he_2019_automatic_labeling_of_topic_models_using_graph_based_ranking

Year

2019

Author(s)

He, Dongbin and Wang, Minjuan and Khattak, Abdul Mateen and Zhang, Li and Gao, Wanlin

Title

Automatic Labeling of Topic Models Using Graph-Based Ranking

Venue

IEEE Access

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel approach

Underlying technique

Redundancy-aware graph-based ranking model (TLRank)

Topic labeling parameters

Considered topic terms: 500

Length of the topic label: 250

Label generation

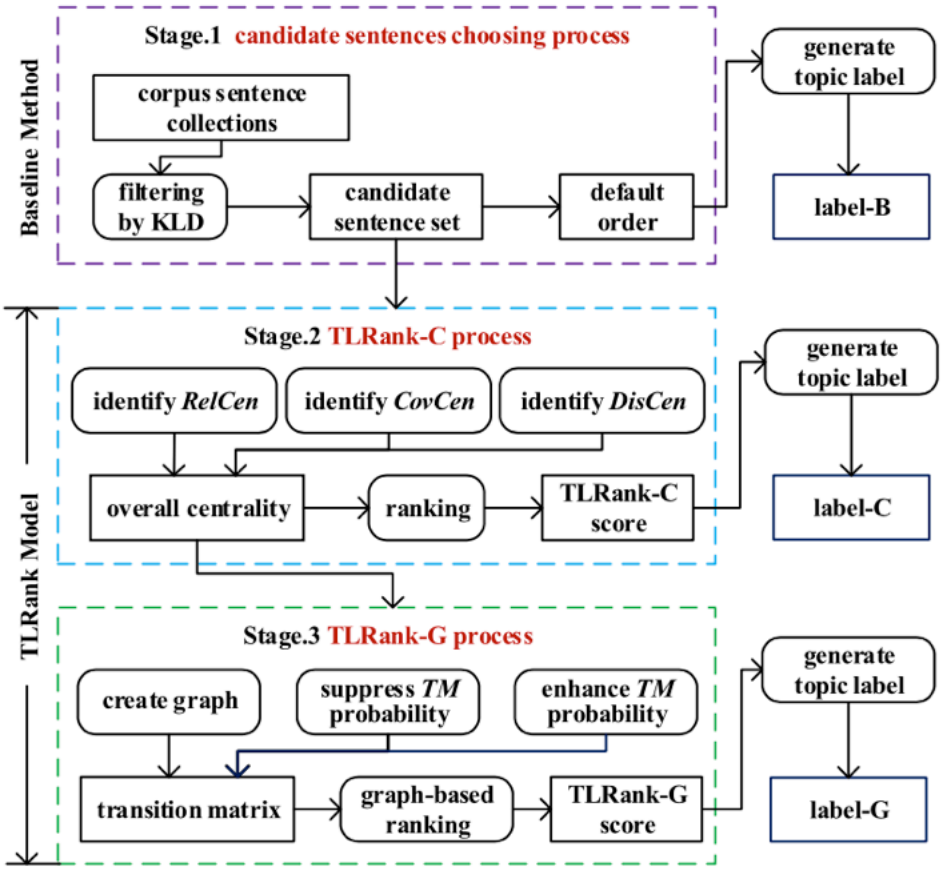

FIGURE 1. An overview of the labeling process of TLRank model.

Topic pseudo-sentence

A centroid is a set of terms to represent a cluster of document collection and illustrate important statistical characteristics.

The discovered topics modeled by the LDA-style approaches can be approximately represented with their major features, i.e. choosing top 500 terms of the discovered topic into a term set $V\_top500\theta$ to alternatively represent the centroid of a discovered topic.

