Increasing Representative Ability for Topic Representation

Rong Yan, Ailing Tang, Ziyi Zhang
College of Computer Science, Inner Mongolia University
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian
Hohhot 010021, China
Email: csyanr@imu.edu.cn

***Abstract*—As for standard topic model, such as LDA (Latent Dirichlet Allocation), each topic is generally depicted by a weighted word set, where the high-ranked words are deemed more representative. Meanwhile, the probability of each word is considered as the ability to represent the semantic contribution for the topic. However, few efforts are focused on enhancing the representative ability of the topic to support fine grained topic representation. In this paper, we propose a Word Topic Ware (WTW) model to take word inherent diversity characteristic into consideration, in order to screen out and enhance the more representative words for topic representation. Experimental results on three large datasets show that our proposed method can increase the representative ability for topic representation. In addition, our work will positively affect improving the quality of topic content analysis.**

***Index Terms*—Topic analysis, Text representation, Topic model, Latent Dirichlet Allocation.**

I. Introduction

Probabilistic topic model (PTM) family [1] offers a promising solution to discover and extract a mixture of latent topic set that occur in a large document collection. Under the bag-of-word assumption, topic modeling approaches, such as Latent Dirichlet Allocation (LDA) [1], are implicitly capture the document-level word co-occurrence patterns to reveal each latent topic as a multinomial distribution over a weighted word set through the statistical techniques [2]. Meanwhile, a weighted topic set that supposed to be semantic or representative is used to summarize and organize the semantic and hidden structures of documents. Thus, each document is forced to represent as the same specific combination of a topic set. The probability value of the specific word in each topic reflects its representative ability that we called semantic contribution degree. Generally, the bigger probability value of the word represents its bigger semantic capacity ability. Ideally, topics discovered by standard PTMs should be independent from each other under the assumption that the topic proportions are randomly drawn from a Dirichlet distribution. In this paper, we call it ‘topic independence’. Therefore, the explanation of each topic should be single-minded and no ambiguity, that is to say, the topic representation results should maintain the topic reliability. Unfortunately, it is commonly seen that the same word often appears in different topics simultaneously in the real dataset. It makes this topic depiction manner unapparent. Meanwhile, it is very difficult to keep ‘topic independence’. But the truth is that the standard PTM is really difficult to improve this scene. In summary, the reason lies in two aspects, including word frequency and word inherent diversity characteristic.

The word with high-frequency in the document collection is bound to appearing in topic-word distribution with higher rankings for most of the topics due to the ‘bag-of-word’ assumption. However, the semantic contribution degree of this kind of words are not in accord with its representative ability, but weaken topic independence. At the same time, they are general and popular in topic description, which are called the common words. However, as it will be seen later, a topic distribution under which a large number of words with higher probability is not always likely to be insignificant. Though most of these words are filtered out in standard PTMs, which are considered as common stop-words, there are still majority of common words undertake the supported role for the topic description, and it should be reserved for avoiding the semantic loss [3]. And at this point, it is inseparable from the word inherent diversity characteristic. In standard PTMs, such as LDA, the top-ranked word set of the specific topic is generally used to represent the topic description. Thus, the word inherent diversity characteristic lies in the semantic representative ability for the specific topic. As we all known, the outward manifestation of word inherent diversity characteristic lies in the probability of the word in each topic. To that end, intensive efforts have been invested on finding the appropriate method to discriminate the word inherent diversity characteristic [4].

However, few previous studies consider this discrimination from the real status of the same word in different topics, even though it is exactly the common word. To overcome the limitations of previously proposed methods, specifically, we propose a Word Topic Ware (WTW) model for taking account of word inherent diversity characteristic, in order to identify and refine the quantification of the meaningful representative words for topic representation. The main idea comes from the answers of the following two questions extensive: (1) Is the same word that in different topics represent the same semantic

DOI reference number:10.18293/SEKE2022-052

meaning? ($Q1$) (2) Is it correct that each latent topic with single semantic? ($Q2$)

The contribution of this paper is as follows. This paper proposed a WTW model to take word inherent diversity characteristic into consideration, in order to screen out and enhance the more representative words for topic representation. To the best of our knowledge, this is the first time considering the word inherent diversity characteristic for topic analysis.

