Automatic Labelling of Topics via Analysis of User Summaries

Lishan Cui[1], Xiuzhen Zhang[1⋆], Amanda Kimpton[2], and Daryl D'Souza[1]

[1] School of Science (Computer Science & Information Technology)
[2] School of Health & Biomedical Sciences
RMIT University, Melbourne 3001, Australia
{lishan.cui,xiuzhen.zhang,amanda.kimpton,daryl.dsouza}@rmit.edu.au

Abstract. Topic models have been widely used to discover useful structures in large collections of documents. A challenge in applying topic models to any text analysis task is to meaningfully label the discovered topics so that users can interpret them. In existing studies, words and bigram phrases extracted *internally* from documents are used as candidate labels but are not always understandable to humans. In this paper, we propose a novel approach to extracting words and meaningful phrases from *external* user generated summaries as candidate labels and then rank them via the Kullback-Leibler semantic distance metric. We further apply our approach to analyse an Australian healthcare discussion forum. User study results show that our proposed approach produces meaningful labels for topics and outperforms state-of-the-art approaches to labelling topics.

Keywords: topic model labels, user generated content, dependency relation parser

1 Introduction

Topic models are useful tools for analysing large collections of documents [13, 17, 20, 22]. With a topic model, each latent *topic* is a multinomial distribution over words, and each document is described as a mixture of latent topic distributions. Several topic models have been proposed in the literature [1, 5], including the popular Latent Dirichlet Allocation (LDA) and other recent developments (e.g., [2]). An important problem is to label the topics produced by topic models so that the topics are understandable and interpretable to humans [7, 8, 12, 15]. The topic below (Topic 0 of Table 4) is generated by the non-parametric LDA [2] from 623 documents (stories) sourced from the Patient Opinion Australia (`https://www.patientopinion.org.au/`), a web forum for patients to post their stories about healthcare and freely express their opinions. Further details of such use of the LDA model appear in [2].

october; appointment; outpatient; referral; letter; answer; confirmed; september; ulcer; telephoned.

⋆ Corresponding author.

The topic is represented by the top ten words as the topic label, ordered by their marginal probability in the topic [1, 5]. However such a conventional topic label is difficult for humans to make sense and interpret the meaning [7, 8, 11, 13, 19]. Specifically words *october* and *september* refer to the temporal information whereas *appointment* and *referral* refer to communication; the single words make it hard for people to infer the semantics of the topic "appointment letter". It is therefore critically important to assign semantically meaningful labels to topics so that humans can interpret and understand the topics.

It has been shown in the literature that phrases as labels for topics are more understandable to human beings [11, 13, 19]. However, existing solutions make the implicit assumption that n-gram phrases are good candidate topic labels. Mei *et al.* [12] used bi-gram phrases from the document collection as candidate labels. Recognising the limitation that words appearing in the document collection cannot represent topics at higher semantic levels, Lau *et al.* [7] proposed to extract words and phrases from external data sources like Wikipedia as candidate topics labels. Still, n-gram phrases extracted from the general knowledge base Wikipedia, are not always meaningful labels for specific domains.

On the other hand, with the development of Web 2.0 technology, user-generated content becomes widely available on the web. User posted documents and comments appear in various discussion forums as well as community question-answer sites like Yahoo!Answers and Stack Overflow. User comments in web forums have been employed for computing seller-buyer trust for E-commerce applications [23], news recommendation [9], and predicting the popularity of online articles [18].

In this paper we propose to utilize user generated summaries for labelling topics for a document collection. The user generated summaries in the same domain as the document collection are ostensibly more relevant to the domain than general external knowledge sources like Wikipedia. Moreover, the *external* user-generated summaries often contain words and phrases at higher semantic levels than words from the *internal* document collection. To the best of our knowledge our work is the first of its kind to take advantage of user-generated content for labelling topic models for document collections.

