IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 1, 2020, accepted March 1, 2020, date of publication April 2, 2020, date of current version April 21, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2985079*

# A Semantic Based Approach for Topic Evaluation in Information Filtering

**YUE XU, HANH NGUYEN, AND YUEFENG LI**
School of Computer Science, Faculty of Science and Engineering, Queensland University of Technology, Brisbane, QLD 4001, Australia
Corresponding author: Yue Xu (yue.xu@qut.edu.au)

**ABSTRACT** Topic Modelling has been successfully applied in many text mining applications such as natural language processing, information retrieval, information filtering, etc. In information filtering systems (IFs), user interest representation is the core part which determines the success of the system. Topics in a topic model generated from a user's documents can be used to represent the user's information interest. However, the quality of a topic model generated from a document collection is not always accurate because the topics of the topic model might contain meaningless or ambiguous words. This ambiguity problem can affect the performance of IFs which use a topic model to represent user information interest. Hence, a topic evaluation method to assess the quality of topics in a topic model is important for ensuring the effectiveness of utilizing the topic model in text mining applications. One method in measuring the quality of a topic model is to match the topical words of the model to concepts in an ontology. However, a limitation of this method is that some topical words in an examined topic cannot be found in the mapping ontology. In this study, we propose a new model to evaluate the quality of topics by matching concepts in an ontology. In particular, word embedding technique is applied to dealing with the ambiguity problem by finding similar concept words based on word embeddings. The assessed topics are then used in an information filtering system for filtering relevant documents for a user. The proposed model was evaluated against some state-of-the-art baseline models in terms of term-based, phrase-based, and topic-based user interest representations, and also some topic evaluation models. The result of the evaluation shows that the new proposed model outperforms the state-of-the-art baseline models.

**INDEX TERMS** Information filtering, concept matching, topic evaluation, topic modeling, user interest representation, word embedding.

## I. INTRODUCTION

The past decade has seen the rapid development of topic modelling in understanding text corpus. Among the state-of-the-art models, Latent Dirichlet Allocation LDA [1]–[3] is the most popular technique, which provides an explicit representation of documents. In LDA, documents can be represented by a probability distribution of topics and each topic is a probability distribution of words. The topic model based document representation has been successfully applied to many text mining applications. However, the topics generated by LDA still have limitations. Ambiguous or meaningless topical words and topics were reported in [4] as a common limitation of topic models in general. Many topical words are ambiguous and noisy [4]. This problem originates from LDA algorithm which generates the inferred topics based on high frequent words occuring in the collection. However, frequent words are not necessarily meaningful. Likewise, topics may contain subtopics which cannot be accurately represented by ambiguous topical words. Therefore, evaluating the quality of topics in order to select good topics to be used in text mining applications plays an important role in improving the performance of topic model based text mining applications.

Currently, the studies on topic model evaluation followed two main directions: automatic evaluation and human judgements. A preliminary work on human judgements was conducted by Chang in [4]. That study discovered intrusive topical words (e.g., meaningless, ambiguous, or irrelevant words) based on human judgements. Unlike Chang, Musat in [5] reported a different method called CRSWN for assessing topics automatically based on predefined knowledge

The associate editor coordinating the review of this manuscript and approving it for publication was Huizhi Liang.

provided by WordNet. This model evaluated topics by measuring semantic relevance among topical words based on knowledge from WordNet. One other method for automatically evaluating topics is based on statistics to measure the semantic co-occurrence between topical words in the modelled documents [6]. Even though these existing works could identify intrusive topical words in the examined topics, they used human judgements to evaluate the effectiveness of their proposed models. The model TRbTCM proposed in [7] evaluates topics by mapping the topics to concepts in an external ontology. However, many of the topical words cannot be directly matched with the concepts in ontology. These unmatched words are often some abbreviations, newly created technical terms, compound words, or some rarely used words. These words are often ambiguous or do not have commonly accepted meaning. Therefore, the topics containing such a kind of words cannot be accurately evaluated based on matching concepts in ontologies. A model called sTRbTCM [8] was proposed to deal with the ambiguity problem by using WordNet to find similar words. But the attempt was unsuccessful because the performance of sTRbTCM is worse than that of TRbTCM. In this paper we propose to explore similar words of the unmatched topical words based on word embeddings in order to evaluate the quality of topics more accurately.

Semantic similarity between two words can be used to measure whether the words are similar when applying them in the same context. Word vector representation is getting more attentions recently for representing semantic meanings of words. Word embedding methods have been proposed [9], [10] to learn words' vector representations (i.e., word embeddings) from a large text corpus using neural networks. The similarity between two words can be measured by the similarity between the word embeddings of the two words. In the proposed method in this paper, word embeddings are used to identify semantically similar words to deal with the unmatched topical word problem.

The main contributions of this study are listed as follows:

(1). Firstly, we propose a new method to assess the quality of a topic by mapping the topic to concepts in an external ontology. A new type of patterns, called semantic patterns, is defined based on matching concepts. The quality of topics in a topic model is evaluated based on the semantic patterns and the matching concepts.

(2) Secondly, a method is proposed to deal with unmatched topical words which cannot be matched with any concepts in the mapping ontology. Word embeddings are used to find concepts which share similar semantic meaning with the unmatched topical words.

(3) We also propose a method to measure the relevance of a document to a user's information interest. A topic model is generated from the user's document collection. Each topic is represented by semantically meaningful topical words, including similar concept words. Relevance of an incoming document to the user information interest is measured based on the quality of the topics that are involved in the incoming document.

Through extensive experiments, the proposed model was compared to some existing works in document representations as well as in topic evaluation. We found that the new proposed model performed better than the word-based topic models such as LDA_words [2] and document representations based on terms and phrases [11], [12]. The new proposed model also outperformed some state-of-the-art models in topic evaluation such as CRSWN and CSM in [5], [6].

This paper consists of five sections. The first section is the introduction part. Section II presents related works. The third section describes the proposed topic evaluation method. Section IV presents a new method to rank document relevancy. Section V is about the experimental results and discussion. The conclusion part is presented in Section VI.

## II. RELATED WORKS

Information filtering systems (IFs) aim to find relevant information to satisfy user's information needs. IFs comprise two main parts which are user interest modeling and filtering part [13]. In user interest modelling, some conventional methods based on terms and phrases are widely used to represent user's interest. One of the popularly used term-based models was BM25 [11]. However, this representation conveys polysemy and synonymy as many of single terms express more than one specific meaning. Phrase-based methods such as the one in [14] were proposed to deal with the polysemy and ambiguity problems. Although phrases are considered to be more specific and representative than single words in document representation, phrases still face the problem of low occurrences. For improving the capability of document representation based on terms and phrases, topic models such as PLSA and LDA in [2], [15] were used which provided statistics based models for document representation in which only high distributing terms are used to represent user interest.

Pointwise Mutual Information (PMI) score was studied for automatically evaluating topics in [6]. Correlation score between two topical words in the topic is investigated in [16]. This study introduced a new method in automatic topic evaluation by discovering correlated information among pairs of high frequent words in a topic. Similarly, the work in [6] also studied the coherence between topical words in a topic. This work compares co-occurrence scores of topical word pairs over three different external resources including Wikipedia, WordNet, and Google. This research compared different methods in evaluating topics based on both external resources and occurrences of topical pair words. A study focus on semantic meaning of topical words was proposed in [4]. This work aimed to discover topical words that are meaningless or nosensical. This work used human judgments to evaluate the performance of the proposed model.

