Enhancing Topical Word Semantic for Relevance Feature Selection

Abdullah Semran Alharbi[1,2]
asaharbi@uqu.edu.sa

Yuefeng Li[1]
y2.li@qut.edu.au

Yue Xu[1]
yue.xu@qut.edu.au

[1]School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Australia

[2]Department of Computer Science
Umm Al-Qura University
Makkah, Saudi Arabia

Abstract

Unsupervised topic models, such as Latent Dirichlet Allocation (LDA), are widely used as automated feature engineering tools for textual data. They model words semantics based on some latent topics on the basis that semantically related words occur in similar documents. However, words weights that are assigned by these topic models do not represent the semantic meaning of these words to user information needs. In this paper, we present an innovative and effective extended random sets (ERS) model to enhance the semantic of topical words. The proposed model is used as a word weighting scheme for relevance feature selection (FS). It accurately weights words based on their appearance in the LDA latent topics and the relevant documents. The experimental results, based on 50 collections of the standard RCV1 dataset and TREC topics for information filtering, show that the proposed model significantly outperforms eight, state-of-the-art, baseline models in five standard performance measures.

In: Proceedings of IJCAI Workshop on Semantic Machine Learning (SML 2017), Aug 19-25 2017, Melbourne, Australia.

1 Introduction

LDA [BNJ03] is currently the most common probabilistic topic model compared to similar models, such as probabilistic Latent Semantic Analysis (pLSA) [Hof01], with a wide range of applications [Ble12]. LDA statistically discovers hidden topics from documents as features to be used for different tasks in information retrieval (IR) [WC06, WMW07], information filtering (IF) [GXL15] and for many other text mining and machine learning applications. LDA represents documents by a set of topics, and each topic is a set of semantically related terms[1]. Thus, it is capable of clustering related words in a document collection, which can reduce the impact of common problems like polysemy, synonymy and information overload [AZ12].

The core and critical part of any text FS method is the *weighting function*. It assigns a numerical value (usually a real number) to each feature, which specifies how informative the feature is to the user's information needs [ALA13]. In the context of probabilistic topic modelling in general and LDA specifically, calculating a term weight is done locally at its document-level based on two components; the term local document-topics distributions and the global term-topics assignment. Therefore, in a set of similar documents, a specific term might receive a different weight in each single document even though this term is semantically identical across all these documents. Such approach does not accurately reflect on the semantic meaning and usefulness of this term to the entire user's information needs. It badly influences the performance of LDA

[1] In this paper, terms, words, keywords or unigrams are used interchangeably.

for FS as it is uncertain and difficult to know which weight is more representative and should be assigned to the intended term. Would it be the average weight? The highest? The lowest? The aggregated? Several experiments in various studies confirm that the local-global weighting approach of the LDA is ineffective for relevant FS [GXL15].

Given a document set that describes user information needs, global statistics, such as document frequency (df), reveal the discriminatory power of terms [LTSL09]. However, in IR, selecting terms based on global weighting schemes did not show better retrieval performance [MO10], because global statistics cannot describe the local importance of terms [MC13]. From the LDA's perspective, it is challenging and still uncertain on how to use LDA's local-global term weighting function in a global context due to the complex relationships between terms and many entities that represent the entire collection. A term, for example, might appear in multiple documents and LDA topics, and each topic may also cover many documents or paragraphs that contain the same term. Therefore, the hard question this research tries to answer is: how to generalise the local topic weight (at document level) and combine it with global topical statistics such as the term frequency in both topics and relevant documents for more discriminative and semantically representative global term weighting scheme?

The aim of this research is to develop an effective topic-based FS model for relevance discovery. The model uses a hierarchical framework based on ERS theory to assign a more representative weight to terms based on their appearance in LDA topics and all relevant documents. Therefore, *two major contributions* have been made in this paper to the fields of text FS and IF: (a) A new theoretical model based on multiple ERS [Mol06] to represent and interpret the complex relationships between long documents, their paragraphs, LDA topics and all terms in the collection, where a function describes each relationship; (b) A new and effective term weighting formula that assigns a more discriminately accurate weight to topical terms that represent their relevance to the user information needs. The formula generalises LDA's local topic weight to a global one using the proposed ERS theory and then combines it with the frequency ratio of words in both documents and topics to answer the question asked by the authors. To test the effectiveness of our model, we conducted extensive experiments on RCV1 dataset and the assessors' relevance judgements of the TREC filtering track. The results show that our model significantly outperforms all used baseline FS models for IF despite the type of text features they use (terms, phrases, patterns, topics or even a different combination of them).