Our approach, Labelling By Summaries (LBS), can produce meaningful topic labels easy to interpret by humans. Looking again at the topic in the previous example, LBS generates the following words and phrases as its generated label:

```
administration; appointment letter
```

Note that `administration` is a single-word expression frequently appearing in user summaries but not in the topic word list. Importantly `administration` summarises nicely the administrative aspect details expressed by several topic words including *october*, *outpatient*, *referral*, *answer*, *confirmed*, *september* and *telephoned*. On the other hand `appointment letter` is a noun phrase frequently appearing in user summaries generated by the dependency relation parser [4] and it summarises meaningfully the topic words *letter* and *appointment*.

The research questions we address are as follows:

– How to generate meaningful phrases from user summaries as candidate labels for topics?
– For a topic, how to measure the semantic association between candidate labels and the topic to rank the candidate labels?

We make several contributions for automatic labelling of topic models. First we propose to apply the NLP dependency relation parser [4] and frequency-based noise filtering to generate meaningful phrases from user summaries as candidate topic labels. To rank candidate labels for a topic, we further propose a metric to evaluate the semantic association between a topic and the candidate labels based on Kullback-Leibler (KL) divergence [6]. We formulate the problem as minimising the KL divergence between a topic and the candidate labels. We apply LBS with the non-parametric LDA topic model proposed by Buntine *et al.* [2] to summarise documents in the healthcare domain – the Patient Opinion Australia (POA) web forum (details in Section 4). We design a user study to evaluate labels for topics. Results show that, compared with state-of-art approaches [8, 12], LBS can generate more meaningful labels that can help understand and better interpret topics.

2 Related Work

Existing studies on automatic labelling for topic models can be categorised into internal and external approaches based on the candidate labels used. The internal approaches use topic words that are sourced and generated by topic models from the document collection [1, 5, 8], whereas external approaches use word sources that are external to the topic words and could be extracted from data sources like Wikipedia or document collection [7, 10, 12]. The standard approach for labelling topics is to use the top N (typically N=10) topic words ranked by their marginal probability [1, 5, 8]. However, it is recognised that the list of top words very often do not present coherent semantics [14, 16] and are difficult for humans to understand and interpret the meaning of topics. To enhance the topic interpretation, Lau *et al.* [8] proposed to select the best topic words to label topics. Several measures are proposed to re-rank the top ten topic words to select the best label for the topic. Still, the assumption is that the best label for a topic can be found among the internal topic words generated by topic models.

In external approaches there are typically two steps involved: generating candidate labels for topics and ranking candidate labels. Lau *et al.* [7] proposed an approach to automatically label topics from topic models via generation of candidate labels from external data sources (Wikipedia) and supervised learning for ranking the candidate labels. The candidate labels generated can be single terms or phrases from Wikipedia. Mei *et al.* [12] proposed an approach to automatically label topics, comprising generating candidate labels by extracting bi-grams or noun chunks from the document collection. To rank the candidate labels for a topic, topics and labels are represented as distributions of words and the KL divergence is used to measure their semantic distance. Magatti *et al.* [10] proposed a method for labelling topics induced by hierarchical topic models. Their

candidate labels are based on the Google Directory hierarchy and the labelling approach relies on a pre-existing ontology and the associated class labels. As a result the approach can only be applied in limited applications.

Although not directly related to topic labelling, Chang *et al.* [3] were one of the first to raise the question of labelling topic models for human understanding. They identified the notion of *intruder* words. Such words (in the top topic words) impede or compromise human understanding of topic meaning and are inconsistent with the semantic meaning in inferred topics.

In the literature, the study closest to our approach is that of Lau *et al.* [7] and Mei *et al.* [12], which are both external approaches. Although our approach in spirit is also an external approach, it has several significant differences from these approaches, in terms of both generating candidate labels and ranking them for labelling: (1) We generate candidate labels from different sources and via a different approach; we employing user summaries and apply dependency relation parsing to generate meaningful candidate labels. (2) We rank candidate labels differently; we represent topics and labels as distributions of documents rather than words to measure the semantic association (distance) between topics and candidate labels.