Ontologies provide relatively reliable knowledge sources for evaluating the quality of topics. Measuring semantic relevance of a topic to concepts in ontology was studied

in [5]. The main idea of the method called CRSWN is to map topical words with senses in WordNet. This research employed the distance between topical words and concepts in WordNet to measure the relevance between topical words in the ontology. Although this approach can measure the conceptual relevance of the examined topic, the main weakness of this study is the failure to address the co-occurrence between topical words inside each concept in assessing the quality of topics. Topic and concept matching, as used in TRbTCM [7], is a recently proposed approach in assessing the quality of topics. This model measured the quality of a topic based on the overlapping parts between the topic and concepts in the ontology. However, there exist unmatched topical words in the examined topic which are considered meaningless or ambiguous topical words because they do not occur in the mapping ontology. sTRbTCM in [8] was proposed to deal with this problem. The main idea is to find similar words of the unmatched topical words using WordNet and use these similar words to match with concepts. However, the performance of sTRbTCM in terms of information filtering was unsatisfactory, which is worse than the performance of TRbTCM.

Semantic similarity is commonly used for paraphrase detection, information retrieval and document classification as in [17], [18]. In general, there are two main groups of methods for measuring similarity between the two given texts, which are corpus-based methods and knowledge-based methods [18]. In corpus-based methods, similarity between two texts is calculated using information content from a large corpora. These include point wise mutual information for information retrieval, shorted as PMI-IR, suggested in [19]. In PMI-IR, word correlation between two words in the corpus is used to measure similarity between them. Regarding to knowledge-based models, WordNet is mostly used to calculate similarity between concepts. Leacock & Chodorow in [20] measures similarity between two concepts using the depth of the least common subsumer (LCS). Resnik in [21] returns the information content of LCS of two concepts for measuring the semantic similarity between them. Similar to Resnik, Jiang & Conrath [22] measures similarity between two concepts using information content of the two concepts with normalization. For semantic relatedness measures between two concepts, Adapted Lesk is used in [23]. Adapted Lesk extends the concept of overlap to include the glosses of words that are related to the target word and its neighbours according to the concept hierarchies provided in WordNet.

Word representation has been studied recently for automatically measuring similarity between two words or two concepts when they are written lexically differently. One of the first study is contextual representation in [24]. In recent years, neural networks have been widely used for modelling language models. In particular, word embedding methods are getting more attentions from researchers in semantic domain [9], [10], [25]. Word2Vec and GloVe in [10], [25] currently are two popular frameworks. Two widely used word2vec mothods are SKIP_Gram and CBOW which can generate word vector representations to capture syntatic and semantic word relationships. In this paper, we utilize word vectors to find concept words which are most similar to a given unmatched topical word. Another word embedding model is GloVe [25] which generates word vectors based on matrix factorization and local context window.

## III. THE PROPOSED TOPIC EVALUATION MODEL

This paper proposes a model, named Semantic based Topic Evaluation (SbTE), to evaluate the quality of topics generated from a document collection based on the semantics of the documents. The main idea of the new model is to match topics in a topic model with concepts in an ontology to understand the semantic meaning of the examined topics. In this paper, ontology LCSH (Library of Congress Subject Headings, https://catalog.loc.gov/ ) is used. The concepts in LCSH are meaningful phrases because they are well-written by librarians. We believe that the matching between topical words and meaningful concepts in LCSH can interpret the semantic meaning of the examined topics. Like other ontologies, LCSH ontology does not cover all words used in our natural languages. Hence, there certainly exists unmatch problem which occurs when a topical word does not match with any concepts in the ontology but it might have similar meaning with some other concepts. For solving this problem, the basic idea is to find the most similar concept words for the unmatched topical words. The proposed model starts with generating a topic model from the given document collection. Therefore, in Section III.A, topic modeling is discussed. Then Section III.B describes our proposed methods to evaluate the quality of topics in a topic model with the unmatched topical words being taken into consideration as well.

### A. TOPIC MODELS

Topic modelling is a group of algorithms to discover hidden topics in a collection of documents. The basic idea of the generic technique is to find high frequent words to represent the topics in the collection. LDA is one of the popularly used techniques for generating hidden topics. Let $D = \{d_1, d_2, \ldots, d_M\}$ be a collection of $M$ documents. The main idea of LDA is that a document is a multinomial distribution over topics. Each topic is a multinomial distribution over words. At document level, each document is represented by topic distribution $\theta_d = \{V_{d,1}, V_{d,2}, \ldots, V_{d,v}\}$, $\sum_{j=1}^{v} V_{d,j} = 1$, $V_{d,j} = P(z_j|d)$, $v$ is the number of topics. In the collection level, $D$ is represented by a set of topics. Each topic is represented by a probability distribution over words. For the $j^{th}$ topic, we have $\Phi_j = \{\phi_{j1}, \phi_{j2}, \ldots, \phi_{jm}\}$, $m$ is the number of words per topic, $\phi_{ji} = P(w_i|z_j)$. The probability of word $w_i$ in the document $d$ can be calculated as $P(w_i|d) = \sum_{j=1}^{v} P(w_i|z_j) * P(z_j|d)$. In terms of words, each topic $z$ is represented as a set of words, denoted as $\mathbb{T}(z) = \{w_1, w_2, \ldots, w_m\}$. The assignments of words to topics mainly based on the probability distribution in which words with high probabilities are sampled as topical words. As a result of this, some of the topical words may be ambiguous or meaningless

even they occur frequently, and thus affect the quality of the topic model. This study aims to assess the quality of topic model by matching topical words with concepts in an external ontology to ensure that the topical words are semantically meaningful. Topical words that match with concepts in the ontology are called matched topical words or explained topical words. The unmatched topical words, also called as unexplained topical words, have no overlapping with any concepts in the ontology. Unmatched topical words can be meaningless or ambiguous. For interpreting the unmatched topical words, based on word embeddings, we can eliminate meaningless words if no similar words can be found and find semantically similar concept words for the ambiguous topical words if any similar words can be found. With the similar concept words, these ambiguous unmatched topical words can be matched with some concepts in the ontology. In next section, the detail of our proposed method to evaluate topic quality based on matching concepts will be explained.

### B. TOPIC EVALUATION BASED ON MATCHING CONCEPTS

As mentioned above, a topic in a topic model is represented by a probability distribution of words, and single words have polysemy and synonymy problems. Patterns are usually considered expressing more specific meaning and less ambiguous. Therefore, in this paper, we propose to generate patterns from the topical words of a topic based on matching concepts in an ontology. These patterns and the matched concepts can be used to evaluate the meaningfulness of the topic. To this end, we define a new type of patterns called Semantic Patterns.

#### 1) SEMANTIC PATTERNS AND MATCHED CONCEPTS

*Definition 1 (Ontologies):* Ontologies can be understood as the concepts of entities that represent human knowledge about things in real world. Ontology can be presented in a tuple $O =< \mathcal{C}, R >$ such that $\mathcal{C}$ is a set of concepts; $R$ is a set of relations.

In LCSH ontology, $\mathcal{C}$ consists of subject headings; $R$ comprises relations between subject headings such as hierarchical, equivalent and association relationships. In this paper, only the concepts of LCSH are used.

*Definition 2 (Matched Concepts):* Given a topic $z$ with its topical words denoted as $\mathbb{T}(z)$, a list of matched concepts between the topic $z$ and concepts in $\mathcal{C}$ of an ontology, denoted as $\Gamma(z)$, is defined below:

$$\Gamma(z) = \{c | c \in \mathcal{C}, c \cap \mathbb{T}(z) \neq \emptyset\} \tag{1}$$

From the definition, a matched concept $c$ in $\Gamma(z)$ obviously has at least one topical word of the topic $z$. For example, given a topic $\mathbb{T}(z) = \{computer, system, data, dutroux\}$, and two concepts: $c_1 =$ "*Computer hardware*" and $c_2 =$ "*Computer system security*". Because both $c_1$ and $c_2$ contain the word "*Computer*" which occurs in topic $z$, both $c_1$ and $c_2$ are matched concepts for the topic, i.e., $\Gamma(z) = \{c_1, c_2, \ldots\}$.