2 Related Works

In the literature, there is a significant amount of work that extends and improves LDA to suit different needs including text FS [ZPH08, TG09]. However, our model is intended for IF, and, to the best of our knowledge, it is the first attempt to extend random sets [Mol06] to functionally describe and interpret complex relationships that involve topical terms and other entities in a document collection to enhance the semantic of topical words for relevance FS. Relevance is a fundamental concept in both IR and IF. IR mainly concerns about document's relevance to a query for a specific subject. However, IF discusses the document's relevance to user information needs [LAZ10]. In relevance discovery, FS is a method that selects a subset of features that are relevant to user's needs and thus removing those that are irrelevant, redundant and noisy. Existing methods adopt different type of text features such as terms [LTSL09], phrases (n-grams) [ALA13], patterns (a pattern is a set of associated terms) [LAA$^+$15], topics [DDF$^+$90, Hof01, BNJ03] or a combination of them for better performance [WMW07, LAZ10, GXL15].

The most efficient FS methods for relevance, are the ones that are developed based on *weighting function*, which is the core and critical part of the selection algorithm [LAA$^+$15]. Using LDA words weighting function for relevance is still limited and does not show encouraging results [GXL15] including similar topic-based models such as the pLSA [Hof01]. For better performance, Gao et al (2015) [GXL15] integrate pattern mining techniques into topic models to discover discriminative features. Such work is expensive and susceptible to the features-loss problem and also might be impacted by the uncertainty of the probabilistic topic model. ERS is proven to be effective in describing complex relations between different entities and interprets them as a function (weighting function) [Li03]. Thus, the ERS-based models can be used to weight closed sequential patterns more accurately and thus facilitate the discovery of specific ones as appears in [ALX14]. However, selecting the most useful patterns is challenging due to a large number of patterns generated from relevant documents using various minimum supports (*min_sup*), and also may lead to feature-loss.

3 Background Overview

For a given corpus C, the relevant long documents set $D \subseteq C$ represents user's information needs that might have multiple subjects. The proposed model uses D for training where each document $d_x \in D$ has a set of paragraphs PS and each paragraph has a set of terms T. Θ is the set of all paragraphs in D and $PS \subseteq \Theta$. A set of terms Ω is the set of all unique words in D.

3.1 Latent Dirichlet Allocation

The proposed model uses LDA to reduce the dimensionality of D to a set of manageable topics Z, where V is the number of topics. LDA assumes that each document has multiple latent topics [GXL15], and defines each topic $z_j{\in}Z$ as a multinomial probability distribution over all words in Ω as $p(w_i|z_j)$ in which $w_i{\in}\Omega$ and $1{\leq}j{\leq}V$ such that $\sum_i^{|\Omega|} p(w_i|z_j){=}1$. LDA also represents a document d as a probabilistic mixture of topics as $p(z_j|d)$. As a result, and based on the number of latent topics, the probability (local weight) of word w_i in document d can be calculated as $p(w_i|d){=}\sum_{j=1}^{V} \big(p(w_i|z_j){\times}p(z_j|d)\big)$. Finally, all hidden variables, $p(w_i|z_j)$ and $p(z_j|d)$, are statistically estimated by the Gibbs sampling algorithm [SG07].

3.2 Random Set

A random set is a random object that has values, which are subsets that are taken from some space [Mol06]. It works as an effective measure of uncertainty in imprecise data for decision analysis [Ngu08]. For example, let Z and Ω be finite sets that represent topics and words respectively. Γ is a set-valued mapping from Z (the evidence space) onto Ω that can be written as Γ: $Z \rightarrow 2^{\Omega}$, and P is a probability function defined on Z, thus the pair (P, Γ) is called a random set [KSH12]. Γ can be extended as ξ :: $Z \rightarrow 2^{\Omega\times[0,1]}$ (also called an extended set-valued mapping), which satisfies $\sum_{(w,p)\in\xi(z)} p{=}1$ for each $z{\in}Z$. Let P be a probability function on Z, such that $\sum_{z\in Z} P(z){=}1$. We call (ξ, P) an extended random set.