3 Labelling Topics with User Summaries

The problem of topic model labelling can be decomposed into two sub-problems: generating candidate labels from user summaries and ranking candidate labels. We describe these steps below.

3.1 Generating candidate labels by dependency relation parsing

User summaries on web forums are generally short, containing single words and short phrases, but also long phrases and sentences. To label the topics for a document collection, we aim to generate user interpretable words and phrases as candidate labels by analysing collectively all user summaries. Words as well as phrases of two or three words frequently appear in user summaries. Examples of such words and phrases include "hospital", "midwife", "comfort", "aged care" and "first class care". An important observation is that *the user generated succinct expressions of single words or short phrases are generally nouns or noun phrases and express meanings easily understandable by humans.* In our application of this observation, we set short phrases to contain at most three words that appear frequently (frequency of at least four by default) in user summaries as candidate labels.

For longer summaries that are sentences and phrases of more than three words, we apply the typed dependency parser [4] to analyse summaries and generate candidate labels. The typed dependency relation parser has been shown to be an effective tool for analyzing short informal text [23]. With typed dependency relation parsing, a sentence is represented as a set of dependency relations between pairs of words in the form of *(head, dependent)*, where content words

Table 1. Examples for dependency relation parsing

User Summary	POS Tag	Typed Dependency	Candidate Label
help from nurses on maternity ward	help/VB, from/IN, nurses/NNS, on/IN, maternity/JJ, ward/NN	root(ROOT-0, help-1), prep_from(help-1, nurses-3), amod(ward-6, maternity-5), prep_on(nurses-3, ward-6)	maternity ward
medical staff at hospital	medical/JJ, staff/NN, at/IN, hospital/NN	amod(staff-2, medical-1), root(ROOT-0, staff-2), prep_at(staff-2, hospital-4)	medical staff

are chosen as heads, and other related words depend on the heads. Table 1 shows some examples of candidate labels generated from user summaries by the dependency relation parser. The numbers after each word indicate the position of the word in the sentence. The word pairs that form dependency relations are listed in the third column, where *root*, *prep_from*, *amod* and *prep_on* are dependency relation types. For example *amod(ward-6, maternity-5)* represents that the word pair "maternity ward" forms an *adjective modifying* dependency relation. Details of the dependency types are described in [4].

After dependency relation parsing, dependency relations for adjacent words, and where at least one word is a noun, are selected as candidate labels. This strategy lends support to the idea that the noun in a dependency relation generally expresses content and adjacent words indicate that the relation forms noun phrases. For example in Table 1, for the user summary "help from nurses on maternity ward", nouns *nurses* and *ward* are identified, from the POS tag. Among the dependency relations on these nouns, which are *prep_from(help-1, nurses-3)*, *prep_on(nurses-3, ward-6)* and *amod(ward-6, maternity-5)*, *maternity ward* is extracted as a candidate label since *ward* is a noun and the two adjacent words *(maternity-5, ward-6)* form a noun phrase.

Noun phrases generated by the dependency relation parser form candidate labels. These candidate labels may contain errors. We apply a simple yet effective strategy to remove noises — candidate labels occurring in less than a frequency threshold (by default 1%) in the user summaries are considered as noise and filtered out.

3.2 Labelling topics with candidate labels

According to Mei et al. [12], a *topic* θ is a probability distribution of words $p(w \in V|\theta)$ where V is a vocabulary, and $\sum_{w \in V} p(w|\theta) = 1$. A topic label, or simply *label*, l for a topic θ is a sequence of words or phrases that is semantically meaningful and covers the latent meaning of θ. The *relevance score* of a label to a topic $s(\theta, l)$ measures the semantic similarity between θ and l. Note that topics are generated from documents, and the candidate labels are generated from user summaries of the documents.