*Definition 3 (Semantic Patterns):* Semantic patterns of a topic $z$ over matched concepts $\Gamma(z)$, denoted as $\mathcal{SP}(z)$, is defined as:

$$\mathcal{SP}(z) = \{p | c \in \Gamma(z), p = c \cap \mathbb{T}(z), p \neq \emptyset\} \tag{2}$$

A semantic pattern $p \in \mathcal{SP}(z)$ is the overlapping part between the topic $z$ and a matched concept $c, c \in \Gamma(z)$. Semantic patterns indicate how much the topic can be explained by the ontology. For example, given a topic $\mathbb{T}(z) = \{computer, system, data, dutroux\}$, and two concepts: $c_1 =$ "*Computer hardware*" and $c_2 =$ "*Computer system security*", one semantic pattern is [*computer*] because it is an overlapping part of $c_1$ and $z$; another semantic pattern is [*computer*, *system*] due to the overlapping between $c_2$ and $z$. These semantic patterns indicate that the mapping ontology can explain these topical words: {*computer*, *system*} in the topic $z$ because the words are presented in the mapping ontology. However, the topical word "*dutroux*" can not be explained because the term "*dutroux*" does not occur in any concepts in the ontology.

The concept matching method can be illustrated in FIGURE 1. In the figure, both concepts $c_2$ and $c_3$ match with topic $z_1$ with two words $w_{11}$ and $w_{12}$. Therefore, $c_2$ and $c_3$ are two matched concepts of topic $z_1$ and $\{w_{11}, w_{12}\}$ is a semantic pattern of topic $z_1$. This means, the two topical words are meaningful because they can be explained by the two matched concepts. The semantic patterns of a topic can have super or sub patterns such as pattern $\{w_{12}\}$ which is a sub pattern of $\{w_{11}, w_{12}\}$, explained by matched concept $c_1$. The purpose of the concept matching is to ensure that the topical words of a topic can be explained by concepts. It is reasonable to look for longer patterns because they can cover more topical words. The maximum semantic patterns defined below are such a kind of patterns which are the longest patterns without super patterns.

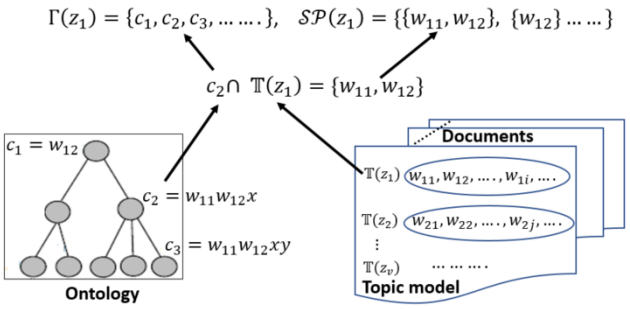


**FIGURE 1.** Generating semantic patterns.

*Definition 4 (Maximum Semantic Patterns):* Maximum semantic patterns of a topic $z$, denoted as $MaxSP(z)$, is defined as:

$$MaxSP(z) = \{p | p \in \mathcal{SP}(z), \nexists \hat{p} \in \mathcal{SP}(z), p \subset \hat{p}\} \tag{3}$$

A pattern $p$ in $MaxSP(z)$ is maximum, i.e., it does not have any super patterns in $MaxSP(z)$. This means that all patterns in $MaxSP(z)$ are the longest patterns. Obviously, a longest semantic pattern for a topic has the highest number of words

in the topic that are matched with a concept. From the concept perspective, a concept that covers a longest pattern of a topic would represent the meaning of the topic more closely than concepts that cover shorter patterns.

For example, given three semantic patterns $\mathcal{SP}(z) = \{[computer], [system], [computer, system]\}$, the pattern $[computer, system]$ is the longest pattern among the three patterns, covering both patterns $[computer]$ and $[system]$. In concept meaning perspective, concepts that cover the pattern $[computer, system]$ convey more specific information than concepts that cover shorter patterns like $[computer]$ or $[system]$. For instance, a concept "*Computer System Security*" which covers the pattern $[computer, system]$ is more specific than concept "*Computer*".

If a pattern can be matched or explained by multiple concepts such as pattern $\{w_{11}, w_{12}\}$ in FIGURE 1 which can be explained by concepts $c_2$ and $c_3$. Among the two concepts, $c_2$ would be more close to $\{w_{11}, w_{12}\}$ in terms of their meaning than $c_3$ is, because $c_2$ has a higher percentage of its words (2 out of 3) matched with $\{w_{11}, w_{12}\}$ than $c_3$ (2 out of 4 words) does. Therefore, the shortest concept which matches with a maximum semantic pattern can be considered most closely explain the pattern.

*Definition 5 (Closest Matched Concepts):* Closest Matched Concepts of a topic $z$, denoted as $\mathbb{CMC}(z)$, are the shortest concepts in ontology that closely cover the maximum semantic patterns in $MaxSP(z)$. $\mathbb{CMC}(z)$ is a subset of matched concepts $\Gamma(z)$ and defined below:

$$\mathbb{CMC}(z) = \{c | c \in \Gamma(z), \exists p \in MaxSP(z), \\ p \subset c, \nexists \hat{c} \in \Gamma(z), \hat{c} \subset c \;\; p \subset \hat{c}\} \quad (4)$$

A concept $c$ in $\mathbb{CMC}(z)$ covers at least one of the longest semantic patterns and it does not has a sub concept which also covers the same longest semantic pattern. In other words, each concept in $\mathbb{CMC}(z)$ must be the shortest concept which covers one of the patterns in $MaxSP(z)$. Assume $\{w_{11}, w_{12}\}$ in FIGURE 1 is a maximum semantic pattern, $c_2$ would be the closest matched concept of $\{w_{11}, w_{12}\}$ if $c_2$ and $c_3$ are the only two concepts of this pattern.

### 2) DISAMBIGUATION BY FINDING SIMILAR WORDS

Although concepts in an ontology contain a relatively large number of terms, they still cannot cover all the words in text documents. For example, the topical word "*dutroux*" does not occur in the mapping ontology. The underlying reason is that the topical words might be meaningless or lexically written differently to concepts in the mapping ontology or newly created novel words which have not been collected in the ontology. For instance, the topical word "*newsroom*" is said to be unmatched word as it does not occur in any concepts in the ontology. However, that word is very similar to a concept word "*editor*" which is in the ontology. Hence, one solution for the mentioned problem is to search for concept words which are most similar to the unmatched topical words. This solution is believed to enhance the interpretation of the unmatched topical words in the examined topic. In this study, word embeddings are used to represent unmatched topical words and concept words. Based on the word embeddings' similarity, similar concept words of an unmatched topical word can be found.

Given a topic $z$ with its topical words $\mathbb{T}(z)$, we can divide $\mathbb{T}(z)$ into two separate sets: a set of matched topical words, denoted as $MT(z)$, each of which shares a part with at least one concept in the ontology; and a set of unmatched topic words, denoted as $UT(z)$, which do not overlap with any concepts in the ontology. The topical words in $UT(z)$ would be considered meaningless or ambiguous. However, if there exist similar words in the ontology which are similar with the unmatched topical words in $UT(z)$, these unmatched words can be considered as explainable topical words.

