Topical Term Weighting based on Extended Random Sets for Relevance Feature Selection

Abdullah Alharbi
Queensland University of Technology
School of EECS
2 George St.
Brisbane, Queensland 4000, Australia
asaharbi@uqu.edu.sa

Yuefeng Li
Queensland University of Technology
School of EECS
2 George St.
Brisbane, Queensland 4000, Australia
y2.li@qut.edu.au

Yue Xu
Queensland University of Technology
School of EECS
2 George St.
Brisbane, Queensland 4000, Australia
yue.xu@qut.edu.au

ABSTRACT

Selecting relevant features from long documents that describe user's information needs is challenging due to the nature of text, where synonymy, polysemy, noise and high dimensionality are common problems. Traditional feature selection (FS) methods assume that long documents discuss only one topic. Such assumption would be too simple knowing that long documents can discuss multiple topics. Topic-based techniques, such as the LDA, relax this assumption and have been developed on the basis that a document can exhibit multiple latent topics. However, LDA does not show encouraging results in FS for relevance, because LDA calculates a term weight based on its local document and does not generalise it globally on the entire documents collection. To address this problem, we propose an innovative and effective extended random set (ERS) model to generalise LDA weight for local document terms. The proposed model is used as a term weighting scheme for relevance FS. It can assign a more discriminately accurate weight to terms based on their appearance in the latent topics and relevant documents. The experimental results, based on the standard RCV1 dataset and the TREC topics, show that our model significantly outperforms eight state-of-the-art baseline models in five different and popular performance measures.

CCS CONCEPTS

• **Information systems → Content analysis and feature selection**; *Document representation*; • **Computing methodologies → Latent Dirichlet allocation**;

KEYWORDS

Feature Selection, Term Weighting, Topic Modelling, Latent Dirichlet Allocation, Extended Random Set, Text Mining, Information Filtering.

ACM Reference format:
Abdullah Alharbi, Yuefeng Li, and Yue Xu. 2017. Topical Term Weighting based on Extended Random Sets for Relevance Feature Selection. In *Proceedings of 2017 IEEE/WIC/ACM International Conference on Web Intelligence, Leipzig, Germany, August 2017 (WI'17),* 8 pages.
DOI: 10.475/123_4

WI'17, Leipzig, Germany

DOI: 10.475/123_4

1 INTRODUCTION

In the age of big data, text documents grow exponentially and constitute more than 80% of the unstructured data available on the web or private storage [6]. Unstructured text can be in the form of emails, tweets, reports, articles, logs, reviews and more. These documents contain invaluable information that needs to be automatically extracted for the success of many organisations and businesses. However, it is a big challenge for text mining and machine learning techniques to extract useful information from text data due to the size and the nature of text where synonymy, polysemy and noise are commonly inherited problems [29, 31]. Feature selection, as a dimensionality reduction technique, plays a major role in the Knowledge Discovery in Text (KDT) by improving accuracy and reducing the complexity of many machine learning algorithms [5]. This can be done by selecting a subset of features that are relevant and removing those that are irrelevant, redundant and noisy. Topic modelling algorithms such as the Probabilistic Latent Semantic Analysis (pLSA) [19] and Latent Dirichlet Allocation (LDA) [9] are proven to be effective in reducing the total dimensionality of text to a set of manageable topics [16].

Unlike the pLSA, LDA is the most popular with many applications [8]. LDA can statistically identify hidden topics from text corpus to improve different tasks in information retrieval (IR) [50, 51], information filtering (IF) [16], Multi-document Summarization [55], collection visualisation [11], Personalised Ontology Learning [7] and many other text mining and machine learning applications. LDA represents documents by a set of topics, where each topic is a set of semantically related terms. Thus, it is capable of clustering related words[1] in documents collection, which can reduce the negative impact of common problems like polysemy, synonymy and information overload [1]. However, in reality, LDA treats topics as multinomial distributions over words and documents as a probabilistic mixture over a pre-defined number of latent topics.

Selecting relevant terms from a collection of long documents that describe user's information needs is important for many applications including, but not limited to, information retrieval [17], information filtering [16], text classification [13] and clustering [32]. The core and critical part of any text feature selection 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 [2]. 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 distribution and the global

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

term-topics assignment. Therefore, in a set of similar documents, a specific term might receive a different weight in each individual document even though this term is semantically identical across all 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 for feature selection as it is uncertain and difficult to know which weight is more representative and should be assigned to the intended term. 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 feature selection [16].