The remainder of this paper is organized as follows. Section 2 reviews the related work on topic quality analysis. Section 3 analyses the incoherence and bias in PTMs. Section 4 presents our proposed method (WTW). Section 5 describes our experiments. In Section 6, we make a conclusion.

II. Related Work

In this paper we mainly discuss the topic quality problem of PTM, so we review the related research effort in this section. The current automated evaluations of PTM topical quality research mainly focus on two aspects: topic groups quality and individual topic quality.

Much effort has been devoted to automated evaluation approach for topic groups quality, and the mainly metrics include perplexity and topic coherence. The perplexity value lies to estimate the generalization capability of the model fit [5], and the lower value represents a higher performance of the model. However, this metric pays less attention to the semantic interpretability of the words composed a specific topic [6]. The real fact is that the lower perplexity of topics is not necessarily correlated to better coherence of topics, even negatively with topic interpretability [6]. Thus, topic coherence is considered as a supplement metric to evaluate the topic groups quality emphasised on understandability and interpretability [7]–[11]. Topic coherence can be estimated by the semantic similarity of topic words [7], [10], [11] or topic documents [11].

For the PTM family, the determination or selection the number K of the most appropriate latent topics is extremely critical and directly effects the quality of estimated topic set. Up to now, it is still an open-ended problem in topic modeling. While large topic number K means lower descriptor ability of the topic model, as well as intensify the 'forced topic' problem [12]. For being avoid this selection dilemma, topic significance emphasises on evaluating the individual topic quality to serve for the topic groups quality. AlSumait et al. [13] devoted to measuring the distances between three categories 'junk topics' comprised of insignificant word groupings and the legitimate topics. Chang et al. [6] considered the interpretability of a topic as a word intrusion task, and designed topic significance to measure the topic quality in terms of semantic interpretability of the words composed a specific topic. Soon, Lau et al. [14] modified and automated the work of Chang et al. [6] via an improved formulation of Newman [7] based on normalized pointwise mutual information (NPMI). Recently, Chi et al. [15] tried to reranking the top-ranked word set in topic description in order to find the more representative words in topics.

III. Incoherence and Bias

In fact, the existing approaches about topic coherence assumed that the topic coherence correlates with the coherence is based on top-N highest ranked word set assigned to the topic, and we let W denote the word set with N words as topic description, W=$(w_1, w_2, \cdots, w_N)$. As an example, Table. I lists the same topic with the represented word set (20 and 200) discovered by LDA on Reuters-10[1] dataset (category: interest) with topic number K is set to 60.

For the first glance, from Table.I, we can intuitively see that the two descriptions of the topic (top-20: W_1, top-200: W_2) are very similar because of some top-ranked words, such as '*credit*, *finance*', and we can easily give 'Finance' category label to the topic. However, we find that this is not the case through further observation. It is obvious that there are lots of non-relevant words appeared in topic description with larger word number being selected, such as in W_2, which makes the topic incoherent. With further investigation, we find that some words are real relevant to the topic, such as '*bonds*, *treasury*' in W_2. Nevertheless, they are generally considered low semantic contribution for the topic representation due to the low-ranking. Intuitively, we can conclude that the representative ability in W_2 works better. In fact, it will choose a small word number to construct W in standard topic modeling, and we find it is inappropriate. Based on these observations, we make an assumption that whether we can promote the rankings of these 'real' words so as to accomplish the representative ability of the topic description, as well as alleviating the dilemma of the word frequency. Furthermore, the subsequent experiments confirm this assumption.

The fact is that not all topics are high coherent, the incoherent topics will intensify the inexpressibility of topic representability. However, in particular, we note that the representative ability of topic becomes will be clarity when being select a large-scale word set. Meanwhile, it will also increase the risk of the redundancy of the word set. That is to say, the word with low probability has representative ability instead [3]. As for the general PTM, the selected word number N is always set to be a constant, and it has been paid less attention in relevant research.