Let $CW = \{w^c | w^c \in c, c \subset \mathcal{C}\}$ be all the concept words of all the LCSH concepts after removing special characters and stop words such as: the, of, about, in, etc., let $sim(w, w^c)$ be a semantic similarity between a topical word $w$ of topic $z$ and a concept word $w^c$, a set of concept words that are similar to the topical word $w$ with the maximum similarity, denoted as $sw_z(w)$, is defined below, where $sw_z(w).sim$ is the similarity value between $w$ and its similar words in $sw_z(w)$.

$$sw_z(w) = argmax_{w^c \in CW}(sim(w, w^c)) \\ sw_z(w).sim = max_{w^c \in CW}(sim(w, w^c)) \quad (5)$$

Similarity is symetric. For each $w^c \in sw_z(w)$, $sw_z(w^c).sim = sw_z(w).sim$. For a concept word $w^c$, it might be a similar concept word for unmatched topical words in different topics. $sw_z(w^c).sim$ provides the similarity value between $w^c$ and a corresponding unmatched word in topic $z$. Similarly, $w^c$ could be a similar concept word for multiple unmatched topical words in the same topic. In this case, $sw_z(w^c).sim$ is set to the maximum similarity among those unmatched topical words in this topic, as defined below.

$$sw_z(w^c).sim = max_{w \in \mathbb{T}(z), w^c \in sw_z(w)}(sw_z(w).sim)$$

The topical word $w$ could be a matched topical word or an unmatched topical word. When $w$ is a matched topical word, because $w$ is in $CW$, therefore, $sw_z(w) = \{w\}$ and $sw_z(w).sim = 1$. As a topical word might be similar to multiple concept words with the same maximum value, the number of similar concept words to the topical word can be greater than 1. On the other hand, there could be no similar concept words to the topical word, i.e., $|sw_z(w)| \geq 0$. If $|sw_z(w)| = 0$, $w$ is considered as a meaningless word because it doesn't have any similar words, otherwise, $w$ is considered as a matched word. Especially, when $sw_z(w).sim < 1$, $w$ would be a matched ambiguous word, $sw_z(w)$ provides its similar words with the same maximum similarity. The next section presents how to compute similarity between two words based on word embeddings.

#### 3) SEMANTIC SIMILARITY USING WORD EMBEDDINGS

In everyday conversation, two words are synonymous if they convey the same meaning in a certain context. However, how to determine whether two words are synonymous or not if the two words are written differently is a problem needing solutions. People believe that contextual information can help to determine synonymous. Contextual information of a word is the context that the word is used [24]. Recently, word embedding methods have been proposed to generate word representations, i.e., word embeddings, such as methods Skip-gram and CBOW [15]. The word vector representations are called word embeddings. In this study, Skip-Gram is used to generate word embeddings to represent concept words and unmatched topical words.

The learnt word vectors can be used to measure similarity between two words. Specifically, let two vectors $W_z$ and $W_c$ represent two words $w_z$ and $w_c$, in the experiments reported in Section V, the cosine similarity is applied over $W_z$ and $W_c$ for computing similarity between the two words,, i.e., $sim(w_z, w_c)$, used in Equation 5.

#### 4) TOPIC QUALITY

Let $SW(UT(z))$ be the set of similar concept words for a given set of unmatched topical words $UT(z)$, $SW(UT(z))$ is defined as following:

$$SW(UT(z)) = \bigcup_{w \in UT(z)} sw_z(w) \tag{6}$$

After searching for the most similar concept words for all the unmatched topical words in $UT(z)$, the topical words in the examined topic $z$ will be extended by adding the similar concept words to the topic, and all the unmatched topical words are excluded from the topic. The modified set of topical words for topic $z$, denoted as $\mathbb{T}^*(z)$, is defined as:

$$\mathbb{T}^*(z) = (\mathbb{T}(z) - UT(z)) \cup SW(UT(z)) \tag{7}$$

As the set of topical words in topic $z$ has changed to $\mathbb{T}^*(z)$, the matched concepts, semantic patterns and maximum semantic patterns generated using equations 1, 2, and 3 are all generated based on $\mathbb{T}^*(z)$, and the generated closest matched concepts based on $\mathbb{T}^*(z)$ will contain both directly matched concept words and similar concept words. Let $\mathbb{CMC}^*(z)$ be the closest matched concepts based on $\mathbb{T}^*(z)$. In this paper, we propose to measure the quality of a topic $z$, denoted as $\mathcal{Q}^*(z)$, based on the matched concepts in $\mathbb{CMC}^*(z)$, as defined below:

$$\mathcal{Q}^*(z) = [\frac{1}{|\mathbb{CMC}^*(z)|} \sum_{c \in \mathbb{CMC}^*(z)} \mathcal{M}(c)] \times E(z) \tag{8}$$

There are two parts in $\mathcal{Q}^*(z)$. In the first part, $\mathcal{M}(c)$ measures the percentage of concept words in a matched concept $c$ that match with the topical words. $\mathcal{M}(c)$ is defined below:

$$\mathcal{M}(c) = \frac{|c \cap \mathbb{T}^*(z)| \times sim(c, \mathbb{T}^*(z))}{|c|} \tag{9}$$

where $sim(c, \mathbb{T}^*(z))$ is the average similarity that $c$ to the topical words in $\mathbb{T}^*(z)$. $sim(c, \mathbb{T}^*(z))$ is defined as

$$sim(c, \mathbb{T}^*(z)) = Avg_{w \in c \cap \mathbb{T}^*(z)}(sw_z(w).sim)$$

By taking all the closest matched concepts in $\mathbb{CMC}^*(z)$ into consideration, the first part measures the average percentage of concept words that match with topical words in topic $z$. The higher the percentage, the more relevant the matched concepts to the topic. In the extreme case when all matched concept words are topical words, the first part would equal to 1, which means all the matched concept words are topical words without ambiguous words. This is almost impossible since there are always some concept words which are not topical words, e.g., word *security* in concept $c_2$ of the example in Section III.B.1 is not a topical word, while $c_2$ is a closest matched concept in that example.

In an inverse manner, the second part, defined in Equation (10), is to measure the percentage of topical words which can be explained by the matched concepts.

$$E(z) = \frac{|\bigcup_{c \in \mathbb{CMC}^*(z)} c \cap \mathbb{T}^*(z)|}{|\mathbb{T}^*(z)|} \tag{10}$$

The higher the $E(z)$ is, the more the topical words in $z$ that can be explained by concepts. In the extreme case when all topical words occur in some of the matched concepts, the second part would equal to 1.

Algorithm 1 below depicts the major steps for measuring the quality of a given topic $z$ by incorporating semantic similarity to deal with unmatched topical words.

**Algorithm 1** Algorithm 1: Topic Quality Measurement

**Input:** topic $z$, Google embedding set,
Concepts in LCSH $\mathcal{C}$
**Output:** Quality of topic $z$, $Q^*(z)$
**Initialize:**
Parse LCSH concepts for concept words $CW$
Topical words in topic $z$ is $\mathbb{T}(z)$
1. Find Unmatched topical words $UT(z)$.
2. Find similar words $SW(UT(z))$ for the unmatched topical words
$UT(z)$ based on word embeddings using equations (5) and (6).
3. Expand set of words in topic $z$ with $SW(UT(z))$
$\mathbb{T}^*(z) = (\mathbb{T}(z) - UT(z)) \cup SW(UT(z))$.
4. Mapping words in $\mathbb{T}^*(z)$ to concepts $\mathcal{C}$ in the ontology LCSH
Generate set of matched concepts $\Gamma^*(z)$, using Equation (1).
5. Generate semantic patterns $\mathcal{SP}^*(z)$ using Equation (2)
6. Calculate Maximum semantic Patterns $MaxSP^*(z)$ using Equation (3)
7. Calculate Closest Matched Concept $\mathbb{CMC}^*(z)$ using Equation (4)
8. Calculate quality of topic $Q^*(z)$ using Equation (8).
9. **End.**

**TABLE 1.** Some extracted subject headings in LCSH ontology.

| ID | Subject headings in LCSH |
|---|---|
| $C_1$ | Reporting |
| $C_2$ | Editorals |
| $C_3$ | Presidents |
| $C_4$ | Labor market–Effect of international trade |
| $C_5$ | Regionally important geological geomorphological sites (Great Britain) |
| $C_6$ | Retail trade–Marketing |
| $C_7$ | RIGGS (Regionally important geological and geomorphological sites) |
| $C_8$ | Important bird areas |

#### 5) EXAMPLE OF COMPUTING TOPIC QUALITY

Given a topic $z$ with a set of terms $\mathbb{T}(z) = \{$"*trade*", "*newsroom*", "*president*", "*markets*", "*reporters*", "*important*"$\}$.