Given a collection of documents that describe user information needs, the terms global statistics such as the document frequency (df) reveal the discriminatory power of terms [27]. However, in information retrieval, selecting terms based on global weighting schemes did not show better retrieval performance [33], because global statistics cannot describe the local importance of terms [35]. From the LDA's perspective, it is challenging and still uncertain on how to use LDA's local-global term weighting function in the 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 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 do we combine the global term weight (df) and the LDA's local weight component together for a better and more discriminative global term weighting scheme?

The aim of this research is to develop an effective topic-based feature selection model for relevance discovery. The model uses a hierarchical framework based on the ERS theory to assign a more representative weight to topical 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 feature selection and information filtering: (a) A new theoretical model based on multiple extended random sets(ERS) [37] to represent and interpret the complex relationships between long documents, their paragraphs, LDA topics and all terms in the collection, where a probability 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's information needs. The formula generalises the LDA's local term weight to a global one using the proposed ERS theory and then combines it with another global weight (the df)) to answer the previous question in the last paragraph. To test the effectiveness of our model, we conducted substantial experiments on the Reuters Corpus Volume 1 (RCV1) and the assessors' relevance judgements of the TREC filtering track. The results show that our model significantly outperforms all used state-of-the-art baseline feature selection models for information filtering despite the type of text features they use (terms, phrases, patterns, topics or even a different combination of them).

The rest of this paper is organised as follows: section 2 provides an overview of the related works; essential details about the LDA are explained in Section 3; section 4 introduces the extended random set theory, while Section 5 presents our proposed model and the term weighting equations. The used dataset, experiment, baselines, performance measures and results are introduced in Section 6 (Evaluation) followed by concluding remarks and future works in the last section.

2 RELATED WORKS

In the literature, there is a significant amount of work that extends and improves LDA [9] to suit different needs, including feature selection for text classification [48, 58]. However, our model is intended for information filtering, and, to the best of our knowledge, it is the first attempt to extend random sets (ERS) [37] to probabilistically describe and interpret complex relationships that involve topical terms and other entities in a documents collection. The model is used, then, to generalise the local term weight at the document level in LDA's term weighting function for more relevant term selection. Relevance is a fundamental concept in both information retrieval and information filtering. Information retrieval is mainly concerned with the document's relevance to a query about a specific subject. However, information filtering discusses the document's relevance to user's information needs [31]. In relevance discovery, feature selection is a method that selects a subset of features that are relevant to user's information needs and it removes those that are irrelevant, redundant and noisy. Existing feature selection methods adopt different type of text features such as terms [27], phrases (n-gams) [2], patterns [29], topics [9, 12, 19] or combinations of them for better performance [16, 31, 50].

Term-based feature selection methods like TF*IDF [27], Mutual Information (MI) [34], Information Gain (IG) [57], Gini-index [60], Ch-Square (X^2) [21], BM25 [42], Rocchio [44], LASSO [49], ranking SVM [22] and others are efficient and have been developed based on sophisticated mathematical and statistical weighting theories [4, 31]. However, these methods are sensitive to noise and suffer from synonymy and polysemy problems [29]. Further, term-based methods ignore word order in documents. Thus, they miss the semantic relationships between these words [23]. Phrase-based models, on the other hand, use phrases (n-grams) because they are more discriminative and contain semantic information better than individual words [31]. Nevertheless, phrases are less frequent and can be redundant and noisy. Further, published phrase-based experiments do not show encouraging results [39, 45]. To overcome the limitations of phrase-based and term-based methods, different pattern-based techniques have been introduced in [29–31, 53, 54].

A pattern, as a set of associated terms, carries more semantic information than individual words and are more frequent than phrases [31]. Frequent patterns are susceptible to redundancy and noise, but some data mining techniques such as the closed, maximal and master patterns have been developed to remove noisy and redundant patterns [18, 38, 56]. However, pattern-based feature selection models that use these enhanced types of patterns still suffer from their low-frequency. Overall, feature selection models that use terms, phrases, patterns or even combinations of them (called hybrid or mix-based models) have been developed based on the assumption that user's information needs can be described by a single topic (theme) only. However, in reality, they contain multiple semantically related topics or sub-topics [14]. Probabilistic topic modelling algorithms such as pLSA [19] and LDA [9] can overcome

this issue by discovering some topics (or themes) that can represent user's information needs.