Besides incoherence, we also note that a second problem may play an important role. In particular, the semantic of the topic may be biased towards the meaning of the top ranked word set with high probability, which covering the dominate semantic of the topic. Meanwhile, the low-ranking word set actually acts as a supplement role because of the low probability. In addition, the ambiguity of the word is another influence factor.

From Table. I, we can observe that the topic description in W_2 has two distinct profiles: 'finance' and 'politics', and the description words of them are intertwined in W_2 even though we know that finance and politics are always inseparable distinctly. However, the truth is that it will let the dominate profile to assign the category of the specific topic [16], [17]. Just like

[1] http://kdd.ics.uci.edu/database/reuters21578/reuters21578.html

TABLE I: Topic description examples for the same topic on Reuters-10 (category: interest) (The first line (W_1) is the original top-20 word set, the second line (W_2) is the original top-200 word set, and the third line (W_3) is the new top-200 word set which re-ranking W_2 using our method).

W_1	credit, rates, card, committee, six, finance, Canadian, group, limit, balances, trade, report, Imperial, subcommittee, Citicorp, transaction, legislation, State, previously, American
W_2	credit, finance, move, fee, group, state, It, banking, hopes, financial, committee, Visa, balance, Express, transaction, Citicorp, charges, statement, ct, amount, reduce, responding, Service, subcommittee, OPTIMA, marketing, Chia, Hockin, quarterly, billing, yearly, market-related, allowed, low, example, threatened, represented, stay, workers, returns, news, levels, sees, factor, pressures, Switzerland, Nova, suspended, important, Italys, departments, aid, positions, nation-wide, association, speculated, curve, expenses, AXP, war, features, issuer, two-to-one, link, Braddock, individual, ones, television, overriding, defend, dominant, cardholders, monthly, delighted, Dallas, Quebec, enaction, expired, touch, Taiwan, middle-class, entitlement, recognised, deciding, resumed, talking, Japan, expected, cut,base, lending, rate, institutions, pct, part, recent, pact, major, industrial, nations, Paris, Finance, Ministry, sources, said, based, revision, Trust, Fund, Bureau, Law, approved, parliament, March, abolishing, minimum, interest, deposits, bureau, channels, funds, government, public, works, official, uses, bodies, Development, Bank, Peoples, Corp, corporations, local, enterprises, usually, moves, tandem, long-term, prime, rates, However, impossible, follow, January, legally, set, ministry, abolish, introduce, resolve, problem, stimulate, domestic, economy, Tuesday, bankers, record, effective, February, suggested, reached, agreement, depositors, postal, savings, system, Posts, Telecommunications, welfare, annuity, Health, Welfare, ministries, trying, determine, market, considered, setting, bureaus, deposit, Coupon, new, year, bonds, minus, percentage, points, likeliest, choice, added, Italian, treasury, annual, coupon, payable, two, issues, certificates, CCTs, four, compared
W_3	credit, finance, move, fee, group, state, transaction, banking, hopes, financial, committee, Visa, balances, Hockin, Express, Citicorp, charges, statement, ct, amount, reduce, responding, Service, subcommittee, OPTIMA, marketing, Chia, quarterly, billing, It, yearly, market-related, allowed, low, example, threatened, represented, stay, workers, returns, cardholders, news, levels, sees, factor, pressures, Nova, suspended, important, departments, aid, welfare, annuity, positions, nation-wide, association, speculated, curve, expenses, AXP, war, features, issuer, link, Braddock, individual, ones, television, overriding, defend, dominant, monthly, delighted, Quebec, expired, touch, Switzerland, middle-class, entitlement, recognised, deciding, resumed, talking, expected, cut, base, lending, rate, institutions, pct, part, recent, pact, major, industrial, nations, Paris, Finance, Ministry, sources, said, based, revision, Trust, Fund, Bureau, Law, approved, parliament, March, abolishing, minimum, interest, deposits, bureau, channels, funds, government, public, works, official, uses, bodies, Development, Bank, Peoples, Corp, Taiwan, Italys, two-to-one,corporations, local, enterprises, usually, moves, tandem, long-term, prime, rates, However, impossible, follow, January, legally, set, ministry, abolish, introduce, resolve, problem, stimulate, domestic, economy, Tuesday, bankers, record, effective, February, suggested, reached, agreement, depositors, postal, savings, system, Posts, Telecommunications, Health, Welfare, ministries, trying, determine, market, considered, setting, bureaus, deposit, Coupon, new, year, bonds, minus, percentage, points, likeliest, Japan, choice, added, Italian, treasury, annual, coupon, payable, two, issues, certificates, CCTs, four, compared

the topic in Table. I, the category will be assigned 'Finance'. In this paper, we take an assumption that each specific topic estimated by PTM has a single sematic interpretation but with multi-sematic aspects or profiles, and that is the motivation of our work.

