A Semantic Similarity based Topic Evaluation for Enhancing Information Filtering

Hanh Nguyen
Science and Engineering Faculty
Queensland University of Technology
Brisbane, Australia
trandiemhanh.nguyen@hdr.qut.edu.au

Yue Xu
Science and Engineering Faculty
Queensland University of Technology
Brisbane, Australia
yue.xu@qut.edu.au

Yuefeng Li
Science and Engineering Faculty
Queensland University of Technology
Brisbane, Australia
yuefeng.li@qut.edu.au

***Abstract* - Topic Modelling has been applied in many successful applications in data mining, text mining, machine learning and information filtering. The limitation is that the quality of topics generated from modelled corpus are not always good because many topics contain intrusive and ambiguous words. This negative drawback would affect the performance of text based application systems based on topic models. Hence, topic evaluation to assess and to rank the topics is really important for the good quality topics before applying those topics to text based applications. In this study, we proposed an ontology-based topic evaluation method for enhancing information filtering, named as STRbTCM. This new model assesses the quality of topics by matching topic models with headings in Library Congress Subject Heading (LCSH) ontology. To evaluate the effectiveness of our proposed model, we compare the model with two existing topic evaluation methods applied to information filtering system. In addition, we also compare our proposed model to term-based model BM25 and two other models based on topics: TNG and LDA_words. Through extensive experiments, we find that our proposed model performed better than other baseline models according to four main evaluating measures.**

Keywords: Information filtering, topic modelling, information retrieval, semantic data, data mining, knowledge base, ontology.

I. INTRODUCTION

The past decade has seen the rapid development of topic modelling in understanding text corpus. Among the state-of-the-art algorithms on understanding text documents, Latent Dirichlet Allocation [1, 2] is the most popular topic modelling technique, which provides an explicit representation of documents. In LDA, words with high frequency in the modelled documents are likely to be chosen to represent topics. LDA can discover main themes of unstructured documents such as free text documents without any markups [3]. LDA was also successfully used to discover main themes for unstructured documents such as books, news, abstracts, and metadata as reported in [4].

Despite LDA's considerable achievements in text applications, the topics generated by LDA still have limitations. Word intrusion and topic intrusion were reported in [5] as common limitations of topic models. Many of the topics are ambiguous and noisy due to ambiguous topic words [4]. These problems originate from the LDA algorithm, which generates inferred topics based on high frequent words occurring throughout the collection. Likewise, the ambiguous topics may contain other subtopics which cannot be accurately represented by1 their topical words [6]. Therefore, it becomes important to evaluate the quality of the topics in order to select good topics so that the qualified topics can help to improve performances of text based applications.

In the past few years, a considerable amount of literature has been published on topic evaluation. These studies follow two main directions: automatic evaluation and human judgements. A preliminary work on human judgements was conducted by Chang in [5]. That study discovered intrusive topic words based on human judgements which indicates the limitation of topic models. Unlike Chang, Musat in [7] reported a different method for assessing topics, which is automatic topic evaluation based on predefined knowledge of WordNet. This method evaluates topics by calculating semantic relevance scores among topic words which can be found in WordNet. One other method for automatically evaluating topics is based on statistics to measure the semantic co-occurrence between topic words as in [8]. This research was not based on the knowledge from external resources, but the co-occurrences of topic words in the modelled documents. To our knowledge, although those researches have been carried out on assessing the quality of topics, no single research has been reported that their methods can work effectively with large datasets. Even though those prior researches could indicate intrusive topic words in the examined topics but failed to assess the effectiveness of their methods in any real application contexts.

The major objective of this study was to investigate how to assess the quality of topics by utilizing external knowledge bases such as LCSH ontology. Firstly, we develop a method to evaluate the semantic expressive capability of a topic model by matching the topics in the topic model with concepts in the LCSH ontology, based on which to choose meaningful topics in order to enhance the quality of the topic model. Secondly, we apply the proposed topic evaluation method to information filtering systems to improve information filtering accuracy based on the enhanced topics. Through extensive experiments, results of the proposed model are compared with those previous studies including the method based on concepts in WordNet as in [7] and statistical method, called co-occurrence score in [8] . Moreover, we also find that our proposed method performs better than term based representations, phrase based representations, and LDA[2, 3, 9]. The contributions of this research is to provide a reliable model to automatically evaluate the quality of topics in a topic model.