4 The Proposed Model

The proposed model (Figure 1) deals with the local weight problem of terms that is assigned by the LDA probability function (described in section 3.1) by exploring all possible relationships between different entities that influence the weighting process. The targeting entities in our model are documents, paragraphs, topics, and terms. The possible relationships between these entities are complex (a set of one-to-many relationships). For example, a document can have many paragraphs and terms; a paragraph can have multiple topics; a topic can have many terms. Inversely, a topic can cover many paragraphs, and a term can appear in many documents and topics.

In this model, we proposed three ERSs to describe such complex relationships, where each ERS can be interpreted as a function by which we can determine the importance of the main entity in the relationship. Then, the proposed ERS theory is used to develop a new weighting scheme to accurately weight topical words by generalising the topic's local weight, and, then, combine it with the frequency ratio of words in both documents and topics.

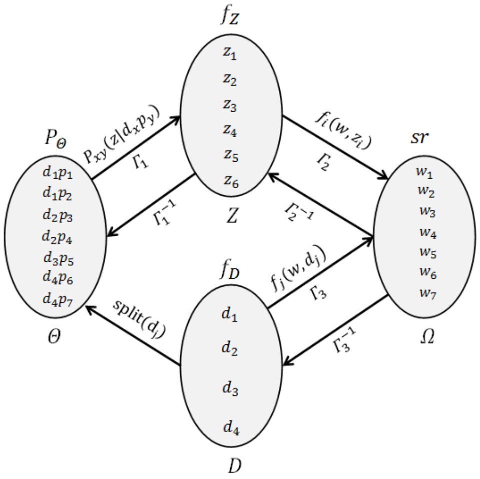

Figure 1: our proposed model

4.1 Extended Random Sets

Let assume we have a set of topics $Z{=}\{z_1, z_2, z_3, \ldots, z_V\}$ in Θ and let $D{=}\{d_1, d_2, d_3, \ldots, d_N\}$ is a set of N relevant long documents. Each document d_x consists of M paragraphs such as $d_x{=}\{p_1, p_2, p_3, \ldots, p_M\}$. A paragraph p_y consists of a set of L words, for example, $p_y{=}\{w_1, w_2, w_3, \ldots, w_L\}$. A word w is a keyword or unigram, where the function $words(p)$ returns a set of words appear in paragraph p. A topic z can be defined as a probability distribution over the set of words Ω where $words(p){\subseteq}\Omega$ for every paragraph $p{\in}\Theta$.

For each $z_i{\in}Z$, let $f_i(w, z_i)$ be a frequency function on Ω, such that $\Gamma(z_i){=}\{w|w{\in}\Omega, f_i(w, z_i){\geq}0\}$ while the inverse mapping of Γ is defined as $\Gamma^{-1} : \Omega \rightarrow 2^Z$; $\Gamma^{-1}(w){=}\{z{\in}Z|w{\in}\Gamma(z)\}$. Also, for each $d_j{\in}D$, let $f_j(w, d_j)$ be a frequency function on Ω, such that $\Gamma(d_j) = \{w|w{\in}\Omega, f_j(w, d_j){>}0\}$ while the inverse mapping of Γ is defined as $\Gamma^{-1} : \Omega \rightarrow 2^D$; $\Gamma^{-1}(w){=}\{d{\in}D|w{\in}\Gamma(d)\}$. These extended set-valued mappings can decide a weighting function on Ω, which satisfies $sr :: \Omega \rightarrow [0, +\infty)$ such that

$$sr(w) = \sum_{d_j\in\Gamma^{-1}(w)} \left[\frac{1}{f_j(w,d_j)} \cdot \Big(\sum_{z_i\in\Gamma^{-1}(w)} \big(P_z(z_i) \times f_i(w,z_i)\big)\Big)\right] \tag{1}$$

where $sr(w)$ is the combined weight of topical word w at the collection level.

The extended random set Γ_1 is proposed to describe the relationships between paragraphs and topics using the conditional probability function $P_{xy}(z|d_xp_y)$ as $\Gamma_1 : \Theta \to 2^{Z\times[0,1]}$; $\Gamma_1(d_xp_y)=\{(z_1, P_{xy}(z_1|d_xp_y)), \ldots\}$.