IV. Methodology

In this section, we explain the detail our methodology for increasing representative ability of the topic. Furthermore, we analyse the effects of incoherence or bias that standard PTMs suffered. Once we determine which sematic profile is being covered by each topic, we propose to promote and identify 'good' words within the topic description.

For a given topic description word set, the aim of this paper is to re-rank the word set in order to obtain a better topic representation. Inspired by the diversification scheme in information retrieval research, we focus on selecting the words that are both relevant to the topic and different from the words already selected. The common principle of diversification is to select as diverse results as possible from a given set of retrieved documents [18]. In our case, we aim to select a semantic representation word set to represent the topic with less redundancy among them as much semantic expression aspects as possible, as well as avoiding the ambiguity of the word. We formulate our identification scheme as a re-ranking task, which is similar to the work of [15]. But different from the work of [15], we devote to identifying and refining the quantification the meaningful representative words for topic representation from the diversification point view for avoiding semantic loss of the topic description.

In this section, we elaborate on a general-based WTW (Word Topic Ware) method to identify the meaningful representative word set for increasing representative ability to a specific topic. Thus, we choose the classical implicit diversification approach to realize. Our approach is based on a similar principle to Maximal Marginal Relevance (MMR) [19], which aim is to take both relevance and redundancy into account for the selected documents.

In this paper, each topic is presented to a top-N most probable word set W from the word-topic distribution Φ to represent the topic t. In this paper, we elaborate MMR algorithm to promote the representative ability of the specific word in topic description, to remedy the semantic contribution degree of the specific word for the topic. In the experiments, we iteratively select and order top-ranked N words with new weight scores by using MMR scheme.

The new semantic contribution degree of each word $w_i \in W$ of the specific topic t is calculated by Eq.(1):

$$weight(w_i,t) = \lambda degree(w_i,t) – (1–\lambda) \max_{w_j \in S} sim(w_i,w_j) \quad (1)$$

where S is the selected set of words in W, $degree(w_i,t)$ determines the original representative degree of each word w_i in topic t, and $sim(w_i,w_j)$ determines the similarity between two words pair. λ denotes the interpolation parameter which controls the tradeoff between the relevance and the diversity. In the experiments, we empirically set λ=0.5.

As for a fixed topic number K, we consider that for the specific topic t_m ($m \in [1,K]$), the representative degree value of the word $degree(w_i,t_m)$ should be with high marginal probability in topic t_m, as well as possessing low marginal probability in other topics, which is calculated by Eq.(2):

$$degree(w_i, t_m) = \phi_{t_m,i} \cdot log \frac{\phi_{t_m,i}}{K\sqrt{\prod_{k=1}^{K} \phi_{k,i}}} \quad (2)$$

where $sim(w_i, w_j)$ denotes the similarity between each node pair, and we use Word2vec [2] to accomplish it.

V. Experiments

In this section, we present the evaluation results that we obtained by applying our proposed framework WTW.

A. Datasets

In this section, we evaluate our topic representation scheme on three large widely datasets: 20NG-bydate [3], Reuter-10 and OHSUMED87-91 [4]. we do use offline topic model so that we can easily extend our work on large collections and avoid the challenge of choosing the topic number.