We present our metric to measure the semantic association between candidate labels generated from user summaries and the topics produced by topic mod-

Table 2. The number of stories with different types of user summaries

Type of User Summaries	# of Stories (%)
Positive & negative user summaries	135 (21.67%)
Only positive summaries	333 (53.45%)
Only negative summaries	125 (20.06%)
No summaries	30 (4.82%)
Total	623

els, in order to rank candidate labels for topics. We employ Kullback Leibler (KL) divergence [6] as the measure of the difference between two probability distributions.

On social media websites like discussion forums, user summaries or comments are for specific documents. As a result phrases generated from user summaries as candidate labels are associated with a multinomial distribution of documents. On the other hand, from the document-topic probability distribution matrix a topic can also be represented as a distribution of documents. Given θ and l, let P_θ and Q_l denote respectively the document probability distribution for θ and l. The relevance score of a candidate label, l, for topic, θ, is defined as the opposite of the KL divergence between the document distributions of l and θ:

$$s(\theta, l) = -D_{KL}(P_\theta || Q_l) = -\sum_i P_\theta(i) * ln \frac{P_\theta(i)}{Q_l(i)}\,. \tag{1}$$

In the above definition, i denotes the documents under consideration.

Given a topic θ and a set of candidate labels, $s(\theta, l)$ is computed for each candidate label, and the top k labels with the highest relevance scores for topic θ become the labels for the topic. By default $k = 2$, as it is shown in the literature that two phrases are preferred by human annotators as labels and generally have high semantic consistency [12].

4 Analysing the Patient Opinion Australia Stories

The Patient Opinion Australia (POA) is a web forum wherein patients post their stories about health services. In addition to user stories (documents), the POA website allows patients to post summaries (called "Story summary"), that are divided into two types: positive summaries are entered in the field "What is Good", and negative summaries are entered in the field "What could be improved". As shown in Table 2, documents may have one or both types of summaries.

We crawled data from the POA website to evaluate our approach, Labelling By Summaries (LBS). We applied the non-parametric LDA topic model [2] to generate 50 topics for collected documents. In later discussions, these 50 topics are numbered as 0..49 as the results of the topic modelling output. Following the topic modelling convention, each topic is represented by the top ten topic words with high marginal probability.

Table 3. Five sample user summaries

	Summary
Doc. 5	GP; bad care; disgusting; emotional wellbeing; fobbed off; lack of compassion; pain; pain relief; rude consultant; understanding.
Doc. 24	addressing my concerns; earlier surgical appointment; good care; ***private gynaecologist listening to my concerns***.
Doc. 36	admission; anaesthetist; good experience; greeted warmly; ***I felt I was in good hands***; nurse attitude; nurse care.
Doc. 145	initial service; My GP; ***staff in surgical ward***; doctor attitude; doctor care; lack of understanding; ***treatment in emergency department***.
Doc. 567	***communication within RSL Care***; consistancy; ***employing people who know how to clean properly***.

We benchmarked LBS against three other approaches:

- *Top Words*: the conventional approach that labels topics by the top k (k=10) topic words with high marginal probability.
- *Top Word Re-rank*: the straight forward approach Re-rank where top topic words are re-ranked by their frequency in user summaries.
- *PHR*: the approach by Mei *et al.* [12] where the top 1000 bi-gram phrases from documents are candidate labels.

4.1 Data, topics and manual labels

We crawled 623 stories and 593 summaries from the POA website in early August 2014. The number of stories that contained different types of user summaries is shown in Table 2. Stories can have positive ("What is good") or negative ("What could be improved"), or mixed user summaries. A small portion (4.82%) did not have any user summaries. In terms of the language features, summaries include single words, short or long phrases, and complete sentences. The average length of user summaries was four words, while the longest summary contained 29 words. Five sample user summaries are listed in Table 3, where long phrases of at least four words and sentences are highlighted.