There is one unmatched topical word which is *newsroom*. By applying semantic similarity method over the unmatched topical word *newsroom*, the concept word *editor* is similar to *newsroom* with similarity value is 0.560. Finally, the set of closest matched concepts after solving the unmatched topical words in the given topic is $\mathbb{CMC}^*(z) = \{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8\}$, as shown in Table 1.

The maximum semantic patterns for the model is $MaxSP^*(z) = \{[market, trade], [important], [report], [president], [editor]\}$. There is one maximum semantic pattern, $[editor]$, which is the most similar word to the unmatched topic word *newsroom* according to their word embeddings. Finally, the quality of topic $z$ is valued as 0.477. The quality of this topic is not very high. The reason is mainly because of the word *important*, which cannot be well explained. *important* is an adjective and matched to three concepts $C_5, C_7, C_8$. Each of the three concepts contains many non-topical words, which will largely reduce the value of the first part of $\mathbb{Q}^*(z)$ and thus reduces $\mathbb{Q}^*(z)$ quite significantly.

A similar example is showed in Table 3 in Section V (i.e., the experiment section). Topic $z_2$ has a low quality of 0.265 ('*important*' is one of its topical words) even all of its topical words can be matched with concepts in the ontology. However, because some of the matched concepts contain many non-topical words meaning that these concepts may not be able to explain the topic, the topic $z_2$ is not evaluated with a high quality value.

## IV. DOCUMENT RANKING MODEL

Information Filtering systems (IFs) aim to retrieve information that satisfies user's needs in information. In this paper, topics generated from a user's training documents by LDA are used to represent the user's interest. For a new incoming document $d$, the basic idea is to determine the relevance of the document $d$ to the user's interest based on the explained topical words occurring in $d$. Specifically, in this section, we propose a ranking method to rank incoming documents based on three topic related measures: topic probability distribution, quality of topics and the significance of topics in the examined document $d$. The ranking scores are used to filter relevant documents from an incoming document stream.

*Topic Probability Distribution:* Let's $V_{D,j}$ be the average topic probability distribution of all documents in the training collection $D$, $\theta_D = (V_{D,1}, V_{D,2}, \ldots, V_{D,v})$, $\sum_{j=1}^{v} V_{D,j} = 1$ and $V_{D,j}$ is measured as:

$$V_{D,j} = \frac{1}{|D|} \sum_{d \in D} P_r(z_j|d) \tag{11}$$

*Quality of a Topic:* The quality of a topic measures how accurate and representative the topic to represent the semantic content of the training collection. As explained in section III.B, the quality of a topic is measured based on the topic's matched concepts in the ontology calculated using Equation (8).

*Significance of Topic Based on the Explained Topical Words:* The significance of topic $z$ in document $d$ is measured by Equation (12) below where $ET(z)$ is the set of explained topical words including the matched topical words and ambiguous topical words which have similar concept words in the ontology. Unexplained topical words will not be included in $ET(z)$ and thus removed. $ET(z)$ represents the user's information interest. Significance of topical word $w_i$ in topic $z$, denoted as $sig(w_i|z)$, is defined as $sig(w_i|z) = m_i * Pr(w_i|z)$, $m_i = Pr(w_i|z)/avgPr(z)$, $Pr(w_i|z)$ is the probability of $w_i$ in topic $z$, $avgPr(z)$ is the average probability of topical words of $z$. In this study, the topical words with $m_i > 1$ are selected to represent the topic, i.e., the topical words whose probability is larger than the average probability.

$$sig(z, d) = \sum_{\substack{w_i \in d, w_i \in ET(z) \\ pr(w_i) > avgPr(z)}} sig(w_i|z) \tag{12}$$

For a new incoming document $d$, the relevance score of $d$ to the training collection $D$ with $v$ topics is measured using Equation (13) as follows:

$$rank(d|D) = \sum_{j=1}^{v} sig(z_j, d) \times \mathbb{Q}^*(z_j) \times V_{D,j} \tag{13}$$

## V. EXPERIMENTS

These experiments are designed for verifying the proposed topic evaluation model SbTE. The correctness of the topic quality calculated using Equation (8) will be assessed in terms of information filtering using Equation (13) to rank and filter relevant documents. There are two aims in the experiments. The first aim is to verify that word embedding based disambiguation can improve the meaningfulness of topic representations. The second aim is to verify that the proposed document relevance ranking based on explained topical words can improve the performance of IFs. The following subsections will explain the datasets and baseline models first, then the experimental results and discussion.

### A. DATASETS

LCSH is a large ontology which is built up and regularly updated over a long period of time, covering a large amount

of information over many different domains. Information in the ontology is organized in terms of meaningful phrases called concepts or subject headings. These concepts are usually coded by librarians. In the experiments, a database [26] containing 498474 LCSH subject headings is used. This is a raw and new RDF file which is updated in 2017. In this paper, concepts and subject headings are used interchangely.

The Reuter Corpus Volume 1 (RCV1) dataset [27] includes articles, collected by Reuters from the year of 1996 to 1997. This dataset covers a variation of domains and comprises of a large number of documents, totally including 806,791 news stories. Dataset RCV1 comprises of 100 collections and divided into two parts: the training set and the testing set. The first 50 collections were composed by human assessors and the remained collections were generated by artificially combining the remained collections together. In this paper, we used the first 50 collections for the experiments.

### B. BASELINE MODELS

The proposed model involves topic evaluation and user interest representation. Therefore, three groups of existing methods are chosen as baseline models to compare with the proposed model in terms of topic evaluation, phrase-based representation and term-based representation. The first group is about topic evaluation models, including TRbTCM, sTRbTCM, CRSWN and CSM. Phrase-based representation TNG is used in the second group. The third group is term-based representation, containing models BM25, PLSA_words and LDA_words. Details of the baseline models are described as below:

#### 1) TOPIC EVALUATION MODELS

**TRbTCM** [7]: TRbTCM evaluates topic quality based on matching concepts of LCSH. However, it uses only matched topical words and ignores unmatched topical words.

**sTRbTCM** [8]: Similar to TRbTCM, sTRbTCM evaluates topic quality based on matching concepts of LCSH. It takes unmatched topical words into consideration by using WordNet to solve the ambiguity problem.

**CSM** [6]: Topic evaluation based on co-occurrence between topical words. CSM measures correlation between pairs of topic words to evaluate the topic. In short, given a topic $z$, correlation score between a topical word $w$ and other topical words in $z$ is defined as $C(z) = \sum_{m=2}^{M}\sum_{l=1}^{m-1} log\frac{D(t_m^{(z)},t_l^{(z)})+1}{D(t_l^{w})}$ where $D(t_1, t_2)$ is the document frequency of pair $(t_1, t_2)$ in the examined collection; $t_1$ and $t_2$ are topical words of $z$. $D(t)$ is the document frequency of word $t$ in the collection. Then, average co_occurrence scores over all topics in the collection $D$ is denoted as $T_D$, where $v$ is the number of topics. $T_D$ is defined as: $T_D = \frac{1}{v}\sum_{j=1}^{v} C(z_j)$.