- **20NG-bydate**: It is a widely used dataset for text classification research. It is highly balanced since each category has about 1000 texts. We use the *bydate* version of this dataset with a total of 18,846 articles that are organized into twenty different categories. This version has been divided into training (60%) and test (40%) set, respectively, and we follow this in the experiment. In addition, we keep the text contained in *Content* field for topic modeling.
- **Reuters-10**: It is another benchmark dataset typically used in the research field of text classification. It contains 21,578 documents in 135 categories. But this dataset is very imbalanced and the variation of category size is quite large. Hence, in the experiment, we left the documents belonging to merely one category and use the 10 largest categories in the dataset (Reuters-10) with a total of 7,285 documents. We use the standard split, 5,228 documents is used as train set and 2,057 documents is used as test set, respectively. In the experiment, we use the *BODY* field for topic modeling.
- **OHSUMED87-91**: It is a widely used dataset for text retrieval and text classification research. It contains five years (1987-1991) relatively short abstracts of references from medical journals in the MEDLINE database with 348,566 documents. In the experiment, we use the *abstract* field for topic modeling and we manually eliminate the invalid documents and left 233,445 documents in fact. We select 119,828 documents used as test and the rest of 113,617 documents used as train set.

For all datasets, we removed HTML tags, stop words, rare words and the word with length less than two or occur in less than five documents. In addition, in 20NG-bydate, the *Content* field of some documents are empty, and we cull these documents manually in the experiment. Table. II gives the detailed statistics information of three datasets.

TABLE II: Statistics of the datasets.

Dataset	train word number	test word number
20NG-bydate	49,446	34,913
Reuters-10	33,340	22,880
OHSUMED87-91	402,058	441,400

B. Experimental Setting

In order to evaluate our approach, we require a topic model. We apply directly standard LDA to obtain the initial topic description results from each dataset. In the experiments, we use Gibbs sampler to generate the topic-word distributions Φ, and the iteration number of Gibbs sampler adopt a fixed value 1000. During modeling training, the Dirichlet hyperparameters α and β are set to 0.1 and 0.01, respectively. For each topic t be denoted as a list with top-ranked N words, and the value of N ranges from 10 to 400, step is set to 5.

C. Results and Analysis

Topic coherence metric is a popular automatical metric to evaluate the coherence of the topics learnt by topic models. However, it is unfit for our work due to the word-frequency essential peculiarity of the language being used. In order to analysis how effect of the topic representative ability is, with the increasing of the topic representation word number N being selected, we take the *perplexity* to evaluate the performance of topic representation.

As shown in Fig. 1, we exhibit the lowest value of *perplexity* comparisons (*top-N* 20 and *top-N* 200) on three datasets.

From Fig. 1, we can observe that the lowest value of perplexity has fluctuation to some extent with the selected word number increasing. Experimental results shows that the topic representative ability is instable due to the scale of the top-ranked word set. It is the proof of the influence of word frequency on text modeling. Thus, we can conclude that it is an objective existence phenomenon in general PTMs, and it is also proved that the necessity and significance for increasing the representative ability to topic representation.

In the third line of Table. I shows the new topic description W_3 for the same topic examples. For the first glance, the description of W_3 is similar to W_2, that is to say, they indeed depict the same semantic.

On the one hand, when we make a detailed observation, we find that there are some words, such as '*transaction*, *Hockin*', appeared in W_3 with higher rankings compared to W_2, which makes the semantic representative ability of W_3 better than W_2.