It is well recognised that topics produced from topic models may sometimes be statistically important but not convey much semantically consistent information [14, 16]. These topics should be discarded rather than labelled. We aim to automatically remove the semantically incoherent topics based on user summaries. Intuitively, individual words in phrases from user summaries are semantically associated. So, topics that break a phrase in user summaries – not every word in a phrase appears as a top topic word for a topic – indicates that the topic lacks semantic coherence. Generally the more phrases broken by a topic the less likely that the topic is semantically coherent. To control noise in user summaries, only frequent phrases (frequency is set to 4) in user summaries are considered. In total we extracted 94 two-word phrases from user summaries.

Table 4. Sample topic labels by different approaches

Topic	Top 10 topic words	LBS	Top Words	Top Word Re-rank	PHR	Human
0	october; appointment; outpatient; referral; letter; answer; confirmed; september; ulcer; telephoned	administration; appointment letter	appointment; letter	appointment; referral	appointment within; first appointment	appointment letter; referral letter
7	neck; wrist; thumb; treatment; brace; report; painkiller; agony; file; skin	medical record; post surgery	neck; wrist	x-ray; neck	neck brace; neck pain	medical record; surgery care
8	healthy; program; kate; lifestyle; men; programme; encouraging; healthier; shed; meet	lifestyle program; continuity of care	healthy; program	program; lifestyle	lifestyle program; healthy lifestyle	healthy lifestyle; lifestyle program
22	wife; hearing; urgency; ENT; surgeries; genetic; westmead; loss; costs; aids	hospital; initial diagnosis	wife; hearing	ENT; loss	hearing loss; hearing aids	ENT emergency; emergency service

Applying the frequent user phases to filter incoherent topics, 9 out of 50 topics were filtered out. The remaining 41 topics were then manually labelled by five human annotators as follows: for each topic, the top 10 candidate labels generated by each of the approaches, LBS, Top Words (or Top Word Re-rank), and PHR, were pooled together. As a result approximately 30 (there may be duplicates among the three models) candidate labels (words or phrases), were randomised and suggested to annotators for manual labelling. The full-text stories (documents) were also presented at the side to help understand the context. It is hard for annotators to unanimously choose the same best label for a topic and so each annotator was asked to choose the top two candidates for each topic and the top two candidates frequently chosen by annotators were deemed the label for the topic. In equal frequency cases, the topic label was resolved by a

Table 5. Ratings by assessors of topic labels by different approaches. Rating 3 indicates "very good label" whereas 0 indicates "completely inappropriate label".

Topic	Approaches	Ratings					Avg Rating
0	LBS	3	2	3	1	3	2.4
	Top Words	2	2	2	2	2	2.0
	Top Word Re-rank	2	2	2	2	2	2.0
	PHR	1	1	0	0	1	0.6
7	LBS	3	1	3	3	3	2.6
	Top Words	2	1	1	2	2	1.6
	Top Word Re-rank	2	2	1	2	1	1.6
	PHR	1	2	2	1	1	1.4
8	LBS	3	3	3	3	3	3.0
	Top Words	2	2	2	2	2	2.0
	Top Word Re-rank	2	1	2	2	1	1.6
	PHR	3	3	3	3	3	3.0
22	LBS	2	2	1	1	2	1.6
	Top Words	0	0	0	0	1	0.2
	Top Word Re-rank	2	1	2	2	1	1.6
	PHR	2	3	3	3	1	2.4

discussion by the first two authors. Table 4 presents some sample topics and their manual labels (indicated as Human). It is worth noting that all human-suggested labels are phrases, which is consistent with previous finding that phrases are preferred labels for topics and easier for humans to understand [7, 12].

4.2 LBS labels: quantitative analysis

We benchmarked LBS against other approaches for automatic labelling. For all approaches, the top two ranked words/phrases were used to label each of the 41 topics. To quantitatively measure the quality of automatic labels, we asked human assessors to rate the labels by each approach. To avoid bias due to presentation, we presented the four types of labels by different methods in random orders. To help a human assessor to interpret the topics, the top three documents with the highest marginal probability for the topic were also presented. Given the labels for a topic, an assessor was asked to rate the labels according to the quality of the labels. The rating levels are: *3: very good label; 2: reasonable label; 1: somewhat related, but bad as a topic label; 0: completely inappropriate topic label.* Each topic was assessed by five assessors who were volunteers, different from the previous assessors for human labelling. For each topic, the ratings of five assessors were averaged to compute the rating for each approach.