The most efficient feature selection methods for relevance are the ones that are developed based on weighting function, which is the core and critical part of the selection algorithm [29]. Using LDA words probability to represent the relevance of these words is still limited and does not show encouraging results [16] including similar topic-based models such as the pLSA [19]. For better performance, Gao et al. (2015) [16] integrated pattern mining techniques into topic models to discover discriminative features. Apart from being effective, such work can be expensive and susceptible to the features-loss problem. It also might be impacted by the uncertainty of the probabilistic topic model (the LDA in their case). The extended random set is proven to be effective in describing complex relationships between different entities and interprets them by a probability function (weighting function) [28]. Thus, the ERS-based model is used to weight closed sequential patterns more accurately and, thus, facilitates the discovery of specific ones as appears in Albathan et al. study [3]. However, selecting the most useful patterns is challenging due to the large number of patterns generated from relevant documents using various minimum supports (*min_sup*), and may also lead to feature-loss. To avoid such a problem, our approach ranks features based on their importance and does not exclude any terms from relevant documents before the weighting process takes place.

3 LATENT DIRICHLET ALLOCATION (LDA)

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 terms in D.

The proposed model uses LDA to reduce the dimensionality of the relevant documents D to a set of manageable topics Z, where V is the number of topics. LDA adopts the bag-of-words model to represent documents for topics discovery [9]. Thus, no relations between words are assumed, and each document is assumed to have multiple latent topics [16]. LDA defines each topic $z_j \in Z$ as a multinomial probability distribution over all terms in Ω as $p(t_i|z_j)$ in which $t_i \in \Omega$ and $1 \leq j \leq V$ such that $\sum_i^{|\Omega|} p(t_i|z_j) = 1$. LDA also represents an individual 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 term t_i in document d can be calculated as $p(t_i|d) = \sum_{j=1}^{V} p(t_i|z_j)p(z_j|d)$. Finally, all hidden variables, $p(t_i|z_j)$ and $p(z_j|d)$, are statistically estimated by the Gibbs sampling algorithm [47]. Interested readers can refer to articles included in the reference for more about LDA, specifically [9, 16, 47].

From a different view, LDA generates two distinct outputs that can be looked at from two levels. At the document level (or the paragraph level as in our case), LDA represents each paragraph p_y by proportions of topics distribution $\theta_{p_y} = (\vartheta_{1,y}, \vartheta_{2,y}, \vartheta_{3,y}, \ldots, \vartheta_{V,y})$. At the collection level, which is the set of relevant documents D in our model, LDA represents D by a set of topics Z where each topic is a probability distribution over all terms in D, ϕ_j for topic z_j and $\Phi = \{\phi_1, \phi_2, \phi_3, \ldots, \phi_V\}$ for all topics. Commonly, different studies [7, 16] use only the top ten terms from each topic based on their probability distribution calculated by $p(t|z)$. However, the proposed model considers all terms in all topics. A third output that LDA can produce is the term-topic assignment, where a set of terms is assigned to a specific topic. Our model considers only the first two representations and makes no use of the third one.

4 EXTENDED RANDOM SETS (ERS)

A random set is a random object that has values, which are subsets that are taken from some space [37]. Random sets, as general mathematical models with many applications, work as an effective measure of uncertainty in imprecise data for decision analysis [40].

Let Z and Ω be finite sets. Z is also called the evidence space. To generalise the local weight of term t in document d that is calculated by the LDA, the set-valued mapping $\Gamma : Z \rightarrow 2^{\Omega}$ is proposed. If Γ is a set-valued mapping from Z onto Ω, and P is a probability function defined on the evidence space. In this case, the pair (P, Γ) is called a random set [25]. The set-valued mapping $\Gamma : Z \rightarrow 2^{\Omega}$ can be extended to an extended set-valued mapping [28] $\xi :: Z \rightarrow 2^{\Omega \times [0,1]}$ which satisfies $\sum_{(t,p)\in\xi(z)} p = 1$ for each $z \in Z$, where Z is a set of topics (or evidences) and Ω is a set of terms (objects) as defined previously.

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. For each $z_i \in Z$, let $P_i(t|z_i)$ be a conditional probability function on Ω, such that $\Gamma(z_i) = \{t | t \in \Omega, P_i(t|z_i) > 0\}$ while the inverse mapping of Γ is defined as $\Gamma^{-1} : \Omega \rightarrow 2^{Z}$; $\Gamma^{-1}(t) = \{z \in Z | t \in \Gamma(z)\}$. The extended set-valued mapping can decide a probability function on Ω, which satisfies $pr :: \Omega \rightarrow [0, 1]$ such that

$$pr(t) = \sum_{z_i \in \Gamma^{-1}(t)} (P(z_i) \times P_i(t|z_i)) \quad (1)$$

where $pr(t)$ is the generalised weight of term t at the collection level that LDA does not calculate.