The paper has been divided into five parts. The first section of this paper is the introduction part. Section 2 presents related works in the area of topic evaluation and information filtering. The third section is concerned with the proposed model in matching topics with LCSH ontology and solves the disambiguated problems of unmatched topic words. Section 4 analyses the comparisons between our proposed results and the baseline results. The result and discussion part is presented in section 5.

II. RELATED WORKS

The task of topic evaluation generally follows two main streams of manual and automatic evaluation. Topic Log Odds was introduced in [5] to measure the intrusion topics by human judgements. However, human being based manual

evaluation is time consuming and needs much effort for generating and assessing results.

For the topic evaluation based on topic models and the modelled documents, semantic co-occurrence between words in topic models is investigated by Mimno in [8] which introduced a correlation score between pairs of topic words basing on the occurrence of the pair in the collection. This method ignored the meaning of the co-occurring topic words. Similarly, the work in [10] also studied the coherence between words in topics. This work compares co-occurrence scores of topic word pairs over three different external resources and uses external resources and word co-occurrence including Wikipedia, WordNet, and Google. This research compared different methods in evaluating topics based on both external resources and occurrences of topic pair words.

On the other hand, ontologies provide knowledge sources for evaluating the quality of topics. Measuring semantic relevance of a topic to concepts in ontology was studied in [7]. The main idea of the method called CRSWM is to map topic words with senses in WordNet. This research employed the distance between topic words and concepts in WordNet to measure the relevance of the words to the concepts. Although this approach can measure the concept relevance of the examined topics based on the distance to the concepts in the ontology, the main weakness of this study is the failure to address the co-occurrence between topic words inside each concept.

Together, these studies have provided insights into topic evaluation. Human judgements on topic models seem to be more biased, time consuming and human efforts while automatic evaluations on topic models proved as a promising method in assessing topics. However, there is still limitation in the most current researches in topic assessment as was explained above. In the next section, an innovative approach in topic evaluation, based on a large controlled vocabulary LCSH, will be introduced. The method to evaluate our model in topic evaluation will be discussed in section 4.

III. THE PROPOSED MODEL

In this paper, we proposed an automatic topic evaluation model based on external knowledge resources LCSH. The models' name is Semantic Topic Ranking based on Topic-Concept Matching, shorted as STRbTCM. The main idea of the model is to match topics in a topic model with concepts in LCSH to measure the interpretations of the examined topics. The concepts in LCSH are meaningful phrases because they are well-written by librarians. For examples, some meaningful concepts in LCSH are "Acorn Electron Microcomputer", "Agent-based model Computer software". Therefore, we believe that the matching between topic words and the meaningful concepts in LCSH can interpret the semantic meaning of the examined topics. The basic idea in our model is to match topic words with concepts in LCSH. However, not all topic words can be matched with concepts in LCSH ontology as some topic words are ambiguous or using different words in LCSH concepts. For the topic words that do not match with any concepts in LCSH, we replaced those unmatched words with semantically similar words and match these similar words to the concepts in LCSH. Specifically, this paper solved two main problems of topic evaluation. Firstly, we defined the term Matching Degree to estimate the level of matching between topical words and concepts in LCSH in order to measure the meaningfulness of the topic words. Secondly, we proposed a model to solve the ambiguity problems for unmatched topic words. In general, if a topic has more words matched with a concept in LCSH, that topic is more meaningful than the topics which do not contain matched topical words. For evaluating our proposed topic evaluation method, we propose a document relevance estimation method based on topic models for information filtering.

