we asked an expert to annotate the topics and the corresponding topic keywords learned by our methods. The annotation was repeated for each of the five LDA runs on each data set.

4.1.1 Annotation Criteria

For the email data set, the person who owned the email collection helped us annotate the results. She was asked to label each topic as either "very important", "somewhat important" or "unimportant". In addition, for each topic, she was also asked to label each of the top 50 keywords as either "relevant" or " irrelevant". When asked about how she ranked these topics, the email owner summarized her criteria as: (1) A "very important" topic clearly describes a major project that the email owner heavily involved. (2) A "somewhat important" topic focuses on a specific event, such as writing a paper. (3) An "unimportant" topic either lacks a clear focus or is about very general work-related activities.

For the AIG news data set, we asked a person who was familiar with the recent AIG-related events to help us annotate the topics and keywords. This person determined the importance of a topic as follows: (1) A "very important" topic clearly describes an event directly related to AIG, e.g. the AIG bonus controversy. (2) A "somewhat important" topic focuses on some background events such as the 2009 presidential election or the financial market crisis. (3) An "unimportant" topic is defined as either confusing or irrelevant, e.g. a topic about various advertisements.

Given the annotated topics and keywords, we compared the automatic topic and keyword re-ranking results with the human-provided results using the F_1-measure, a criterion commonly used in information retrieval (IR). Following the IR tradition, in our analysis we categorized our topics annotated by our experts into both "relevant" and "irrelevant". The "relevant" topics are those that are either "very important" or "somewhat important" while "irrelevant" topics are those that are "unimportant". Similarly, based on our topic or keyword re-ranking methods, each topic or keyword can be categorized as either "retrieved" or "not retrieved" depending on the assigned ranks and the cut-off threshold used in each evaluation metric ("top 5" means only the top five keywords are retrieved).

4.1.2 Re-ranking Results

The keyword re-ranking results are shown in Tables 2 and 4. In these Tables, KR_0 is the baseline that uses the LDA estimated parameters $\hat{\varphi}_{k,i}$ directly; KR_1 and KR_2 are defined in section 3.2. We can see that KR_1 performed better than KR_0 and KR_2. It shows that for the two selected data sets, topic proportion sum is better than topic proportion production in weighing the proportion.

The topic re-ranking results for the email data set are shown in Table 3. In the Table, "C.V." represents *Weighted Topic Coverage and Variation*, "L.S." represents *Laplacian Score*, "M.I." represents *Pairwise Mutual Information*, and "T.S." represents *Topic Similarity*. Our results show that the Laplacian score-based method outperformed the other methods. In particular, all the top five retrieved topics were labeled by the email owner as relevant.

The topic re-ranking results for the AIG data set are shown in Table 5. The Laplacian score-based method also outperformed all the other methods using this data set. Overall, the Laplacian score-based re-ranking method seems to capture the essence of an important topic the best. Table 1 shows the details of the Laplacian score-based topic re-ranking results for the AIG data set. It contains the top 10 keywords of all the 20 topics.

Table 1: AIG topics ranked by Laplacian score (bus. means business; contr. means controversy; gov. means government; int'l means international).