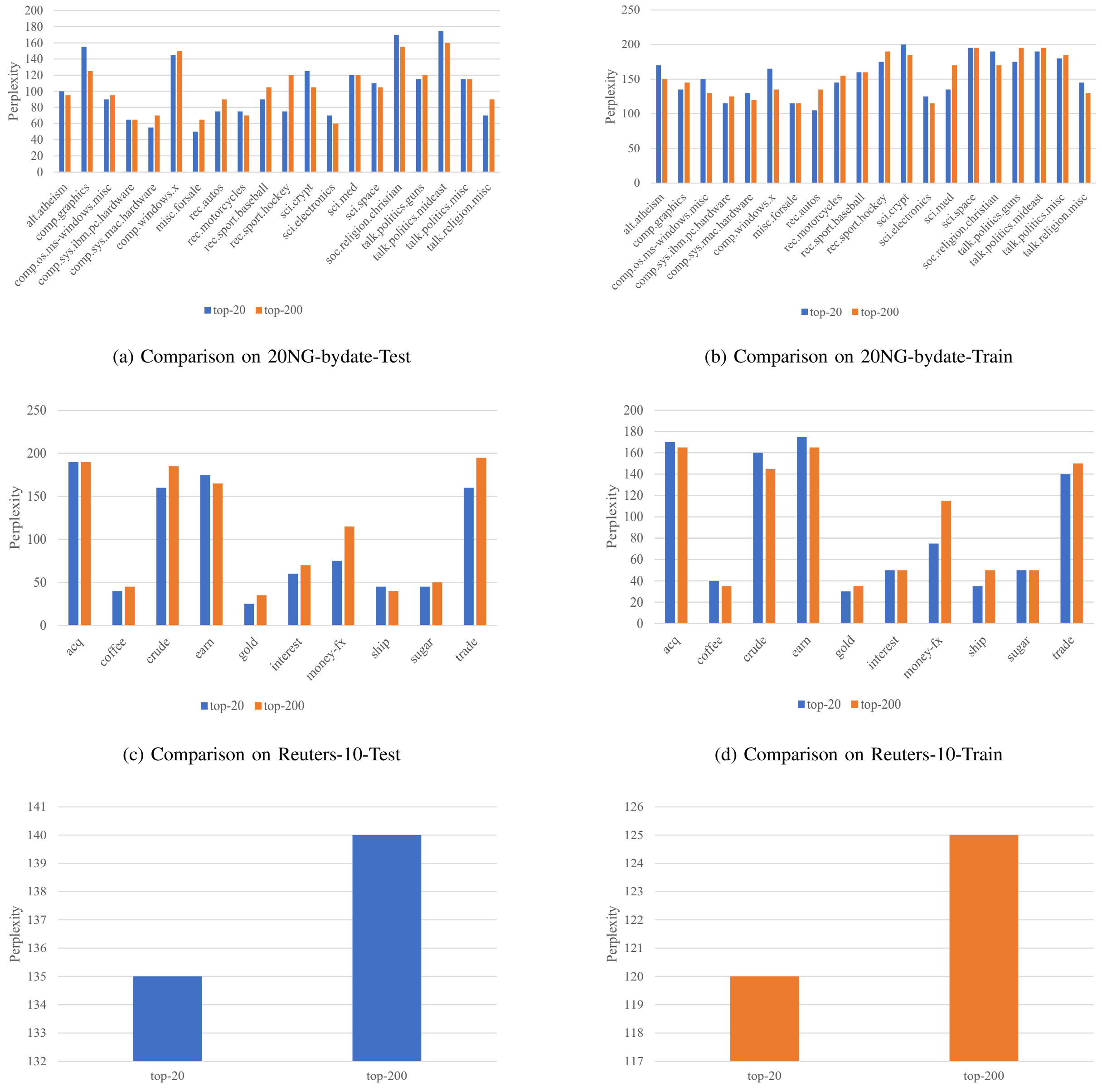
On the other hand, we find that the word with high frequency, such as the word '*It*' in W_2, descend the rankings, i.e, its semantic contribution degree is descend because the little relevance with the selected word set, as well as descending the influence of its word frequency.

Furthermore, in the experiments, we find that with the value of N ascending, the lower-ranked word set has little semantic contribution for topic representation even though they are ranked using our method, and the representative ability of

[2] http://code.google.com/p/word2vec

[3] http://qwone.com/~jason/20Newsgroups/

[4] http://mlr.cs.umass.edu/ml/machine-learning-databases/ohsumed/



(a) Comparison on 20NG-bydate-Test

(b) Comparison on 20NG-bydate-Train

(c) Comparison on Reuters-10-Test

(d) Comparison on Reuters-10-Train

(e) Comparison on OHSUMED87-91-Test

(f) Comparison on OHSUMED87-91-Train

Fig. 1: Examples of perplexity value comparisons on three datasets.

the topic will be confused sometimes. We also find that the value of N is highly depend on the quality of the dataset. If the category of the dataset is relatively distinct, the influence of the N value for the topic representative ability is slight, such as 20NG-bydate and Reuters-10. On the contrary, as for OHSUMED, the influence of the N value for the topic representative ability is volatile.