A. Meaningfulness of topics

Topic modelling is a group of algorithms to discover hidden topics in the collection of documents. LDA is one of the generic statistical technique for generating topics. Let $D = \{d_1, d_2, \dots, d_M\}$ 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. Probability of the *ith* word written as w_i in the document d, denoted as $P(w_i|d) = \sum_{j=1}^{\nu} P(w_i|\mathcal{Z}_j) \times P(\mathcal{Z}_j|d)$, ν is the number of topics. At document level, each document is represented by topic distribution $\theta_d = \{\mathcal{V}_{d,1}, \mathcal{V}_{d,2}, \dots, \mathcal{V}_{d,\nu}\}, \mathcal{V}_{d,j} = P(Z_j|d)$, $\sum_{j=1}^{\nu} \mathcal{V}_{d,j} = 1$. In a collection level, D is represented by a set of topics. Each topic is represented by a probability distribution over words. For the *jth* topic, we have $\Phi_j = \{\phi_{j1}, \phi_{j2}, \dots, \phi_{jm}\}$, m is the number of words per topic, $\phi_{ji} = P(w_i|\mathcal{Z}_j)$. In terms of words, each topic $\mathcal{Z}_j$ is represented as a set of words, denoted as $termSet(\mathcal{Z}_j) = \{w_{j1}, w_{j2}, \dots w_{jm}\}$. The assignments of words to topic mainly base on the probability distribution in which words with high probabilities are sampled as topic words. As a result of this, some of the topic words are not representative because those words may be noisy, intrusive or meaningless even they occur frequently. This study aims to identify the interpretation of topics by matching topic words with concepts on ontology. The meaningfulness of a topic, measured as a matching degree, is estimated based on how much the topic words are matched with the ontology concepts. Specifically, matching degree of a topic is calculated by aggregating the topic words matched with concepts in the ontology. Topics with higher matching degree values are better interpreted by the ontology than those with lower matching degrees.

Definition 1 (Ontologies): Ontologies can be understood as the concepts of entities that represent human knowledge about things. Ontology can be presented in a tuple $Ont = \langle \boldsymbol{C}, \boldsymbol{R} \rangle$ such that C is a set of concepts; R is a set of relations. Regarding to LCSH ontology, $\boldsymbol{C}$ consists of subject headings; $\boldsymbol{R}$ comprises of relations between subject headings such as hierarchical, equivalent and association relationships.

Definition 2 (Matched Concepts): Matched Concepts of a topic $\mathcal{Z}_j$, denoted as $\Gamma(\mathcal{Z}_j)$, are a set of concepts C_i in the ontology Ont, which share a number of matched words with the topic $\mathcal{Z}_j = \{w_{j1}, w_{j2}, \dots w_{jm}\}$. Matched concepts is defined below, where $\boldsymbol{C}$ is the set of concepts in the ontology:

$$\Gamma(\mathcal{Z}_j) = \{C_i \mid C_i \in \boldsymbol{C}, C_i \cap termSet(\mathcal{Z}_j) \neq \phi\} \quad (1)$$

From the definition, we can see that a matched concept C_i belong to $\Gamma(\mathcal{Z}_j)$ has at least one word that overlaps with the topic. For example, given a topic $\mathcal{Z}_3$={"computer", "system", "data", "dutroux"}, and two concepts: C1= "Computer hardware" and C2= "Computer equipment". Because both C1 and C2 contain the word "Computer" overlapping with topic $\mathcal{Z}_3$, the matched concept for the topic is $\Gamma(\mathcal{Z}_3) = \{C1, C2\}$.