5 THE PROPOSED MODEL

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 terms, for example, $p_y = \{t_1, t_2, t_3, \ldots, t_L\}$.

A term t is a keyword or unigram, where the function $terms(p)$ returns a set of terms appear in paragraph p. A topic z can be defined as a probability distribution over the set of terms Ω where $terms(p) \subseteq \Omega$ for every paragraph $p \in \Theta$.

The proposed model (Figure 1) deals with the local weight problem of document terms that is assigned by the LDA probability function (described in section 3) by exploring all possible relationships between different entities that influence the term 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; 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 topics.

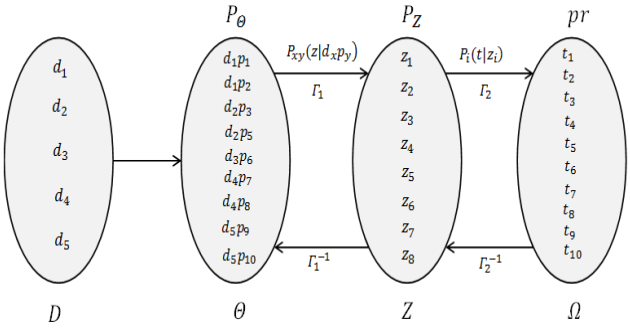

Figure 1: Our proposed model

In this model, we proposed four extended random sets to describe such complex relationships, where each ERS can be interpreted as a probability function by which we can determine the importance of the main entity in the relationship that is described by the defined ERS. The proposed ERS theory is then used to develop new term weighting scheme to generalise LDA's local term probability to a global one that is still descriptive locally and more discriminative when it is combined with the global document frequency (df) as it appears below in Equation 2.

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 $\Gamma_1 : \Theta \rightarrow 2^{Z\times[0,1]}$; $\Gamma_1(d_xp_y) = \{(z_1, P_{xy}(z_1|d_xp_y)), \ldots\}$

Similarly and based on the descriptions in section 4, Γ_2 is also proposed to describe the relationship between topics and terms using the defined conditional probability function $P_i(t|z_i)$ as $\Gamma_2 : Z \rightarrow 2^{\Omega\times[0,1]}$; $\Gamma_2(z_i) = \{(t_1, P_i(t_1|z_i)), \ldots\}$

Based on the inverse mapping specified in section 4, two extended random sets Γ_1^{-1} and Γ_2^{-1} are proposed. 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 the probability function $pr(t)$ such that $\Gamma_2^{-1}(t) = \{z|t\in\Gamma_2(z)\}$.

5.1 Generalised Topical Term Weighting

To calculate the generalised (local to global) term t weight in document d, we need to calculate two probabilities based on Γ_1^{-1} and Γ_2^{-1}. The first one is the probability of each topic $P_z(z_i)$ in each paragraph of document d and similarly for all documents in D in which we assume $P_\Theta(d_xp_y) = \frac{1}{N}$, where N is the total number of paragraphs as follows:

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

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.

Second, for each topic z_i in Z, we need to calculate the conditional probability of term t given topic z_i, $P_i(t|z_i)$ (which is estimated by LDA in our case). Thus, the generalised term weight can be calculated using Equation 1, which can be expanded using Equation 2 as follows:

$$\begin{aligned} pr(t) &= \sum_{z_i\in\Gamma_2^{-1}(t)} (P_z(z_i) \times P_i(t|z_i)) \\ &= \sum_{z_i\in\Gamma_2^{-1}(t)} \left[\left(\frac{1}{N} \sum_{d_xp_y\in\Gamma_1{}^{-1}(z_i)} P_{xy}(z_i|d_xp_y)\right) \times P_i(t|z_i)\right] \\ &= \frac{1}{N} \sum_{z_i\in\Gamma_2^{-1}(t)} \left[P_i(t|z_i) \times \left(\sum_{d_xp_y\in\Gamma_1^{-1}(z_i)} P_{xy}(z_i|d_xp_y)\right)\right] \end{aligned} \tag{3}$$

Finally, The global term weight $w(t)$ at the collection level is calculated as follows:

$$w(t) = pr(t) \times df(t) \tag{4}$$

where $pr(t)$ is the generalised weight of term t that is estimated previously by Equation 3, and $df(t)$ is the document frequency of term t.