In addition, since a centroid of the discovered topic and a topic label generated by labeling method have different representations, it is hard to directly evaluate the quality of the generated topic labels by computing their similarity. In order to address this issue, we convert the centroid terms set to a pseudo-sentence, which consists of top 500 topic terms and does not contain syntax and grammar. In

this way, we can consider a topic pseudo sentence (TPS) as a discovered topic, and it can be described as follows.

$$TPS(\theta) = \{p_\theta(w)\}_{w \in V_{top500\theta}} \Big/ \left\| \{p_\theta(w)\}_{w \in V_{top500\theta}} \right\|$$

where the *p*_θ (*w*) indicates the probability of word *w* in the topic θ, *V*_*top*500θ represents the top 500 terms set of topic θ.

$$s = \{tfidf(w)\}_{w \in S}$$

s represents the centroid of the sentence, and *S* denotes the words set in the sentence.

THE TLRANK MODEL

CANDIDATE SENTENCES CHOOSING PROCESS

In the first stage, we select the most relevant sentences to set up the candidate sentences sets for each discovered topic.

Using KLD measured similarity between the topic label and TPS is an effective approach in topics labeling task

$$KLD(TPS(\theta), s) = \sum_{w \in SW \cup V_{TPS}} p_\theta(w) * \log \frac{p_\theta(w)}{tf(w, s) \Big/ len(s)}$$

where *TPS*(θ) denotes the TPS of discovered topic θ, *s* represents the corresponding topic label or sentence in the corpus, *V*_*Tps* represents the words set of TPS, *SW* denotes the words set of sentence *s* after removing digital, stop words and terms with length less than threshold *m* (minimum term length), *tf* (*w*, *s*) represents the frequency of *w* in *s*, and *len*(*s*) denotes the count of *SW.*

In this stage, we extract the top 500 sentences from the current corpus for each discovered topic, and then take them as candidate sentences sets (*CSSets*).

THE TLRank-C PROCESS

In this section, we first identify the three central features of candidate sentences, and then forge them to overall centrality. According to the ranking order based on the centrality value of sentences, we generate a topic label for each discovered topic.

RELEVANCE CENTRALITY

The Relevance centrality (*RelCen*) is a measurement based on the textual similarity between the candidate sentence and TPS, which is computed using the following equation.

$$RelCen(s, \theta) = \exp(KLD(TPS(\theta), s)^{-1}) \Big/ len(s)^a$$

where *s* represents the candidate sentence, the exponential smoothing parameter *a* is used to optimize the result.

COVERAGE CENTRALITY

It is quite intuitive that a candidate sentence with good Coverage should contain as many different words as possible, which is very similar to the diverse requirements of multi-document summaries. In addition, the sum of probability of different non-repetitive words in the sentence can well reflect the sentence Coverage to the topic. So the Coverage centrality (*CovCen*) equation is shown as follows

$$CovCen(s, \theta) = \sum_{w \in SW} p_\theta(w)\ tf(w, s) \Big/ len(s)^a$$

where *s* represents a candidate sentence.

DISCRIMINATION CENTRALITY

Referring to Eq. (8), we argue that the importance of a sentence *s* belonging to the topic θ can be measured by the ratio of the *Cov* (*s*, θ) to P *Cov* (*s*, θ). The higher the ratio is, the more important to the topic θ and the more discriminative to other topics the sentence *s* is. So, the Discrimination centrality (*DisCen*) equation can be written as follows.

$$DisCen(s, \theta) = \frac{\sum_{w \in SW} p_\theta(w)\ tf(w, s)}{\sum_{\theta^* \in U} \sum_{w \in SW} p_{\theta^*}(w)\ tf(w, s)}$$

where *U* represents a set of all the discovered topics.

OVERALL CENTRALITY

The three centrality features of the candidate sentence (*RelCen*, *CovCen* and *DisCen*) are corresponding to the three criteria for evaluating the quality of topic labels (Relevance, Coverage, and Discrimination). The purpose is to identify the three centrality features that aim to find suitable sentences and hence improve the three evaluating scores of the topic labels generated by our method. In order to use a scale value to measure the sentences in the ranking process, we introduce a new concept, overall centrality (*OC*). This combines the three centrality features to estimate the overall quality of sentences and generates summaries. This way, we achieve a satisfactory balance among the three evaluating criteria. Moreover, the overall centrality of sentence *s_y* (*OC y*) equation is defined as follows.

$$OC_y = \alpha RelCen(s_y, \theta) + \beta DisCen(s_y, \theta) + (1 - \alpha - \beta) CovCen(s_y, \theta)$$

where α and β are proportion parameter to be empirically set, and it has α > 0, β > 0, α + β < 1.