We first examined the variance of ratings for topic labels across assessors. Table 5 lists the ratings by assessors of topic labels by different approaches for the four sample topics in Table 4. The table shows that, although assessors rated an approach differently for different topics, they rated an approach surprisingly consistently for a specific topic. For example, PHR received consistent ratings

Table 6. Quantitative evaluation of LBS labels

	LBS	**Top Words**	**Top Word Re-rank**	**PHR**
Avg rating	1.815	1.322	1.341	1.356
p-value	-	0.0016	0.0009	0.0138
correct labels	34	26	29	24

of two 0s or three 1s for Topic 0 but consistent ratings of five 3s for Topic 8. So, we can safely say that the ratings by assessors are reliable. We next discuss the overall performance of all approaches for 41 topics.

We compared the ratings for LBS with each of the three other approaches for 41 topics using the Wilcoxon signed-rank test [21]. The overall average ratings for each approach and the Wilcoxon signed-rank test p-value for the other three approaches are shown in the first two rows of Table 6. LBS has the highest average of 1.815, indicating that human assessors judge the labels as in between *somewhat related* to *reasonable label*. All three other approaches have significantly lower average ratings ($p < 0.05$) than LBS. In other words the labels are deemed by assessors as nearly *somewhat related but bad as a topic label*. Note that PHR, using phrases as labels, has better average ratings overall than the other two approaches, which are both based on words. This result confirms that using phrases as topic labels are more understandable to humans [11, 13, 19]. On the other hand, as PHR generates topic labels from bi-grams in documents rather than from user summaries, the phrases are not always meaningful and this is reflected in their significantly lower ratings by assessors than those for LBS.

The automatic labels of all approaches are also compared with the human labels. The last row of Table 6 shows the number of correct labels generated by automatic approaches compared with the human-generated labels for the 41 topics. We asked human assessors to judge the labels by each approach compared with the human-generated labels. Each label was assessed by five assessors who were volunteers, different from the previous assessors. A label is considered correct if at least three assessors vote the label as correct. Clearly LBS has the largest number of 34 topics with correct labels, indicating a high level of consistency with the human-generated labels. In contrast, top topic words and PHR has the lowest number of 26 and 24 topics with correct labels, respectively, and a low level of consistency with the human-generated labels.

4.3 LBS labels: qualitative analysis

Table 4 lists automatic labels generated by different approaches compared with the human-generated labels for some sample topics. We can see that the automatic labels generated by LBS very often do not appear at all in the word list for topics. Interestingly the labels generated by LBS can very often conceptually generalise the user phrases captured by topics. For example for Topic 0, the LBS label "appointment letter" generalises conceptually the topic words "referral" and "letter". This result can be attributed to the use in LBS of the dependency

relation parser of user-generated summaries, which can produce meaningful candidate labels that are more understandable to humans.

Comparing the labels by LBS with the manual labels, it is clear from Table 4 that the LBS generated phrase labels can more accurately capture the meaning of topics. For example for Topic 7, the LBS labels of "medical record" and "post surgery" closely match the human-generated labels of "medical record" and "surgery care". As is also shown in Table 6, LBS has the highest number of labels matching the human-generated labels.

Table 4 also sheds light on the performance of automatic labels by other approaches. Take again Topic 7 as an example. Only LBS correctly labelled it as "medical record; post surgery", while the labels by other approaches are not related to "medical reports; surgery care", but rather only describe detailed aspects of medical reports and surgery care. The word-based Top Words and Top Word Re-rank approaches label the topic as "x-ray", "neck", "wrist", which is difficult for humans to interpret. The PHR phrase approach can generate more meaningful labels like "neck pain", which are phrases extracted from the document collection but are not high level summarisation of the document content.