Algorithm 1 describes our ERS-based feature selection model where the term weighting function (Equation 3) is its core. The algorithm begins with an initialisation step for all terms in Ω (steps 2-3). Then, the algorithm splits and labels every paragraph in the training documents D (steps 5-7) after removing stop words and stemming all terms in each paragraph. The paragraph label consists of two parts separated by a delimiter. The first part is the paragraph parent document name (or ID) while the second part is a sequential paragraph number. By splitting the long documents paragraphs, we could implicitly exploit the relationships between terms that are in a similar context [24] during topics extraction (next step). Then, the algorithm uses the LDA algorithm to generate two representations (step 9 and step 10). The first one is the paragraph-topics coverage (paragraph-topics distributions) Θ_{xy}. Secondly, a specified number of latent topics (ten topics in our case) ($V = 10$) is generated from the set of labelled paragraphs Θ. Then, the algorithm calculates the term weight (the term probability based on Equation 1) for each term in Ω (steps 12-23). To do that, the algorithm first applies Equation 2 to calculate the topic probability $P_z(z_j)$ for each topic $z_j\in Z$ in all paragraphs in Θ (steps 13-16). Then, the algorithm continues to calculate the term probability in topics that contain the same term (steps 17-21) based on Equation 3. The previous steps generalise the local LDA's term weight to a global one ($pr(t_i)$). Step 23 combines both global weights ($pr(t_i)$ and the document frequency $df(t_i)$).

Finally, we should mention that paragraphs splitting, stop words removal, terms stemming and LDA topics extraction can be done once and off-line in this model.

Algorithm 1: ERS-based term weighting scheme

Input : A set of relevant documents D, the vocabulary Ω and total number of topics V
Output: A function $pr : \Omega \rightarrow [0, 1]$

$Z := T := \Theta := \emptyset$;
for *each* $t_i \in \Omega$ **do**
  $pr(t_i) := 0$;
// split and label all paragraphs in D, where $d_x p_y$ is the yth paragraph in xth document.
for *each* $d_x \in D$ **do**
  for *each* $p_y \in d_x$ **do**
    $\Theta := \Theta \cup \{d_x p_y\}$;
$N := |\Theta|$;
Generate paragraph-topic proportions $\Theta_{xy} := (\vartheta_{1,xy}, \ldots, \vartheta_{V,xy})$ by applying LDA to Θ;
Generate topics $Z := \{z_1, \ldots, z_V\}$ by applying LDA to Θ;
// calculate $pr(t)$ based on eq.(2)
for *each* $t_i \in \Omega$ **do**
  for *each* $z_j \in Z$ **do**
    $P_{z_j} := 0$;
    for *each* $d_x p_y \in \Theta$ **do**
      $P_{z_j} := P_{z_j} + \vartheta_{j,xy}$;
    if $t_i \in z_j$ **then**
      $w' := \left(\frac{tf(t_i, z_j)}{\sum_{t \in z_j} tf(t)} \right) \times P_{z_j}$;
    else
      $w' := 0$;
    $pr(t_i) := pr(t_i) + w'$;
  // $Df(t_i)$ is the document frequency of term t_i
  $pr(t_i) := \frac{pr(t_i) \times Df(t_i)}{N}$;

6 EVALUATION

To verify the proposed model, we designed two hypotheses. First, our ERS model can effectively generalise the local LDA term weight in each document using the latent topics that are extracted from all documents paragraphs. The generalisation has led to a more accurate term weighting scheme especially when it is combined with document frequency. 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 feature selection models. To support these two hypotheses, we conducted experiments and evaluated their performance.

6.1 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 [46] for TREC[2] in their filtering track. This number of collections is sufficient and stable for better and reliable experiments [10].

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

RCV1 is collections of documents where each document is a news story in English published by Reuters. It is a standard and widely used dataset in testing text mining and machine learning techniques. RCV1 is a large dataset with more than 806,000 documents that cover 100 different subjects. Each collection of the RCV1 has been split into training and testing sets, and each set has some relevant and irrelevant documents to the subject they describe. Each document in the RCV1 is an XML document that has many elements.

Our model uses only the *title* and *text* elements for training and testing, where each element is considered a seperate paragraph. To eliminate bias in our experiments, all meta-data elements have been ignored. Before our model can be trained, some pre-processing steps have to be done on each element. First, all stop-words have to be removed. Second, all keywords are stemmed using the Porter Suffix Stripping algorithm [41]. Lastly, during the training phase of our model, each and every paragraph (element) of the relevant documents are separately split and labelled.

6.2 Baseline models

For better and comprehensive evaluation, we compared the performance of our model to eight different baseline feature selection models that are considered state-of-the-art. 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.