THE TLRank-C PROCESS

In this section, candidate sentences in *CSSets* are used as vertices to create a directed complete graph, and a novel graph-based ranking algorithm is given accordingly. We will accomplish the following tasks. First, a transition matrix (*TM*) is established according to overall centrality value, and then, the transition probability *TM_xy* between vertex-pairs is suppressed according to their relationship and enhanced base on their *Degree* (a characteristic of each vertex). Thence we can ensure that *TM* is the transition matrix of a Markov chain and the random walk algorithm can converge to a stable state after a finite number of iterations. Finally, we generate a topic label for each discovered topic based on the TLRank-G score.

DIRECTED COMPLETE GRAPH

Consider *G* = (*Vertices*, *Edges*) to be a directed complete graph, *x*,*y* ∈ *Vertices*, Cand *x* != *y*, C *edge_xy* ∈ *Edges* represents the weight of edge from *x* points to *y* and its original value can be derived from the overall centrality of vertex *y* (sentence *s_y*). Here, it has *edge_xy* = *OC*_y , which means that vertex *x* votes to *y* based on the overall centrality value of vertex *y*. In this graph, the weights of edges are the critical factor in determining the transition matrix, which directly defines the output of the graph-based ranking algorithm. Therefore, it is feasible to modify the weights of edges in the graph to adjust the ranking results.

SUPPRESS WEIGHT OF EDGE

In order to refrain the redundancy of the topic label during the generating process, the sentence which has the smallest similarity with those existing in the topic label set should be chosen preferably. Therefore, one of the important tasks of our ranking method is to ensure that the sentence in the back position of the ranked order keeps the similarity as small as possible with the ones ahead.
In this paper, we use the Jaccard distance to measure the similarity between two vertices (sentences) in the graph. In general, for two sentences *s_x* and *s_y*, the Jaccard distance equation is defined as follows.

$$similarity\_Jaccard(s_x, s_y) = \cap(s_x, s_y) \Big/ \cup(s_x, s_y)$$

If vertex *x* (*s_x*), and *y* (*s_y*) in the graph are similar, and it has *OC_x* > *OC_y*, then the *edge_xy* should be suppressed, and the suppression coefficient bases on the value of similarity_Jaccard(*s_x* , *s_y*). The *edge_xy* which has been sup- pressed can be described as follows.

$$edge_{xy} = edge_{xy} \Big/ e^{similarity\_Jaccard(s_x, s_y)}$$

ENHANCE WEIGHT OF EDGE

`look at paper`

GRAPH-BASED RANKING PROGRESSING

`look at paper`

Motivation

Although the manual topic label is more interpretive and understandable, labeling topics need considerable human labor to review massive data of the corpus. Moreover, it tends to add subjective opinions to the manual label unconsciously.
Thus, automatic summarization could be the preferable choice to assemble the topics label by the extractive sentences manner.

Topic modeling

LDA

Topic modeling parameters

For the Gibbs method, it is necessary to fix the parameter k = 25 and set the parameters *burnin* = 1000, *thin* = 100, *iter* = 1000
a = 0.85, α = 0.3, β = 0.4, γ = 2.25, t = 0.2 and m = 3

Nr. of topics

25

Label

In this paper, for interpreting the discovered topics, we extract the appropriate sentences to generate a more meaningful and concise topic label for each topic.

Label selection

\

Label quality evaluation

AUTOMATED

RELEVANCE

For each labeling method, we computed the Relevance (using KLD method) between the topic labels and corresponding TPS in the four different cases and then averaged the Relevance of all topics in each case.

TABLE 1. Relevance of topic label; the average of KL divergence between topic label and discovered topic.