Originally, the study in [6] used $T_D$ to compare the quality of the examined topics. Then, human experts are invited to evaluate the topic quality assessment. In our experiments, the accuracy of topic evaluation was assessed by incorporating the quality of topics into document relevancy ranking in information filtering systems.

The relevance ranking model for document $d$ is defined as: $rank(d|D) = \sum_{j=1}^{v} sig(z_j, d) \times V_{D,j} \times T_D$

**CRSWN** [5]: Topic evaluation based on relevance score among topical words. This model uses WordNet as the mapping ontology. CRSWN calculates the relative relevance score of the examined topic based on the distances between matched concepts which contain topical words in the WordNet ontology. Readers can refer to the study in [5] for detail. In summary, given a topic $z$, each topical word $w \in z$ is considered as a concept $c$ in WordNet ontology. Relevance score between two topic words $w_1, w_2$ is calculated via concepts $c_1, c_2$ by this formula: $\phi(c_1, c_2) = w_{cov} * cov(c_1, c_2) + w_{spec} * spec(c_1, c_2)$.

$cov(c_1, c_2)$ is the coverage of a concept $c_1$ over the concept $c_2$, $cov(c_1, c_2) = \frac{|\delta(c_1)\cap\delta(c_2)|}{\delta(c_1)}$, $\delta(c)$ contains all the concepts in WordNet that lead to the concept $c$. The specificity of the concept $c_1$ over concept $c_2$, denoted as $spec(c_1, c_2)$, is defined as $spec(c_1, c_2) = w_h * height(c_1) + w_p * depth(c_1, c_2)$ where $w_h = 0.5$ and $w_p = 0.5$; $height(c_1)$ is the height of the concept and $depth(c_1, c_2)$ is the distance from concept $c_1$ to concept $c_2$. We also set $w_{cov} = 0.5$ and $w_{spec} = 0.5$.

Similar to CSM model, human judgements were used for evaluating the performance of the model. In our experiments, we evaluated the performance of topic evaluation by applying the assessed topics to document ranking in information filtering systems.

The average relevance score of all matched concepts with topical words in $z$ is calculated as $avgRS(z) = \frac{1}{|z|}\sum_{t\in z}\frac{1}{|\delta(t)|}\sum_{c_t\in\delta(t)}\phi(c_t, t)$ where $\delta(t)$ contains all the concepts that lead to the topical word $t$ which is also a concept in WordNet.

$rank(d|D) = \sum_{j=1}^{v} sig(z_j, d) \times V_{D,j} \times avgRS(z_j)$

#### 2) PHRASE-BASED REPRESENTATION

**TNG**: This is a phrase based topic model, n-grams phrases are generated by using the TNG model [12]. In this model, phrases are used to represent user's interest.

#### 3) TERM-BASED REPRESENTATIONS

**LDA_Words**: topical words in LDA topic model are used to represent users' interest [1].

**PLSA_Words**: topical words in pLSA topic models are used to represent users' interest [15].

**BM25**: BM25 [11] is the state-of-the-art model in document representation, in which term $t$ is measured using this following equation: $w(t) = \frac{tf\times(k+1)}{k\times((1-b)+b\frac{DL}{AVDL})+tf} \times log(\frac{N-n+0.5}{n+0.5})$ where $N$ is the total number of documents in the collection; $n$ is the number of documents that contain term $t$; $tf$ is term frequency; $DL$ and $AVDL$ are document length and average document length, respectively; and $k$ and $b$ are the parameters, which are set as 1.2 and 0.75 as used and explained in [28]

### C. EXPERIMENTAL SETTINGS

In the proposed model, for a given document collection which contains a user's interest information, a topic model is firstly generated, and the topical words of the topic model are used to represent user interest; significance of a given topic for an incomming document is estimated using Equation (12), and the relevance of the document to the user's interest is measured using Equation (13).

The MALLET toolkit [29] was used to generate LDA topic models. The initial parameter settings for LDA include $\alpha = 0.5$; *and* $\beta = 0.01$. The number of topics is $v = 10$. For different topics, different number of topical words were chosen depending on the probability distribution over words in that topic. The chosen topical words for a topic in our proposed model are words with probabilities higher than the average word probability of that topic and the maximum number of words per topic is 20. Specifically, the $i^{th}$ topical word in topic $z$ is selected if $Pr(w_i|z) > avgPr(z) * \gamma$, where $avgPr(z)$ is the average word probability in topic $z$, $\gamma = 0.8$.

Similarity between two words is measured using cosine similarity between the two word embeddings. Given two words $w_c$ and $w_z$, which are a LCSH conceptual word and a unmatched topical word correspondingly, the corresponding word embeddings $W_c$ and $W_z$ can be obtained from Google pretrained vectors.[1] The vectors, each of which has 300 dimensions in length and is trained using documents containing over 100 billion words, are considered the highest quality word embeddings as mentioned in [10].

### D. EVALUATION MEASUREMENT

In these experiments, four main evaluation metrics are used to compare performances of the models. The Top-K score evaluates the precision for the first K retrieved documents. In these experiments, Top-20 is used. Mean Average Precision (*MAP*) measures precision at each relevant document first, and averaging precision over all topics afterwards. *MAP* metric provides a very succinct summary of the effectiveness of a ranking algorithm over many different queries. The break-even point $b/p$ indicates the points where precision and recall are equal. This score measures the effectiveness of the system. The higher this value of $b/p$, the better the implemented system. $F1$ scores reflect the harmonic average of the precision and recall. $F1$ emphasizes the effectiveness of retrieved documents.

### E. RESULTS AND DISCUSSION

#### 1) DISAMBIGUATION USING WORD EMBEDDINGS

In this section, we will observe the effectiveness of using word embeddings for assessing topic quality and to boost the performance of information filtering.

Firstly, some statistic information about the ambiguity occurring in the topic model generated from the first 50 collections in dataset RCV1 was analyzed. There are totally 3305 unique topical words in the topic models generated from the 50 collections. Among them, there are about 399 unmatched topical words which could not be matched with any concept words in LCSH ontology. The percentage of unmatched topical words are about 12% of the total number of topical words in the topic models. Most of these unmatched topical words are in the form of verb, variations of verb form, adjective and adverb. For example, some topical words like "*noted*", "*shortly*", "*virtually*", "*unhappy*" are unmatched topical words. Many of the unmatched topical words, which have no similar concept words, are considered as "meaningless" topical words. For dataset RCV1, out of the 399 unmatched topical words, 214 do not have similar words, which is 54%. Table 2 shows some examples of unmatched topical words, similar words, and their similarity values from the training collections.

**TABLE 2.** Examples of similarity between pairs of words: Term $w_z$ is unmatched topical word; Term $w_c$ is the most similar word.

| Term $w_z$ | Term $w_c$ | $Sim(w_z, w_c)$ |
|---|---|---|
| noted | acknowledged | 0.690069973 |
| gartner | apis | 0.470424443 |
| biotech | biotechnology | 0.876297534 |
| shortly | after | 0.612630188 |
| ceo | chairman | 0.528498232 |
| virtually | almost | 0.647622406 |
| unhappy | angry | 0.6646263 |
| predecessor | applied | 0.752986372 |
| apply | applied | 0.752986372 |
| plaintiff | defendant | 0.648491681 |

Let consider the training collection 117 in dataset RCV1. A LDA topic model with 10 topics is generated from this collection and the 10 topics are listed in the left column of TABLE 3. The other two columns in TABLE 3 list the topic quality measures calculated by using or without using disambiguation. TABLE 4 shows a list of unmatched topical words for each corresponding topic in the collection, found similar concept words and similarity values between unmatched topic words and their most similar words.