From the term-based category, we selected the **Okapi BM25** [42] which is considered one of the best unsupervised ranking algorithm in IR. The standard phrase-based model **n-Grams** is used in this paper. 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) [14] in a similar experiment on RCV1. The **Pattern Deploying based on Support (PDS)** [59] is one of the state-of-the-art pattern-based feature selection models. It is an enhanced extension to the PTM [54] and the PDM [53] to overcome the limitations of pattern frequency and usage. From the topic-based category, we selected the **Latent Dirichlet Allocation (LDA)** [9] as the most widely used statistical topic modelling algorithm. It assigns weight to terms based on their appearance in individual documents (local) and hidden topics (global). From the same group we also selected the **Probabilistic Latent Semantic Analysis (pLSA)** [19], which is an enhanced probabilistic model of the LSA [12]; it is similar to the LDA and can deal with the problem of polysemy. Three feature selection models were selected from the mix-based category. The first one is the **Pattern-Based Topic Model (PBTM-FP)** [16], which incorporates topics and frequent patterns **FP** to obtain semantically rich and discriminative representation for information filtering. Secondly, the **(PBTM-FCP)** [16], which is similar to the PBTM-FP except it uses the frequent closed pattern **FCP** instead. Lastly, we selected the **Topical N-Grams (TNG)** [50] that integrates the topic model with phrases (n-grams) to discover topical phrases that are more discriminative and interpretable. TNG is treated as a relevance ranking model in our experiment as it appears in the recent study of Gao et al. (2015) [15].

6.3 Evaluation Measures

The effectiveness of our proposed model is measured based on relevance judgements by five metrics that are well-established and commonly used in the IR and IF research 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) [34].

For even better analysis of the experimental results, the **Wilcoxon signed-rank test** (Wilcoxon T-test) [52] was used. Wilcoxon T-test is a statistical non-parametric hypothesis test used to compare and asses 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.

6.4 Experimental Design

We treated our proposed model as a relevance feature selection model for information filtering (IF) based on the testing system of the TREC filtering track [43]. Therefore, to effectively measure the performance of our model and its baselines, we conducted a series of experiments on the first 50 collections of the standard RCV1 dataset and their TREC relevance judgements that are assessed by domain experts. These experiments have been carried out to prove that our evaluation hypothesis is valid.

For each collection, we train our model on all paragraphs of 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 [14–16]. Then, the proposed model weights 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:

$$weight(d) = \sum_{t \in Q} x, \; if \begin{cases} t \in d, x = weight(t) \\ t \notin d, x = 0 \end{cases} \tag{5}$$

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

6.5 Experimental Settings

In our experiment, we use the MALLET toolkit [36] 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 be 1000 and for the hyper-parameters to be $\alpha = 50/V$ and $\beta = 0.01$ as they were used and justified in [47]. 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) [34]. The settings of the other baseline models are kept the same as they have been reported in their original experiments.

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

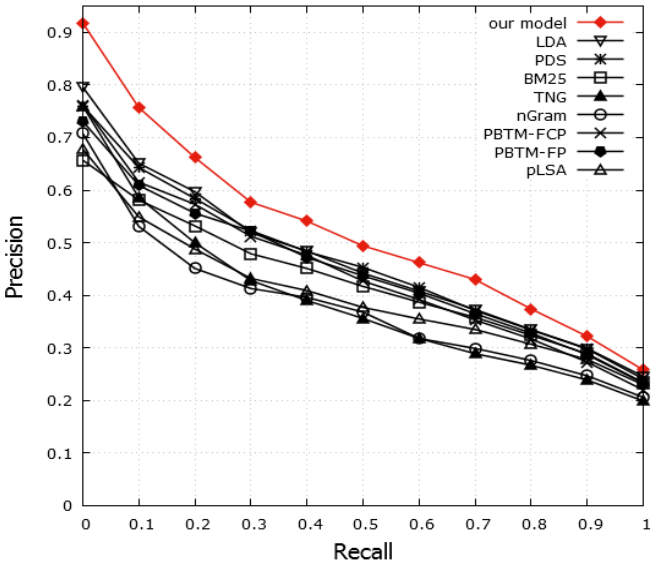

Figure 2: 11-*points* result of our model in comparison with baselines averaged over the first 50 document collections of the RCV1 dataset.

6.6 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 7.7% and 41.4% maximum. Moreover, the 11-*points* result in figure 2 illustrates the superiority of our model and confirms the significant improvements that shown in Table 1.