Similarly Γ_2 is also proposed to describe the relationship between topics and terms using the defined frequency function $f_i(w, z_i)$ as $\Gamma_2 : Z \to 2^{\Omega\times[0,+\infty)}$; $\Gamma_2(z_i)=\{(w_1, P_i(w_1|z_i)), \ldots\}$.

Lastly, Γ_3 is also proposed to describe the relationship between documents and terms using the defined frequency function $f_j(w, d_j)$ as $\Gamma_3 : D \to 2^{\Omega\times[0,+\infty)}$; $\Gamma_3(d_j)=\{(w_1, f_j(w_1, d_j)), \ldots\}$

Based on the inverse mapping described above, we have Γ_1^{-1}, Γ_2^{-1} and Γ_3^{-1}. Γ_1^{-1} describes the inverse relationships between topics and paragraphs using the probability function $P_z(z_i)$ such that $\Gamma_1^{-1}(z)=\{d_xp_y|z\in\Gamma_1(d_xp_y)\}$ while Γ_2^{-1}, on the other hand, describes the inverse relationships between terms and topics using $f_i(w, z_i)$ function such that $\Gamma_2^{-1}(w)=\{z|w\in\Gamma_2(z)\}$. Γ_3^{-1} describes the inverse relationships between terms and documents using $f_j(w, d_j)$ function such that $\Gamma_3(w)=\{d|w \in \Gamma_3(d)\}$

4.2 Generalised Topic Weight

To estimate the generalised topic weight in D, we need to calculate the probability of each topic $P_z(z_i)$ in each paragraph of document d and similarly for all documents in D based on Γ_1^{-1} in which we assume $P_\Theta(d_xp_y) = \frac{1}{N}$, where N is the total number of paragraphs as follows:

$$\begin{aligned} P_z(z_i) &= \sum_{d_xp_y\in\Gamma_1^{-1}(z_i)} (P_\Theta(d_xp_y) \times P_{xy}(z_i|d_xp_y)) \\ &= \frac{1}{N} \sum_{d_xp_y\in\Gamma_1^{-1}(z_i)} P_{xy}(z_i|d_xp_y) \end{aligned} \tag{2}$$

where $P_{xy}(z_i|d_xp_y)$ is estimated by LDA, d_xp_y refers to paragraph y in document x. Γ_1^{-1} is a mapping function defined previously.

4.3 Topical Word Weighting Scheme

To calculate the topical word weight at collection level, we simply substitute $P_z(z_i)$ in Equation 1 by its value from Equation 2. Equation 3 shows the substitution.

5 Evaluation

To verify the proposed model, we designed two hypotheses. First, our ERS model can effectively generalise the topic's local weight that is estimated from all documents paragraphs. The generalisation has led to a more accurate term weighting scheme especially when it is combined with the term frequency ratio in both documents and topics. Second, our model, overall, is more effective in selecting relevant features than most, state-of-the-art, term-based, pattern-based, topic-based or even mix-based FS models. To support these two hypotheses, we conducted experiments and evaluated their performance.

5.1 Dataset

The first 50 collections of the standard Reuters Corpus Volume 1 (RCV1) dataset is used in this research due to being assessed by domain experts at NIST [SR03] for TREC[2] in their filtering track. This number of collections is sufficient and stable for better and reliable experiments [BV00]. RCV1 is collections of documents where each document is a news story in English published by Reuters.

5.2 Baseline models

We compared the performance of our model to eight different baseline models. These models are categorised into five groups based on the type of feature they use. The proposed model is trained only on relevant documents and does not consider irrelevant ones. Therefore, for fair comparison and judgement, we can only select a baseline model that either unsupervised or does not require the use of irrelevant documents.