VI. Conclusion

In this paper, we focus on screening out the more representative words for topic representation. We investigate a re-ranking method WTW to evaluate the representative ability of words over different topics. This work will enhance the ability to support fine grained topics representation for text content mining tasks.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (Grant No. 61866029).

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, mar 2003.

[2] W. Xuerui and M. Andrew, “Topics over time: A non-markov continuous-time model of topical trends,” in *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining*, ser. KDD’06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 424–433.

[3] R. Yan and G. Gao, “Topic analysis by exploring headline information,” in *Proceedings of 21st International Conference on Web Information Systems Engineering*, ser. WISE’20, H. Zhisheng, B. Wouter, W. Hua, Z. Rui, and Z. Yanchun, Eds., vol. 12343. Chem: Springer, oct 2020, pp. 129–142.

[4] Q. Chen, X. Guo, and H. Bai, “Semantic-based topic detection using markov dcision processes,” *Neurocomputing*, vol. 242, pp. 40–50, jun 2017.

[5] W. H. M., M. Iain, S. Ruslan, and M. David, “Evaluation methods for topic models,” ser. ICML’2009, vol. 382. New York, NY, USA: Association for Computing Machinery, jan 2009, pp. 1105–1112.

[6] C. Jonathan, B.-G. Jordan, G. Sean, W. Chong, and B. D. M., “Reading tea leaves: How humans interpret topic models,” in *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, ser. NIPS’09. Red Hook, NY, USA: Curran Associates Inc., 2009, pp. 288–296.

[7] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, “Automatic evaluation of topic coherence,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT ’10. USA: Association for Computational Linguistics, 2010, pp. 100–108.

[8] D. Newman, E. V. Bonilla, and W. Buntine, “Improving topic coherence with regularized topic models,” in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, ser. NIPS’11. Red Hook, NY, USA: Curran Associates Inc., 2011, pp. 496–504.

[9] M. David, W. H. M., T. Edmund, L. Miriam, and M. Andrew, “Optimizing semantic coherence in topic models,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’11. USA: Association for Computational Linguistics, 2011, pp. 262–272.

[10] D. O’Callaghan, D. Greene, J. Carthy, and P. Cunningham, “An analysis of the coherence of descriptors in topic modeling,” *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.

[11] D. Korenčić, S. Ristov, and J. Šnajder, “Document-based topic coherence measures for news media text,” *Expert Systems with Applications*, vol. 114, pp. 357–373, 2018.

[12] X. Li, J. Ouyang, Y. Lu, X. Zhou, and T. Tian, “Group topic model: Organizing topics into groups,” *Information Retrieval*, vol. 18, no. 1, pp. 1–25, feb 2015.

[13] A. Loulwah, B. Daniel, G. James, and D. Carlotta, “Topic significance ranking of lda generative models,” in *Proceedings of Joint European Conference on Machine Learning (ECML) European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 67–82.

[14] J. H. Lau, D. Newman, and T. Baldwin, “Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality,” in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, jan 2014, pp. 530–539.

[15] J. Chi, J. Ouyang, C. Li, X. Dong, X. Li, and X. Wang, “Topic representation: Finding more representative words in topic models,” *Pattern Recognition Letters*, vol. 123, pp. 53–60, 2019.

[16] R. Yan, Q. Chen, and G. Gao, “Dataless text classification with pseudo topic representation,” in *Proceedings of 32nd International Conference on Tools with Artificial Intelligence*, ser. ICTAI’20, nov 2020, pp. 1255–1259.

[17] D. Zha and C. Li, “Multi-label dataless text classification with topic modeling,” *Knowledge and Information Systems*, vol. 61, no. 1, pp. 137–160, oct 2019.

[18] R. Santos, C. Macdonald, and I. Ounis, “Search result diversification,” *Foundations and Trends in Information Retrieval*, vol. 9, no. 1, pp. 1–90, mar 2015.

[19] J. Carbonell and J. Stewart, “The use of mmr, diversity-based reranking for reordering documents and producing summaries,” in *Proceedings of the 21st International ACM Conference on Research and Development in Information Retrieval*, ser. SIGIR’98, Melbourne, Australia, aug 1998, pp. 335–336.