5 Conclusions

The problem of automatic labelling for topic models with human understandable labels is crucial for the wide applications of topic modelling for many text analysis tasks. Existing work on automatic labelling for topic models either relies on the document collection itself or external web-based resources. In this paper we proposed a novel approach of making use of user summaries to label topic models. We made several contributions: (1) We proposed the novel application of dependency relation parsing to extract meaningful and human understandable phrases from user summaries as candidate labels. (2) We proposed KL divergence to measure the semantic similarity between candidate labels and topics and rank candidate labels. (3) We applied our automatic topic labelling approach to analyse a real-world user discussion forum for healthcare. Results show that our approach can generate meaningful and human interpretable topic labels.

For future work we will design further refinements to our approach in terms of deploying user input. It would also be interesting to extend our approach to other web-based forums for products and services.

Acknowledgements. Thanks go to Mr. Bin Lu for crawling the Patient Opinion Australia website and preliminary analysis of the data. Thanks also go to Associate Professor Michael Greco of Patient Opinion Australia for explanation of the data.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3, 993–1022 (2003)

2. Buntine, W.L., Mishra, S.: Experiments with non-parametric topic models. In: Proc. KDD 2014. pp. 881–890. ACM (2014)
3. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Proc. Advances in neural information processing systems. pp. 288–296 (2009)
4. De Marneffe, M.C., Manning, C.D.: The stanford typed dependencies representation. In: Proc. the workshop on Cross-Framework and Cross-Domain Parser Evaluation. pp. 1–8. Association for Computational Linguistics (2008)
5. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences of the United States of America 101(1), 5228–5235 (2004)
6. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics 22(1), 79–86 (1951)
7. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic labelling of topic models. In: Proc. ACL HLT 2011. pp. 1536–1545 (2011)
8. Lau, J.H., Newman, D., Karimi, S., Baldwin, T.: Best topic word selection for topic labelling. In: Proc. COLING 2016: Posters. pp. 605–613 (2010)
9. Li, Q., Wang, J., Chen, Y.P., Lin, Z.: User comments for news recommendation in forum-based social media. Information Sciences 180(24), 4929–4939 (2010)
10. Magatti, D., Calegari, S., Ciucci, D., Stella, F.: Automatic labeling of topics. In: Proc. 9th International Conference on ISDA. pp. 1227–1232. IEEE (2009)
11. Mei, Q., Liu, C., Su, H., Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: Proc. WWW 2006. ACM (2006)
12. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proc. KDD 2007. pp. 490–499. ACM (2007)
13. Mei, Q., Zhai, C.: A mixture model for contextual text mining. In: Proc. KDD 2006. pp. 649–655. ACM (2006)
14. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proc. EMNLP 2011. pp. 262–272 (2011)
15. Newman, D., Karimi, S., Cavedon, L., Kay, J., Thomas, P., Trotman, A.: External evaluation of topic models. In: Proc. ADCS 2009. pp. 1–8 (2009)
16. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Proc. NAACL HLT 2010. pp. 100–108 (2010)
17. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: Proc. KDD 2004. pp. 306–315. ACM (2004)
18. Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., de Amorim, M.D., Fdida, S.: Predicting the popularity of online articles based on user comments. In: Proc. Int. Conf. on Web Intelligence, Mining and Semantics. p. 67. ACM (2011)
19. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proc. KDD 2006. pp. 424–433. ACM (2006)
20. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proc. SIGIR 2006. pp. 178–185. ACM (2006)
21. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics bulletin 1(6), 80–83 (1945)
22. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proc. CIKM 2001. pp. 403–410 (2001)
23. Zhang, X., Cui, L., Wang, Y.: Commtrust: Computing multi-dimensional trust by mining e-commerce feedback comments. IEEE Transactions on Knowledge and Data Engineering 26(7), 1631–1643 (2014)