Wilcoxon T-test results (Table 2) present the p-values of the results of our model compared to all baseline 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 [52].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.

Based on the results presented earlier, we are confident in claiming that our ERS model can effectively generalise the local term

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	0.000752	5.54×10^{-6}	1.10×10^{-5}	1.89×10^{-5}	7.49×10^{-6}
pLSA	6.06×10^{-5}	5.81×10^{-5}	9.17×10^{-7}	1.49×10^{-6}	1.63×10^{-7}
PDS	0.007638	0.001870	0.000295	0.000434	5.46×10^{-5}
n-Gram	5.73×10^{-8}	2.14×10^{-8}	2.04×10^{-9}	2.58×10^{-9}	1.26×10^{-9}
BM25	0.000181	0.002140	0.000254	0.000122	4.73×10^{-5}
TNG	0.003689	0.000226	0.000202	0.000167	2.68×10^{-5}
PBTM-FP	0.001166	6.60×10^{-5}	0.000539	0.000458	3.04×10^{-5}
PBTM-FCP	0.023060	0.005005	0.000413	0.000693	0.000127

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	**0.567**	**0.475**	**0.500**	**0.473**	**0.527**
LDA	**0.492**	**0.414**	**0.442**	**0.437**	**0.468**
pLSA	0.423	0.386	0.379	0.392	0.404
improvement%	**+15.3%**	**+14.8%**	**+13.3%**	**+8.1%**	**+12.5%**
PDS	0.496	0.430	0.444	0.439	0.464
improvement%	**+14.3%**	**+10.4%**	**+12.8%**	**+7.7%**	**+13.5%**
n-Gram	0.401	0.342	0.361	0.386	0.384
improvement%	**+41.4%**	**+38.9%**	**+38.6%**	**+22.5%**	**+37.3%**
BM25	0.445	0.407	0.407	0.414	0.428
improvement%	**+27.4%**	**+16.6%**	**+23.0%**	**+14.2%**	**+23.1%**
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%	**+16.0%**	**+13.1%**	**+17.2%**	**+11.9%**	**+17.2%**

weight at the document level in the LDA term weighting function and, thus, provide a more globally representative weight when it combined with document frequency. Also, our model is more effective in selecting relevant features to acquire user's information needs that represented by a set of long documents.

7 CONCLUSION AND FUTURE WORKS

This paper presents an innovative topic-based feature selection model for relevance discovery. The model extends random sets to generalise the local LDA terms probability at the document level. Then, a term weighting scheme is developed to accurately weight topical terms based on their appearance in the LDA topics distributions and all relevant documents. The 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 information filtering on the standard RCV1 dataset, TREC assessors' relevance judgements, 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.

For future work, we will investigate the following issues:

(1) In the term weighting equation, we assumed that each paragraph in the collection has equal importance, but in reality, they differ in the amount of information they contain. For this issue, we will investigate the use of a clustering technique to group similar documents or paragraphs together for a more accurate assumption.

(2) Our model shows significant performance in weighting terms in long documents. However, its performance in weighing other discriminative features such as phrases, patterns and concepts is still unknown. We will investigate this issue and explore the possibilities of adapting our model inputs to consider these types of features.

(3) As our model uses LDA, we expect it to favour highly frequent terms. These terms are usually general and less discriminative. We will investigate the possibilities of using more discriminative topic models such as DiscLDA [26] and DTM [20] for better performance. Also, we will study the positive and negative impact of using optimising techniques on the LDA during the topics learning phase.

(4) Our model is only trained on relevant documents while the testing part has relevant and irrelevant documents. We will investigate the use of irrelevant documents in the training phase to optimise terms weight and thus reduce the impact of the terms that commonly appear in the relevant and irrelevant documents.

(5) The proposed model shows significant performance in selecting relevant features for information filtering. However, its performance in binary text classification is still unknown, and we will investigate it. Also, we will examine our model performance on social media by using short text datasets.

REFERENCES

[1] Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In *Mining text data*. Springer, 77–128.
[2] Mubarak Albathan, Yuefeng Li, and Abdulmohsen Algarni. 2013. *Enhanced N-Gram Extraction Using Relevance Feature Discovery*. Springer International Publishing, Cham, 453–465.
[3] Mubarak Albathan, Yuefeng Li, and Yue Xu. 2014. Using extended random set to find specific patterns. In *WI'14*, Vol. 2. IEEE, 30–37.
[4] Abdulmohsen Algarni and Nasser Tairan. 2014. Feature Selection and Term Weighting. In *WI'14*, Vol. 1. IEEE, 336–339.