	SIGMOD VEM	SIGMOD Gibbs	APNews VEM	APNews Gibbs
Baseline	0.813665	2.288893	0.401759	**0.996711**
LexRank	**0.634523**	2.553759	0.431941	1.109661
TextRank	0.705769	2.759404	0.535805	1.265098
Submodular	0.814613	3.315231	0.633104	1.57551
TLRank-C	0.698133	2.212693	0.338974	1.006217
TLRank-G	0.696214	**2.204864**	**0.333005**	1.000132

COVERAGE

For each labeling method, we computed the Coverage-the ratio of the top 20 topic terms in the corresponding topic label- for each topic in the four different cases. Then we fetched the average, min, and max from the Coverages of all topics in each case. The reason that choosing the top 20 terms instead of 500 was that the top 20 terms were more significant than the rest and we paid more attention towards the more representative top ones

TABLE 2. Coverage of topic label; the mean, min, and max ratio of the top 20 topic terms in the corresponding topic label.

	SIGMOD VEM			SIGMOD Gibbs			APNews VEM			APNews Gibbs		
	Mean	Min	Max	**Mean**	Min	Max	**Mean**	Min	Max	**Mean**	Min	Max
Baseline	0.676	0.550	0.800	0.700	0.600	0.850	0.662	0.450	0.850	**0.704**	0.450	0.900
LexRank	**0.744**	0.600	0.900	0.626	0.400	0.800	0.616	0.350	0.800	0.624	0.400	0.850
TextRank	0.708	0.450	0.900	0.566	0.300	0.800	0.548	0.200	0.750	0.570	0.250	0.800
Submodular	0.614	0.250	0.850	0.496	0.050	0.750	0.382	0.050	0.700	0.462	0.100	0.800
TLRank-C	0.718	0.550	0.900	**0.708**	0.550	0.850	0.734	0.550	0.900	0.702	0.450	0.900
TLRank-G	0.722	0.550	0.900	0.706	0.550	0.850	**0.742**	0.550	0.900	0.702	0.450	0.900

DISCRIMINATION

For each labeling method, we computed Discrimination (the cosine similarity between the topic label and the corresponding TPS) for each topic in the four different cases. Then we fetched the average, min, and max from the Discriminations of all topics in each case

TABLE 3. Discrimination of topic label; the mean, min, and max cosine similarity between the topic label and corresponding discovered topic (OR TPS).

	SIGMOD VEM			SIGMOD Gibbs			APNews VEM			APNews Gibbs		
	Mean	Min	Max	**Mean**	Min	Max	**Mean**	Min	Max	**Mean**	Min	Max
Baseline	0.09937	0.00275	1.00000	0.04665	0.00636	0.67238	0.04768	0.00026	0.69225	0.02291	0.00000	0.26283
LexRank	0.10032	0.00325	0.62277	0.06258	0.00254	0.64343	0.05343	0.00000	0.46150	0.03477	0.00000	0.28211
TextRank	0.13545	0.00789	0.62671	0.05618	0.00400	0.36151	0.06715	0.00000	0.78909	0.03045	0.00000	0.38393
Submodular	**0.06596**	0.01917	0.24103	0.05077	0.01118	0.19281	**0.03788**	0.00182	0.34464	0.03314	0.00285	0.16205
TLRank-C	0.07152	0.01477	0.24125	0.04651	0.00564	0.36907	0.03820	0.00019	0.49470	**0.02181**	0.00000	0.19026
TLRank-G	0.07136	0.01523	0.24517	**0.04621**	0.00473	0.21219	0.03790	0.00020	0.50407	0.02214	0.00000	0.20545

DUPLICATION

The label length and noisy data can easily interfere with the use of similarity to measure the Discrimination between topic labels. In this regard, the present study proposes an intuitive method to measure the Discrimination quality of the topic label considering the number of duplicated sentences in topic label set for each topic.

The method is described in the following steps: (1) for all topic-label-pairs, count the number of duplicated sentences and sum them up to the given "Dup Number"; (2) accumulate the number of sentences contained in all topic labels to the "Total"; (3) let "Dup Ratio" value is that "Dup Number" divided by "Total", if the value is big, it means that there is serious confusion between different topic labels generated by this method, and user can hardly capture the difference between topic labels of each topic.