As can be seen from the TABLE 4, there are about a half of the unmatched topical words which exist no similar concept words, written as "*NA*". Those words are considered meaningless topical words and removed from the user's interest representation in the later filtering stage of IFs. Even though only some of the unmatched words become explainable based on word embeddings and concepts in ontology, the interpretation of those unmatched words can lift the quality of the corresponding topics and thus make those topics more important than before to represent users' information interest. For example, topic $z_3$ in TABLE 4 has 4 unmatched topical words, three of them can be explained by similar words. This makes the topic quality of $z_3$ increased from 0.663 to 0.773 as shown in TABLE 3. But for topic $z_8$, none of its three unmatched topical words has similar words and thus this topic's quality measure keeps intact.

[1] https://code.google.com/archive/p/word2vec/

**TABLE 3.** List of topics in training collection 117 in dataset RCV1 and topic quality using disambiguation $Q^*(z)$ and without using disambiguation $Q(z)$.

| Topics $z$ | Q*(z) | Q(z) |
|---|---|---|
| $z_1 = \{human, pig, animal, disease\}$ | 0.533 | 0.533 |
| $z_2 = \{affect, finding, viruse, science, molecular, concern, mapped, movement, working, blindness, suffering, shortage, fibrosis, james, cystic, important, alternative\}$ | 0.265 | 0.265 |
| $z_3 = \{transplant, drug, market, year, stop, organ, rejecting, halt, inaccessible, convinced, likelihood, smithkline, dying, voluntary, broadcast, salmon, rejection\}$ | **0.773** | 0.663 |
| $z_4 = \{(ppl, organ, scientist, transplant, research, protein, suitable, human, reduce, animaltohuman, collapse, include, potential, largest\}$ | **0.408** | 0.385 |
| $z_5 = \{aid, virus, report, spte, protease, enzyme, rise, information, animal, herpes, director\}$ | 0.409 | 0.409 |
| $z_6 = \{engine, saving, treating, leading, eye, journal, difficulty, series, fear, shingles, beecham, cycle, suspicions, kidney, member, effective, plc, disposal, attack, timetable\}$ | **0.358** | 0.338 |
| $z_7 = \{company, attack, number, pension, expected, result, internet, technology, hope, clone, report, financial, scottish, monitoring, sale, single, tesco, employee, cause, week\}$ | **0.387** | 0.369 |
| $z_8 = \{family, heart, programme, usa, raise, xenograph, monsantosearle, recipient, access, inhibit, medical, school, prospect, book, statement, boehringer, pennsylvania, bid, therapies, studied\}$ | 0.364 | 0.364 |
| $z_9 = \{stg, million, percent, , group, service, therapeutics, claim, ferry, share, british, bmw, automotive, pension, deal, ayling, wednesday, year, scheme, announce, chairman, \}$ | **0.462** | 0.434 |
| $z_{10} = \{cmv, patient, researcher, problem, pharmaceutical\}$ | 0.402 | 0.402 |

**TABLE 4.** List of unmatched topical words $w_z$, found similar concept words $w_s$, and their similarity value *sim_val* in collection 117.

| z | $w_z$ | $w_s$ | $sim\_val$ |
|---|---|---|---|
| $z_1$ | NA | NA | NA |
| $z_2$ | NA | NA | NA |
| $z_3$ | halt | stop | 0.60826385 |
| | smithkline | NA | NA |
| | convinced | determined | 0.526270926 |
| | likelihood | probability | 0.818094134 |
| $z_4$ | animaltohuman | NA | NA |
| | largest | smallest | 0.560892045 |
| $z_5$ | spte | NA | NA |
| $z_6$ | timetable | blueprint | 0.488592535 |
| | beecham | NA | NA |
| $z_7$ | tesco | birmingham | 0.588748097 |
| $z_8$ | boehringer | NA | NA |
| | monsantosearle | NA | NA |
| | xenograph | NA | NA |
| $z_9$ | percent | percentage | 0.652692556 |
| | ayling | NA | NA |
| $z_{10}$ | NA | NA | NA |

**TABLE 5.** Performance among methods for dataset RCV1.

| Methods | Top-20 | B/P | MAP | F1 |
|---|---|---|---|---|
| **SbTE** | **0.520** | **0.459** | **0.478** | **0.461** |
| sTRbTCM | 0.503 | 0.438 | 0.463 | 0.453 |
| TRbTCM | 0.516 | 0.442 | 0.469 | 0.454 |
| CRSWN | 0.469 | 0.427 | 0.439 | 0.438 |
| CSM | 0.445 | 0.395 | 0.408 | 0.416 |
| $\%Change$ | 0.78% | 3.84% | 1.92% | 1.54% |
| TNG | 0.484 | 0.381 | 0.400 | 0.404 |
| $\%Change$ | 7.43% | 20.47% | 19.5% | 14.11% |
| LDA_words | 0.466 | 0.424 | 0.439 | 0.438 |
| PLSA_words | 0.433 | 0.370 | 0.390 | 0.401 |
| BM25 | 0.345 | 0.337 | 0.330 | 0.359 |
| $\%Change$ | 11.59% | 8.25% | 8.88% | 5.02% |

### 2) DOCUMENT RANKING BASED ON TOPIC QUALITY

In this part, we would like to investigate the contributions of topic quality to information filtering systems. The proposed model was compared to some existing models in terms of using topic evaluation including TRbTCM, sTRbTCM, CRSWN and CSM. The proposed model is also compared to phrase-based topic model TNG, term-based topic models LDA_words and PLSA_words, and term-based representation BM25. TABLE 5 presents the performance results for dataset RCV1. The *%change* shows the percentage of change between the new proposed model and the best result of the other models in the same group. The higher the value of *%change*, the better the improvement of the proposed model is.

*Comparison with existing topic evaluation models*. As shown in TABLE 5, the performance in Top-20 score between the new model SbTE and other models in terms of topic evaluation was improved. In particular, it was 0.520 for the new model while it was 0.516 for the TRbTCM model which was better than the other topic evaluation models, sTRbTCM, CSM and CRSWN. This improvement in comparison to the second best method TRbTCM was nearly 0.78%. Similarly, SbTE outperformed model TRbTCM in MAP score with 0.478 and 0.469 for the two models respectively. The improvement in MAP score was 1.92%.

*Comparison with phrase-based representations*. As shown in TABLE 5, the new model SbTE performed better than phrase based models in all four criteria. In particular, the new model was higher than TNG in Top-20 score, which was 0.520 in the new model and 0.484 in TNG accordingly. In terms of MAP score, SbTE model gained 0.478 which is higher than 0.400 in TNG. This made the improvement to 19.5%.

*Comparison with term-based representations*. LDA_words is a baseline model which used topical words to represent user's interest. BM25 uses term frequency and invert term frequency to represent user interest. As shown in TABLE 5, the largest improvement between SbTE and term-based representation is in Top-20 score. Obviously, it was 0.520 in SbTE while it was 0.466 in LDA_words. This made the improvement to 11.59%. In MAP score, the new model also was higher in LDA_words, which was 0.477 and 0.439 respectively.

### 3) DISCUSSION

As mentioned in the introduction part, the aim of this study is to improve the accuracy of user interest representation based on topic models by evaluating the quality of topics, especially by interpreting the ambiguous topical words. As mentioned in the result section, approximately 10% of topical words are unmatched topical words. Therefore, it can be beneficial when some of the unmatched topical words can be explained by using concept words which are similar to the unmatched topical words. As can be seen in TABLE 5, the model SbTE outperforms the baseline models.