[5] Yindalon Aphinyanaphongs, Lawrence D Fu, Zhiguo Li, Eric R Peskin, Efstratios Efstathiadis, Constantin F Aliferis, and Alexander Statnikov. 2014. A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. *Journal of the Association for IST* 65, 10 (2014), 1964–1987.

[6] Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. 2010. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology* 1, 1 (2010), 4–20.

[7] Md Abul Bashar, Yuefeng Li, and Yang Gao. 2016. A Framework for Automatic Personalised Ontology Learning. In *WI'16*. IEEE, 105–112.

[8] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

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

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

[11] Allison June-Barlow Chaney and David M Blei. 2012. Visualizing Topic Models.. In *ICWSM*.

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

[13] George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research* 3, Mar (2003), 1289–1305.

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

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

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

[17] Xiubo Geng, Tie-Yan Liu, Tao Qin, and Hang Li. 2007. Feature selection for ranking. In *SIGIR'07*. ACM, 407–414.

[18] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. 2007. Frequent pattern mining: current status and future directions. *DMKD* 15, 1 (2007), 55–86.

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

[20] Seungil Huh and Stephen E Fienberg. 2012. Discriminative topic modeling based on manifold learning. *ACM Transactions on TKDD* 5, 4 (2012), 20.

[21] Chuanxin Jin, Tinghuai Ma, Rongtao Hou, Meili Tang, Yuan Tian, Abdullah Al-Dhelaan, and Mznah Al-Rodhaan. 2015. Chi-square statistics feature selection based on term frequency and distribution for text categorization. *IETE Journal of Research* 61, 4 (2015), 351–362.

[22] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *KDD'02*. ACM, 133–142.

[23] Mostafa Keikha, Narjes Sharif Razavian, Farhad Oroumchian, and Hassan Seyed Razi. 2008. Document representation and quality of text: An analysis. In *Survey of Text Mining II*. Springer, 219–232.

[24] Eyal Krikon and Oren Kurland. 2011. A study of the integration of passage-, document-, and cluster-based information for re-ranking search results. *Information retrieval* 14, 6 (2011), 593.

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

[26] Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. 2009. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*. 897–904.

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

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

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

[30] Yuefeng Li, Abdulmohsen Algarni, and Yue Xu. 2011. A pattern mining approach for information filtering systems. *Information Retrieval* 14, 3 (2011), 237–256.

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

[32] Tao Liu, Shengping Liu, Zheng Chen, and Wei-Ying Ma. 2003. An evaluation on feature selection for text clustering. In *Icml*, Vol. 3. 488–495.

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

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

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

[36] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. (2002).

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

[38] Carl H Mooney and John F Roddick. 2013. Sequential pattern mining–approaches and algorithms. *ACM CSUR* 45, 2 (2013), 19.

[39] Alessandro Moschitti and Roberto Basili. 2004. Complex linguistic features for text classification: A comprehensive study. In *ECIR'04*. Springer, 181–196.

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

[41] Martin F Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.

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

[43] Stephen E Robertson and Ian Soboroff. 2002. The TREC 2002 Filtering Track Report.. In *TREC*, Vol. 2002. 5.

[44] Joseph John Rocchio. 1971. Relevance feedback in information retrieval. *THE SMART RETRIEVAL SYSTEM* (1971).

[45] Sam Scott and Stan Matwin. 1999. Feature engineering for text classification. In *ICML*, Vol. 99. Citeseer, 379–388.

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

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

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

[49] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. (1996), 267–288.

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

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

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

[53] Sheng-Tang Wu, Yuefeng Li, and Yue Xu. 2006. Deploying approaches for pattern refinement in text mining. In *ICDM'06*. IEEE, 1157–1161.

[54] Sheng-Tang Wu, Yuefeng Li, Yue Xu, Binh Pham, and Phoebe Chen. 2004. Automatic pattern-taxonomy extraction for web mining. In *WI'04*. IEEE, 242–248.

[55] Yutong Wu, Yuefeng Li, Yue Xu, and Wei Huang. 2016. Mining Topically Coherent Patterns for Unsupervised Extractive Multi-document Summarization. In *WI'16*. IEEE, 129–136.

[56] Yue Xu, Yuefeng Li, and Gavin Shaw. 2011. Reliable representations for association rules. *DKE* 70, 6 (2011), 555–575.

[57] Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *Icml*, Vol. 97. 412–420.

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

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

[60] Weidong Zhu and Yongmin Lin. 2013. Using GINI-index for feature weighting in text categorization. (2013).