Definition 3 (Matched patterns). Matched patterns of a topic $\mathcal{Z}_j$ over matched concepts, denoted as $\mathcal{MP}(\mathcal{Z}_j)$, is defined as:

$$\mathcal{MP}(\mathcal{Z}_j) = \{P \mid \forall C_i \in \Gamma(\mathcal{Z}_j), P = C_i \cap termSet(\mathcal{Z}_j), P \neq \emptyset\} \quad (2)$$

A matched pattern of a topic over a concept is the overlapping part between that topic and the concept. Matched patterns indicate the most closely matched concepts and can identify how much the topic can be explained by the ontology. For example, given a topic $\mathcal{Z}_3$={"computer", "system", "data", "dutroux" } and two concepts: C1= "Computer hardware" and C2= "Computer system security", the matched pattern between C1 and $\mathcal{Z}_3$ is [computer]; the matched pattern between C2 and $\mathcal{Z}_3$ is [Computer, system]. Based on these matched patterns, we can identify the most closely matched concepts and these concepts indicate how much the topic can be explained by the ontology.

Definition 4 (Maximum Matched patterns): Maximum Matched patterns of a topic $\mathcal{Z}_j$, denoted as $Max\mathcal{MP}(\mathcal{Z}_j)$, is defined as:

$$Max\mathcal{MP}(\mathcal{Z}_j) = \{P \mid P \in \mathcal{MP}(\mathcal{Z}_j), \nexists P' \in \mathcal{MP}(\mathcal{Z}_j), P \subset P'\} \quad (3)$$

Each pattern in $Max\mathcal{MP}$ is maximum, i.e., it does not have super patterns in $Max\mathcal{MP}$. This means that each pattern in $Max\mathcal{MP}$ is a longest pattern. Obviously, the longest matched pattern for a topic has the highest number of words in the topic that are matched with a concept. In concept perspective, a concept that covers the longest pattern is closer to the topic than the concepts that cover shorter patterns.

For example, given three matched patterns $\mathcal{MP} = \{[\text{Computer}], [\text{System}], [\text{Computer}, \text{system}]\}$, the pattern [Computer, system] is the longest pattern, covering both patterns [computer] and [system]. In topic interpretation, concepts cover the pattern [Computer, system] are more specific than concepts covering shorter patterns like [Computer] and [System]; for instance, a concept "Computer System Security" is more specific than concept "Computer".

Definition 5 (Closest Matched Concepts): Closest Matched Concepts of a topic $\mathcal{Z}_j$, denoted as $\mathbb{CMC}(\mathcal{Z}_j)$, include the shortest concepts in ontology that closely cover the maximum matched patterns in Max$\mathcal{MP}(\mathcal{Z}_j)$. $\mathbb{CMC}(\mathcal{Z}_j)$ is a subset of matched concepts $\Gamma(\mathcal{Z}_j)$ and satisfy the following conditions:

$$\mathbb{CMC}(\mathcal{Z}_j) = \{C \mid P \in \text{Max}\mathcal{MP}(\mathcal{Z}_j), \exists C \in \Gamma(\mathcal{Z}_j), \nexists C' \in \Gamma(\mathcal{Z}_j), P \subset C, P \subset C', C' \subset C\} \quad (4)$$

As in this definition, the Closet Matched Concepts must satisfy the following conditions.

A concept in $\mathbb{CMC}$ covers one of the longest matched patterns. In other words, each concept in $\mathbb{CMC}$ must cover one of the patterns in Max$\mathcal{MP}$. For example, given a concept C1= "Computer systems--Security measures" and a concept C2= "Computer system security"; the longest pattern is [Computer, system]. In this step, C1 and C2 could be in $\mathbb{CMC}$ because both of them cover the longest matched pattern.

Each concept in $\mathbb{CMC}$ must be a shortest concept, which means there does not exist any other concept which is shorter than this concept and covers the same maximum matched pattern. In the last example, the concept C2 is the smallest one; the concept C1 covers concept C2, so C1 is not in $\mathbb{CMC}$. Hence, the closest matched concept is $\mathbb{CMC}$ ={C2}.