We selected *Okapi BM25* [RZ09], which is one of the best term-based ranking algorithm. The phrase-based model *n-Grams* is selected. It represents user's information needs as a set of phrases where $n = 3$ as it is the best value reported by Gao et al. (2015) [GXL15]. The *Pattern Deploying based on Support* (PDS) [ZLW12] is one of the pattern-based models. It can overcome the limitations of pattern frequency and usage. We selected the *Latent Dirichlet Allocation* (LDA) [BNJ03] as the most widely used topic modelling algorithm. From the same group we also selected the *Probabilistic Latent Semantic Analysis* (pLSA) [Hof01]; it is similar to the LDA and can deal with the problem of polysemy. Three models were selected from the mix-based category. First, we selected the *Pattern-Based Topic Model* (PBTM-FP) [GXL15] that incorporates topics and frequent patterns *FP* to obtain semantically rich and discriminative representation for IF. Secondly, the *PBTM-FCP* [GXL15], which is similar to the PBTM-FP except it uses the frequent closed pattern *FCP* instead. Lastly, we selected the *Topical N-Grams* (TNG) [WMW07] that integrates the topic model with phrases (n-grams) to discover topical phrases that are more discriminative and interpretable.

[2] http://trec.nist.gov/

$$sr(w) = \frac{1}{N} \sum_{d_j \in \Gamma_3^{-1}(w)} \left[\frac{1}{f_j(w, d_j)} \times \left(\sum_{z_i \in \Gamma_2^{-1}(w)} \Big(f_i(w, z_i) \times \big(\sum_{d_x p_y \in \Gamma_1^{-1}(z_i)} P_{xy}(z_i | d_x p_y) \big) \Big) \right) \right] \tag{3}$$

5.3 Evaluation Measures

The effectiveness of our model is measured based on relevance judgements by five metrics that are well-established and commonly used in the IR and IF communities. These metrics are the *average precision* of the top-20 ranked documents (top-20), *break-even point* (b/p), *mean average precision* (MAP), *F-score* (F_1) measure, and *11-points interpolated average precision* (IAP). For more details about these measures, the reader can refer to Manning et al (2008) [MRS08]. For even better analysis of the experimental results, the *Wilcoxon signed-rank test* (Wilcoxon T-test) [Wil45] was used. Wilcoxon T-test is a statistical non-parametric hypothesis test used to compare and assess if the ranked means of two related samples differ or not. It is a better alternative to the student's t-test, especially when no normal distribution is assumed.

5.4 Experimental Design

For each collection, we train our model on all paragraphs of relevant documents D in the training part of the collection. We use LDA to extract ten topics because it is the best number for each collection as it has reported in [GXL13, GXL14, GXL15]. Then, the proposed model scores documents' terms, ranks them and uses the *top-k* features as a query to an IF system. The IF system uses unknown documents (from the testing part of the same collection) to decide their relevance to the user's information needs (relevant or irrelevant). However, specifying the value of k is experimental. The same process is also applied separately to all baseline models. If the results of the IF system returned by the five metrics are better than the baseline results, then we can claim that our model is significant and outperforms a baseline model.

The IF testing system uses the following equation to rank the testing documents set:

$$weight(d) = \sum_{t \in Q} x, \; if \begin{cases} t \in d, x = weight(t) \\ t \notin d, x = 0 \end{cases} \tag{3}$$

where $weight(d)$ is the weight of document d.

5.5 Experimental Settings

In our experiment, we use the MALLET toolkit [McC02] to implement all LDA-based models except for the pLSA model where we used the Lemur toolkit [3] instead. All topic-based models require some parameters to be set. For the LDA-based models, we set the number of iterations for the Gibbs sampling to 1000 and for the hyper-parameters to $\beta = 0.01$ and $\alpha = 50/V$ as they were justified in [SG07]. We configured the number of iterations for the pLSA to be 1000 (default setting). For the experimental parameters of the BM25, we set $b = 0.75$ and $k_1 = 1.2$ as recommended by Manning et al. (2008) [MRS08].

5.6 Experimental Results

Table 1 and figure 2 show the evaluation results of our model and the baselines. These results are the average of the 50 collections of the RCV1. The results in Table 1 have been categorised based on the type of feature used by the baseline model and the *improvement*% represents the percentage change in our model's performance compared to the best result of the baseline model (marked in bold if there is more than one baseline model in the category). We consider any improvement that is greater than 5% to be significant.

Table 1 shows that our model outperformed all baseline models for information filtering in all five measures. Regardless of the type of feature used by the baseline model, our model is significantly better on average by a minimum improvement of 8.0% and 39.7% maximum. Moreover, the 11-*points* result in figure 2 illustrates the superiority of the proposed model and confirms the significant improvements that shown in table 1.

Table 1: Evaluation results of our model in comparison with the baselines (grouped based on the type of feature used by the model) for all measures averaged over the first 50 document collections of the RCV1 dataset.