job related	market info	int'l market	unclear focus	insurance bus.	retreat contr.	general bus.	bonus contr.	ads	Spitzer probe
aig	crude	hbos	rating	quote	retreat	aia	bonus	gps	spitzer
ng	stock	bank	insurer	farm	spa	insurer	taxpayer	diamond	settlement
jobcircle	oil	japan	business	travel	regis	policyholders	payments	shirts	investigation
description	price	european	subsidiaries	progressive	resort	company	compensation	cellular	fraud
employer	percent	asia	assets	car	bancorp	exposure	executive	garden	greenberg
resume	trading	london	reserve	cheap	executives	clients	employees	jeans	ceo
location	investors	brothers	monday	homeowners	committee	capital	ceo	ipod	executive
apply	dollar	cash	capital	insurance	lawmakers	income	administration	shoes	chairman
experience	gold	lehman	federal	life	bailout	commercial	bailout	ringtones	products
title	financial	crisis	american	medical	event	million	dollars	silver	services
insurance bus.	election	credit swap	no focus	financial company	no focus	web related	gov. and crisis	ads	congress
quotation	republican	swap	thing	stanley	amp	story	capitalism	movie	fortune
universal	mccain	mortgage	remember	morgan	quot	post	paulson	clip	voted
health	voters	hedge	pretty	jpmorgan	nbsp	comment	bernanke	video	legislation
policy	obama	derivatives	idea	merrill	ldquo	google	economy	drug	republicans
protection	presidential	sell	people	goldman	message	thread	war	game	leaders
care	election	bonds	save	bank	www	click	crisis	sitemap	democrats
coverage	candidate	fund	hard	lehman	public	article	growth	crack	lawmakers
benefits	clinton	risk	work	lynch	read	blog	freddie	music	bush
medical	bush	investment	place	bankruptcy	investments	read	debt	torrent	congress
life	economic	credit	problem	wall	business	daily	congress	porn	paulson

Table 2: Email Keyword Re-ranking Results.

Retrieved	Top 10	Top 20
KR_0	0.535 ± 0.055	0.442 ± 0.048
KR_1	**0.701 ± 0.067**	**0.600 ± 0.043**
KR_2	0.616 ± 0.078	0.551 ± 0.049

Table 3: Email Topic Re-ranking Results.

Retrieved	Top 5	Top 10
C.V.	0.800 ± 0.000	0.620 ± 0.028
L.S.	**1.000 ± 0.000**	**0.780 ± 0.028**
M.I.	0.760 ± 0.106	0.740 ± 0.035
T.S.	0.440 ± 0.057	0.480 ± 0.028

Table 4: News (AIG) Keyword Re-ranking Results.

Retrieved	Top 10	Top 20
KR_0	0.466 ± 0.111	0.445 ± 0.083
KR_1	**0.662 ± 0.094**	**0.614 ± 0.054**
KR_2	0.403 ± 0.079	0.343 ± 0.048

Table 5: News (AIG) Topic Re-ranking Results.

Retrieved	Top 5	Top 10
C.V.	0.640 ± 0.057	0.68 ± 0.028
L.S.	**0.760 ± 0.057**	**0.76 ± 0.035**
M.I.	**0.760 ± 0.057**	0.74 ± 0.035
T.S.	0.720 ± 0.069	0.70 ± 0.045

5. CONCLUSIONS

In this paper, we have presented several topic and keyword re-ranking methods. Our experiments show that among the topic keywords re-ranking methods that we investigated, KR_1 outperformed others on both of our test data sets. Moreover, among all the topic re-ranking methods that we tested, the Laplacian score-based method performed the best.

We are also working on several areas to further improve the current re-ranking methods. One area is to allow users to interactively define their topic or keyword ranking criteria. We are also interested in exploring the use of various application-specific features to build more accurate topic summarization systems.

6. ACKNOWLEDGMENT

We would like to thank David Mimno, Tom Griffiths, and Gregor Heinrich for their help and advices on LDA and its implementation.

7. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.

[2] D. Blei and J. Lafferty. *Topic Models*, chapter: Topic Models. Taylor and Francis, 2009. (in Press).

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

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

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

[6] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *Proceedings of NIPS*, 2005.

[7] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of UAI*, pages 289–296, 1999.

[8] M. Lan, C. Tan, H. Low, and S. Sung. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *WWW (Special interest tracks and posters)*, pages 1032–1033, 2005.

[9] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topics-based visual text summarization and analysis. In *Proceedings of the CIKM*, 2009.

[10] A. K. McCallum. Mallet: A machine learning for language toolkit. 2002.

[11] P. Mitra, C. Murthy, and S. Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, 2002.

[12] M. Sahami. *Using Machine Learning to Improve Information Access*. PhD thesis, Department of Computer Science, Stanford University, USA, 1998.

[13] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[14] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of ICML*, pages 412–420, 1997.