B. Disambiguation to improve the meaningfulness of the topic

Given a topic $\mathcal{Z}$ with $termset(\mathcal{Z}) = \{w_1, w_2, \dots, w_m\}$ where m is the number of terms in the topic, we assume that not all m topical words can be found in the LCSH ontology because the LCSH ontology does not cover all the words in our natural language. Therefore, we can divide the topic $\mathcal{Z}$ into two separate parts: (1) a set of matched topic words, denoted as $\mathcal{MZ}(\mathcal{Z})$, which shares a certain overlapping part with at least one concept in the ontology as in definition 2; and (2) a set of unmatched topic words, denoted as $\mathcal{UZ}(\mathcal{Z})$, which does not overlap with any concepts in the ontology. Hence, the topic becomes the union of two sets $\mathcal{MZ}(\mathcal{Z})$ and $\mathcal{UZ}(\mathcal{Z})$, $\mathcal{Z} = \mathcal{MZ}(\mathcal{Z}) \cup \mathcal{UZ}(\mathcal{Z}), |\mathcal{MZ}(\mathcal{Z})| \leq m$ & $|\mathcal{UZ}(\mathcal{Z})| \leq m$. For example, in the given topic $\mathcal{Z}$ ={"information", "soliday", "safety", "aviation", "trust", "urged", "air", "gain", "share", "told"}, words such as "soliday", "urged", and "told" cannot be matched with any concepts. Therefore, $\mathcal{UZ}(\mathcal{Z})$={ "soliday", "urged", "told"} is a set of unmatched topic words. If we just use LCSH ontology to interpret the topic $\mathcal{Z}$, the topic words in $\mathcal{UZ}(\mathcal{Z})$ would be considered meaningless or ambiguous. However, if we find some similar words in LCSH ontology that share a certain similarity with the words in $\mathcal{UZ}(\mathcal{Z})$, we can interpret the words in $\mathcal{UZ}(\mathcal{Z})$ by using the similar concept words. For instance, the unmatched topic word "urged" has maximal similarity with some words like "itch" and "impulse". If we replace the topic word "urged" with "itch" or "impulse", we can find matched concepts which have overlapping with the similar words. Hence, in this case, we can interpret the unmatched topic word in $\mathcal{UZ}(\mathcal{Z})$. In this section, we will propose a model for disambiguating those unmatched topic words. Each word in WordNet Ontology[11] is associated with a set of similar words, called word senses or synonyms. For each unmatched word in a topic, the WordNet ontology is used to find a set of senses which exist in LCSH. Based on the similarity the unmatched topic word and its senses. We can find the most similar sense word for the unmatched word.

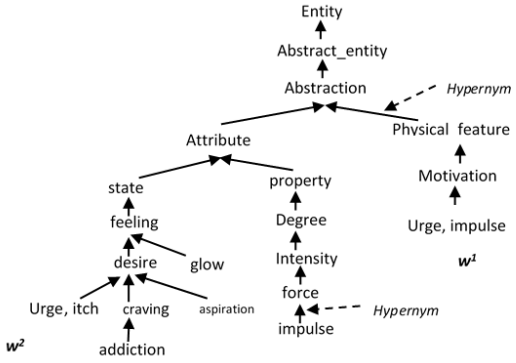

Figure 1. Fragment of WordNet ontology

In WordNet, $\mathcal{R}$ contains a set of hypernym or hyponymy relations and $\mathcal{C}$ is the set of senses represented in term of words. In general 'senses' are concepts, $\mathcal{R}$ contains a set of 'hypernym or hyponymy concepts'. There are four main types of senses in WordNet such as NOUN, VERB, ADJECTIVE, and ADVERB. For example, a sense represented by a word "urge" belongs to two more general senses which are about "Physical_feature" and about "Attribute" as illustrated in Fig 1 above. WordNet has a generic tree structure of nodes and edges. Nodes in WordNet are synsets containing senses which are synonym with each other. Take the node containing the sense "urge" as an example, that node also contains synonymy senses like "itch" or "impulse". Edges in WordNet represent hierarchical relationships between synsets of senses. The Fig

1 above illustrates the fragment of WordNet ontology regarding to the given concepts.