Model	Top-20	b/p	MAP	$F_{\beta=1}$	IAP
our model	**0.560**	**0.471**	**0.502**	**0.475**	**0.526**
LDA	**0.492**	**0.414**	**0.442**	**0.437**	**0.468**
pLSA	0.423	0.386	0.379	0.392	0.404
improvement%	**+13.9%**	**+13.8%**	**+13.7%**	**+8.5%**	**+12.3%**
PDS	0.496	0.430	0.444	0.439	0.464
improvement%	**+12.9%**	**+9.5%**	**+13.2%**	**+8.0%**	**+13.4%**
n-Gram	0.401	0.342	0.361	0.386	0.384
improvement%	**+39.7%**	**+37.8%**	**+39.1%**	**+22.9%**	**+37.1%**
BM25	0.445	0.407	0.407	0.414	0.428
improvement%	**+25.8%**	**+15.6%**	**+23.5%**	**+14.6%**	**+22.9%**
PBTM-FCP	**0.489**	**0.420**	0.423	0.422	0.447
PBTM-FP	0.470	0.402	**0.427**	**0.423**	**0.449**
TNG	0.447	0.360	0.372	0.386	0.394
improvement%	**+14.5%**	**+12.1%**	**+17.7%**	**+12.2%**	**+17.1%**

Wilcoxon T-test results (Table 2) present the p-values of the results of our model compared to all base-

[3] https://www.lemurproject.org/

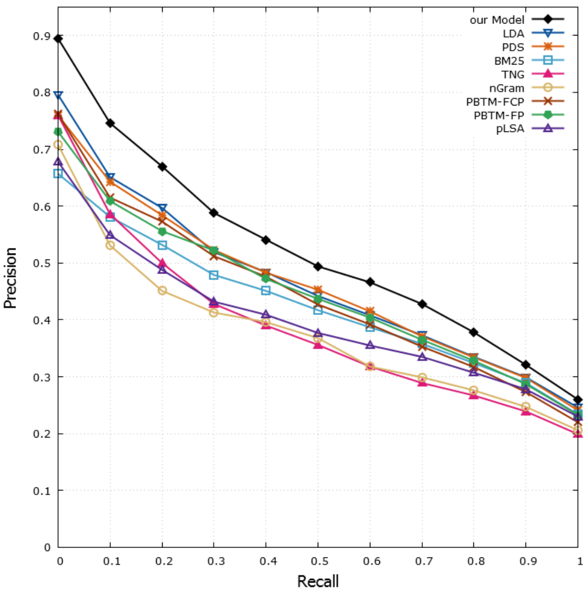

Figure 2: 11-*points* result of our model in comparison with baselines averaged over the first 50 document collections of the RCV1 dataset.

line models on all performance measures. A model's result is considered significantly different from other model's if the p-value is less than 0.05 [Wil45].Clearly, the p-value for all metrics is largely less than 0.05 confirming that our model's performance is significantly different from all baselines. This shows that our model gains substantial improvement compared to the used baseline models.

Table 2: Wilcoxon T-test *p-values* of the baseline models in comparison with our model's.

Model	Top-20	b/p	MAP	$F_{\beta=1}$	IAP
LDA	0.004165	0.000179	7.00×10^{-6}	8.96×10^{-6}	6.71×10^{-6}
pLSA	1.48×10^{-4}	1.49×10^{-4}	6.65×10^{-7}	5.86×10^{-7}	1.72×10^{-7}
PDS	0.008575	0.003034	0.000194	0.000140	4.53×10^{-5}
n-Gram	7.46×10^{-8}	1.05×10^{-7}	1.71×10^{-9}	1.86×10^{-9}	1.23×10^{-9}
BM25	0.000353	0.008264	0.000279	0.000117	5.68×10^{-5}
TNG	0.010360	0.000607	0.000180	0.000137	3.76×10^{-5}
PBTM-FP	0.003442	7.19×10^{-4}	0.000382	0.000235	5.81×10^{-5}
PBTM-FCP	0.048010	0.033410	0.000306	0.000289	0.000180

Based on the results presented earlier, we are confident in claiming that our extended random sets model can effectively generalise the local topic weight at the document level in the LDA term scoring function and, thus, provide a more globally representative term weight when it combined the term frequency in document and topics. Also, our model is more effective in selecting relevant features to acquire user's information needs that represented by a set of long documents.