*Feature Significance:* Both TRbTCM and SbTE use highly frequent topical words (called features) for representing user's interest. However, some features might not be meaningful. In TRbTCM, all topical features are used to represent user's interest whatever the features are meaningful or meaningless. This leads to the problem of the existence of meaningless features in user's interest representations. On the other hand, SbTE only chooses significantly meaningful features which occur in concepts in the mapping ontology or have similar concept words in the ontology. Hence, SbTE can provide more meaningful terms in user interest representations where meaningless topical words are not included. User interest representations with meaningless topical words usually make noise in filtering relevant documents. Hence, the performance of TRbTCM is lower than that of SbTE as can be seen in TABLE 5.

It is worth to mention that, sTRbTCM has also attempted to interpret unmatched topical words using the synsets of WordNet which contain synonyms of a given word. The Jaccard similarity between two words was calculated in terms of the two words' positions in the WordNet ontology and used to find similar words for an unmatched topical word. The Jaccard similarity reveals less semantic information in comparison with the semantic similarity based on word embeddings which is used in SbTE. This may expain that the performance of sTRbTCM is worse than that of SbTE, even worse than the performance of TRbTCM.

*Topic Quality:* The meaningless topical words in topic models usually make noise to information filtering systems when used to represent user interest because it might lead to retrieve irrelevant documents. Hence, identifying the meaningless topical words is believed to be useful in enhancing the performance of IFs. Because SbTE deals with the ambiguity problem more effectively, it can help to discover more meaningful topical words as well as meaningful matched patterns to represent user's interest than models TRbTCM and sTRbTCM. As a result, SbTE outperforms the models TRbTCM and sTRbTCM in enhancing the performance of IFs.

When compared to the existing topic evaluation method based on correlation among topical words in CRSWN, the model SbTE outperforms that baseline model. The reason is that SbTE represents user's interest using the words in semantic patterns generated based on the matching between the examined topic and concepts in LCSH ontology. CRSWN utilizes specificity and coverage of overlap parts between concepts in WordNet and topics to evaluate the topics. This may indicate that, topic evaluation based on concepts in LCSH provides higher performance than WordNet lexical database.

## VI. CONCLUSION

In conclusion, topics in a topic model generated from a collection of documents provide a statistical representation for the document collection. However, the quality of topics is not always good because of meaningless and ambiguous topical words. Hence, there is a need to identify and remove the meaningless words from the user's interest representation. It is possible that ambiguous topical words can be explained by using similar words from concepts in ontologies. These tasks can be performed through a topic evaluation process. This study proposed a new model named as SbTE to assess the quality of topics based on an external knowledge base and use the assessed topics to filter out irrelevant documents from an incoming document stream. In particular, for solving the ambiguity problem, the model SbTE has used word embeddings to represent concept words and topical words which are then used to measure similarities between the words. In addition, we proposed a method to determine topic significance in an incoming document and a new method to rank relevance of the incoming documents over the training collection. Finally, the experiments were conducted on benchmark datasets RCV1 to compare performances between the new model and the baseline models. Four comparison metrics were used to assess the performances of the models. As can be seen in the experimental results, the new proposed model outperformed not only the baseline models in topic evaluation but also the state-of-the-art models such as BM25, pLSA, and LDA. In summary, the model SbTE has proven to be more effective than baseline models in enhancing IF performance.

## REFERENCES


[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," in *Proc. J. Mach. Learn. Res.*, pp. 993–1022, Jan. 2003.

[2] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, p. 77, Apr. 2012.

[3] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 448–456.

[4] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 288–296.

[5] C. C. Musat, J. Velcin, S. Trausan-Matu, and M.-A. Rizoiu, "Improving topic evaluation using conceptual knowledge," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1866–1871.

[6] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 100–108.

[7] H. Nguyen, Y. Xu, and Y. Li, "An ontology-based topic evaluation method for enhancing information filtering," in *Proc. IJCAI*, 2018, pp. 1–8. [Online]. Available: https://eprints.qut.edu.au/128771/

[8] H. Nguyen, Y. Xu, and Y. Li, "A semantic similarity based topic evaluation for enhancing information filtering," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Dec. 2018, pp. 150–157.

[9] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, no. Feb, pp. 1137–1155, 2003.
[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[11] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in *Proc. 13th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2004, pp. 42–49.
[12] X. Wang, A. McCallum, and X. Wei, "Topical N-grams: Phrase and topic discovery, with an application to information retrieval," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 697–702.
[13] U. Hanani, B. Shapira, and P. Shoval, "Information filtering: Overview of issues, research and systems," *User Model. User-Adapt. Interact.*, vol. 11, pp. 203–259, Aug. 2001.
[14] W. B. Cavnar and J. M. Trenkle, "N gram based text categorization," *Ann Arbor MI*, vol. 48113, no. 2, pp. 161–175, 1994.
[15] T. Hofmann, "Probabilistic latent semantic indexing," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 211–218, Aug. 2017.
[16] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 262–272.
[17] S. Fernando and M. Stevenson, "A semantic similarity approach to paraphrase detection," in *Proc. 11th Annu. Res. Colloq. UK Special Interest Group Comput. Linguistics*, 2008, pp. 45–52.
[18] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proc. 21st Nat. Conf. Artif. Intell. (AAAI)*, vol. 6. Palo Alto, CA, USA: AAAI Press, 2006, pp. 775–780.
[19] P. D. Turney, "Mining the Web for synonyms: PMI-ir versus ISA on TOEFL," in *Proc. Eur. Conf. Mach. Learn.* Springer, 2001, pp. 491–502.
[20] C. Leacock and M. Chodorow, *Combining Local Context and Wordnet Sense Similarity for Word Sense Identification. Wordnet, an Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.
[21] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," 1995, *arXiv:cmp-lg/9511007*. [Online]. Available: https://arxiv.org/abs/cmp-lg/9511007
[22] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. 10th Res. Comput. Linguistics Int. Conf.* Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 1997, pp. 19–33.
[23] S. Patwardhan, S. Banerjee, and T. Pedersen, "Using measures of semantic relatedness for word sense disambiguation," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Springer, 2003, pp. 241–257.
[24] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Lang. Cognit. Processes*, vol. 6, no. 1, pp. 1–28, Jan. 1991.
[25] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[26] LCSH. (2017). *Library of congress*. [Online]. Available: https://www.loc.gov/
[27] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.*, vol. 5, no. Apr, pp. 361–397, 2004.
[28] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, Mar. 2002.
[29] A. K. McCallum. (2002). *Mallet: A machine learning for language toolkit*. [Online]. Available: http://mallet.cs.umass.edu

**YUE XU** is currently an Associate Professor with the School of Computer Science, Queensland University of Technology, Australia. She is also an active Researcher in the areas of data mining, machine learning, and Web intelligence, and has made important contributions to related research areas. She has published over 150 refereed articles over the past ten years, some of which have been published in top journals, such as *Information Science*, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the *European Journal of Information Systems, ACM Transactions on Intelligent Systems and Technology*, and *Knowledge-Based Systems*, and major conferences, such as ICDM, CIKM, WWW, and ICIS. Her current research interests include text mining, recommender systems, Web intelligence, and user behavior modeling.

**HANH NGUYEN** is currently pursuing the Ph.D. degree with the School of Computer Science, Queensland University of Technology. She has published some articles in conferences such as IJCAI workshop Data Science meets Optimization, Web Intelligence, and Australian-AI. Her research interests focus on data mining, web intelligence, and artificial intelligence. In particular, her current research focuses on topic evaluation and pattern discovery.

**YUEFENG LI** is currently a Professor with the School of Computer Science, Queensland University of Technology, Australia. He has published more than 170 refereed journals and conference papers. He has demonstrable experience in leading large-scale research projects and has achieved many established research outcomes that have been published and highly cited in top data mining journals and conferences, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, CIKM, and ICDM. His research interests include text and data mining, ontology learning, and Web intelligence. He is currently an Editor-in-Chief of *Web Intelligence*, an International journal.