Path between concepts to root concept. Given two meaningful concepts c_i and c_{root} in a tree, $c_i \neq c_{root}$, there are at least one path in WordNet between the two concepts. Usually a concept c_i relates to multiple senses. There is a path from each sense of c_i to the root. Therefore, there are often multiple paths from c_i to the root. All paths from concept c_i to concept c_{root} are defined as $P_{i,root}(c_i, c_{root}) = \{P_i^1, P_i^2, \dots, P_i^v\}$ where $P_i^k = \langle c_i^k, c_{i+1,\dots}, c_{i+t}, \dots, c_{root}\rangle$ is the path from k^{th} sense of the concept c_i^k to the root, which is represented as a vector of concepts; where c_i^k present k^{th} sense of the concept c_i in the ontology.

Similarity between two senses in WordNet. For measuring the similarity between t^{th}sense of concept c_i denoted as c_i^t and r^{th} sense of concept c_j noted as c_j^r.There are some methods of measuring the similarity between two senses in WordNet as in [11] . In this study, we employed the Jaccard similarity coefficient [12] between two senses c_i^t and c_j^r where $c_i^t \neq c_{root}$ and $c_j^r \neq c_{root}$. Firstly, we find all the paths from concept c_i and concept c_j to concept root c_{root}, denoted as $P_{i,root}(c_i, c_{root}) = \{P_i^1, P_i^2, \dots, P_i^t\}$ and $P_{j,root}(c_j, c_{root}) = \{P_j^1, P_j^2, \dots, P_j^r\}$, with t and r be the number of paths from concept c_i and c_j to the concept root c_{root} respectively. The asymmetric binary similarity between path P_i^t and P_j^r is calculated based on this formula: $Jacc(P_i^t, P_j^r) = \frac{f(1,1)}{f(0,1)+f(1,0)+f(1,1)}$, where $f(1,1)$ represents is the number of concepts which appear in both paths P_i^t and P_j^r; $f(1,0)$ represents the number of concepts which appear in P_i^t and not appear in P_j^r; $f(0,1)$ represents the number of concepts which does not appear in P_i^t and appear in P_j^r.

For example, the asymmetric binary similarity between the concepts "addiction" and "urge", noted as $Jacc(\text{urge}, \text{addiction})$, is 0.7. Obviously, the hierarchical path from concept "addiction" to concept root is $P_{\text{addiction},root}$ = {"addiction", "carving", "desire", "feeling", "state", "attribute", "abstraction", "abstract_entity", "entity"} and the hierarchical path from concept "urge" to root concept is $P_{urge,root}$ ={"urge", "desire", "feeling", "state", "attribute", "abstraction", "abstract_entity", "entity"}. Fig 1 illustrates the paths in detail. $Jacc(\text{urge},\text{addiction}) = \frac{7}{1+2+7} = 0.7$.

Table I below presents similarities of an unmatched topic word "Urge" and some sample concept words in LCSH by following the Jaccard similarity coefficient. Readers can refer to Fig 1 about the paths of the concepts to the root "entity".

TABLE I. EXAMPLES OF CONCEPT SIMILARITIES

Concepts in WordNet	Similarity
Urge - motivation	0.833
Urge - Aspiration	0.778
Urge - impulse	1.000
Urge - itch	1.000
Urge - feeling	0.750
Urge - addiction	0.700
Urge - glow	0.667

Matching the unmatched topic words.

Let $\mathcal{CW}(Ont) = \{w_1^c, w_2^c, \dots, w_n^c\}$ are all the conceptual words extracted from all the LCSH concepts after removing special characters and stop words such as: the, of, about, in, etc. Let $\mathcal{TW}(\mathcal{D}) = \{t_1^{\mathcal{D}}, t_2^{\mathcal{D}}, \dots, t_m^{\mathcal{D}}\}$ be the set of all the topic words in the training collection $\mathcal{D}$. Similarly, let $\mathcal{UZ}(\mathcal{Z}_j)$ be a set of topic words in topic $\mathcal{Z}_j$ which cannot be matched with any concepts of LCSH. $\mathcal{UTW}(\mathcal{D}) = \bigcup_{j=1}^{v} \mathcal{UZ}(\mathcal{Z}_j)$ is represented for the set of all unmatched topic words in the training collection $\mathcal{D}$.