TABLE 4. Duplication of topic label; for all topic-label-pairs, count the number of duplicated sentences and sum them up to the "Dup Number", accumulate the number of sentences contained in all topic labels to the "Total", and the "Dup Ratio" indicate the ratio of "Dup Number" against "Total".

	SIGMOD VEM			SIGMOD Gibbs			APNews VEM			APNews Gibbs		
	Dup Number	Total	**Dup Ratio**	Dup Number	Total	**Dup Ratio**	Dup Number	Total	**Dup Ratio**	Dup Number	Total	**Dup Ratio**
Baseline	572	76	7.52632	2	195	**0.01026**	70	164	0.42683	2	183	0.01093
LexRank	1398	283	4.93993	88	312	0.28205	202	271	0.74539	50	295	0.16949
TextRank	1790	328	5.45732	60	373	0.16086	596	384	1.55208	58	371	0.15633
Submodular	28	209	**0.13397**	6	270	0.02222	16	257	**0.06226**	0	180	**0.00000**
TLRank-C	426	406	1.04926	4	238	0.01681	60	198	0.30303	2	196	0.01020
TLRank-G	452	410	1.10244	4	260	0.01538	68	202	0.33663	4	213	0.01878

MANUAL

We evaluated the results of four different labeling methods (LexRank, TextRank, Submodular, and TLRank) via manual evaluation in the two different cases, SIGMOD VEM and SIGMOD Gibbs.

We had four human annotators who manually score the topic labels generated. In addition to offering the collection of relevant documents that is necessary to help understand the topics, we provide each human annotator with the top 20 terms of each topic and corresponding topic labels generated by

different methods. It should be noted that the topic labels we provide to each annotator are anonymous. The annotators do not know which labels were generated by which methods.
Besides, we require each annotator to consider the following three aspects when they are scoring the topic labels:

1. the Relevance between the label and the corresponding topic
2. the Coverage of the topic label
3. the Discrimination between the different topic labels.

It needs annotators to score the topic labels in three aspects separately, and then average the three scores for each topic label.
The scores of topic labels generated range from 0 to 3, with 0 representing the worst, and 3 the best. Besides, the floating point number is allowed as a score.
Finally, we average the scores across all topics from the four annotators.

TABLE 6. Our Likert scale of four points.

Score	Description
3	The topic label is perfect.
2	The topic label is reasonable.
1	The topic label is semantically related to the topic, but it is not a good label.
0	The topic label is the worst.

TABLE 7. Manul evaluation, the average of scores between topic label and discovered topic in the cases of sigmod vem and sigmod Gibbs.

	SIGMOD VEM	SIGMOD Gibbs
LexRank	1.567	1.288
TextRank	1.476	1.148
Submodular	1.650	1.094
TLRank	**1.662**	**1.458**

Assessors

four human annotators

Domain

Paper: Topic labeling

Dataset: Scientific literature, News

Problem statement

This article introduces a novel graph-based ranking model (TLRank), to find a meaningful topic label with high Relevance, Coverage, and Discrimination. The model applies a specific strategy that suppresses or enhances the matrix transition probability according to the textual similarity between vertices (sentences) and the characteristics of vertices respectively. Moreover, to boost diversity and enhance performance, TLRank scores the candidate sentences and refrains redundancy of topic labels simultaneously in a single labeling process. In our experiments, the evaluation results showed that the TLRank model significantly and consistently outperformed the prevailing state-of-the-art and classic models in topic labeling task.

Corpus

Origin: SIGMOD and APNews

Nr. of documents: 3016 and 2246

Details:

Document

Pre-processing

stemming, removing stop words

```
@article{he_2019_automatic_labeling_of_topic_models_using_graph_based_ranking,
    author = {He, Dongbin and Wang, Minjuan and Khattak, Abdul Mateen and Zhang,
Li and Gao, Wanlin},
    doi = {10.1109/ACCESS.2019.2940516},
    journal = {IEEE Access},
    pages = {131593-131608},
    title = {Automatic Labeling of Topic Models Using Graph-Based Ranking},
```

```
      volume = {7},
      year = {2019},
      bdsk-url-1 = {https://doi.org/10.1109/ACCESS.2019.2940516}}
```

#Thesis/Papers/FS