6 Conclusion

This paper presents an innovative and effective topic-based feature ranking model to enhance the semantic of topical words to acquire user needs. The model extends random sets to generalise the LDA topic weight at the document level. Then, a term weighting scheme is developed to accurately rank topical terms based on their frequent appearance in the LDA topics distributions and all relevant documents. The new calculated weight effectively reflects the relevance of a term to user's information needs and maintains the same semantic meaning of terms across all relevant documents. The proposed model is tested for IF on the standard RCV1 dataset, TREC topics, five different performance measurement metrics and eight state-of-the-art baseline models. The experimental results show that our model achieved significant performance compared to all other baseline models.

References

[ALA13] Mubarak Albathan, Yuefeng Li, and Abdulmohsen Algarni. *Enhanced N-Gram Extraction Using Relevance Feature Discovery*, pages 453–465. Springer International Publishing, Cham, 2013.

[ALX14] Mubarak Albathan, Yuefeng Li, and Yue Xu. Using extended random set to find specific patterns. In *WI'14*, volume 2, pages 30–37. IEEE, 2014.

[AZ12] Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer, 2012.

[Ble12] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[BV00] Chris Buckley and Ellen M Voorhees. Evaluating evaluation measure stability. In *SIGIR'00*, pages 33–40. ACM, 2000.

[DDF+90] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[GXL13] Yang Gao, Yue Xu, and Yuefeng Li. Pattern-based topic models for information filtering. In *ICDM'13*, pages 921–928. IEEE, 2013.

[GXL14] Yang Gao, Yue Xu, and Yuefeng Li. Topical pattern based document modelling and relevance ranking. In *WISE'14*, pages 186–201. Springer, 2014.

[GXL15] Yang Gao, Yue Xu, and Yuefeng Li. Pattern-based topics for document modelling in information filtering. *IEEE TKDE*, 27(6):1629–1642, 2015.

[Hof01] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.

[KSH12] Rudolf Kruse, Erhard Schwecke, and Jochen Heinsohn. *Uncertainty and vagueness in knowledge based systems: numerical methods*. Springer Science & Business Media, 2012.

[LAA$^+$15] Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana. Relevance feature discovery for text mining. *IEEE TKDE*, 27(6):1656–1669, 2015.

[LAZ10] Yuefeng Li, Abdulmohsen Algarni, and Ning Zhong. Mining positive and negative patterns for relevance feature discovery. In *KDD'10*, pages 753–762. ACM, 2010.

[Li03] Yuefeng Li. Extended random sets for knowledge discovery in information systems. In *RSFDGrC'03*, pages 524–532. Springer, 2003.

[LTSL09] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE TPAMI*, 31(4):721–735, 2009.

[MC13] K Tamsin Maxwell and W Bruce Croft. Compact query term selection using topically related text. In *SIGIR'13*, pages 583–592. ACM, 2013.

[McC02] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. 2002.

[MO10] Craig Macdonald and Iadh Ounis. Global statistics in proximity weighting models. In *Web N-gram Workshop*, page 30. Citeseer, 2010.

[Mol06] Ilya Molchanov. *Theory of random sets*. Springer Science & Business Media, 2006.

[MRS08] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[Ngu08] Hung T Nguyen. Random sets. *Scholarpedia*, 3(7):3383, 2008.

[RZ09] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[SG07] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[SR03] Ian Soboroff and Stephen Robertson. Building a filtering test collection for trec 2002. In *SIGIR'03*, pages 243–250. ACM, 2003.

[TG09] Serafettin Tasci and Tunga Gungor. Lda-based keyword selection in text categorization. In *ISCIS'09*, pages 230–235. IEEE, 2009.

[WC06] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR'06*, pages 178–185. ACM, 2006.

[Wil45] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[WMW07] Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM'07*, pages 697–702. IEEE, 2007.

[ZLW12] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu. Effective pattern discovery for text mining. *IEEE TKDE*, 24(1):30–44, 2012.

[ZPH08] Zhiwei Zhang, Xuan-Hieu Phan, and Susumu Horiguchi. An efficient feature selection using hidden topic in text categorization. In *AINAW'08*, pages 1223–1228. IEEE, 2008.