In this paper, we proposed a method to identify the meaning of these unmatched topic words by matching their similar words with LCSH concepts. The following mapping maps a topic word to its most similar word in LCSH, $\mathcal{RP}(\mathcal{Z}_j) = \{c_{j1}^w, c_{j2}^w, \dots, c_{jr}^w\}$ are concept words in $\mathcal{CW}(Ont)$, before calculating the matching degree, these unmatched topic words will have to be matched with concepts.

Let W_z be a set of senses of a word w_z, the sense t^{th} of w_z is denoted as w_z^t, i.e. $w_z^t \in W_z$. The path-based similarity between word senses w_z^t (w_z's t^{th} sense) and w_c^v (w_c's v^{th} sense) is calculated by following Jaccard similarity coefficient as $sim(w_z^t, w_c^v) = Jacc(P_z^t, P_c^v)$, where P_z^t, P_c^v are the paths from w_z^t and w_c^v to the root of the tree which contains both w_z^t and w_c^v, respectively. Path-based similarity between word w_z and word w_c, i.e, $sim(w_z, w_c)$ is defined as the maximum of path-based similarities of all senses in the two words w_z and w_c as following:

$$sim(w_z, w_c) = \max_{\substack{w_z^t \in W_z \\ w_c^v \in W_c}} (sim(w_z^t, w_c^v))$$

For a given topical word w, its most similar word in $\mathcal{CW}$ can be calculated as $arg\max_{w^c \in \mathcal{CW}} (sim(w, w^c))$ and its corresponding similarity is $\max_{w^c \in \mathcal{CW}} (sim(w, w^c))$. Let $\mathcal{SW}$ be a mapping from a topical word to a concept word in LCSH which is the most similar word to the topical word, $\mathcal{SW}$: $\mathcal{TW} \rightarrow \mathcal{CW}(Ont) \times [0,1]$; The most similar word with the similarity of topic word w can be defined as the following mapping:

$$\mathcal{SW}(w) = \begin{cases} \langle arg\max_{w^c \in \mathcal{CW}} (sim(w, w^c)), \max_{w^c \in \mathcal{CW}} (sim(w, w^c)) \rangle & w \in \mathcal{UTW} \\ \langle w, 1 \rangle & otherwise \end{cases}$$

$\mathcal{SW}(w)$ returns a pair of a similar word and a similarity value, i.e., $\mathcal{SW}(w) = < w_s, sim(w, w_s) >$. $\mathcal{SW}(w).ws$ and $\mathcal{SW}(w).sim$ denote the most similar word to w and the similarity value, respectively.

Definition 6 (Matching Degree): Matching Degree of a topic $\mathcal{Z}_j$ with concepts in an ontology, denoted as $\mathcal{MD}(\mathcal{Z}_j)$, $0 \leq \mathcal{MD}(\mathcal{Z}_j) \leq 1$; N is the number of similar words.

$$\mathcal{MD}(\mathcal{Z}_j) = \frac{1}{|\mathbb{CMC}(\mathcal{Z}_j)|}\left(\sum_{c \in \mathbb{CMC}(\mathcal{Z}_j)} \frac{|c \cap termSet(\mathcal{Z}_j)| * \underset{w \in c}{Avg}(\mathcal{SW}(w).sim)}{|c|}\right) \times \frac{\left|\bigcup_{c \in \mathbb{CMC}(\mathcal{Z}_j)} c \cap termSet(\mathcal{Z}_j)\right| + |N|}{|termSet(\mathcal{Z}_j)|} \quad (5)$$

This matching degree emphasizes the level of interpretation of topic $\mathcal{Z}_j$ with the human defined knowledge coded in ontology. If all topic words can be matched with concepts, i.e., those topic words can be interpreted by concepts in the ontology, the matching degree will be high. In contrast, meaningless topic words match with no concepts in the ontology, these words make no contribution to the matching degree.

Algorithm. *Calculate Matching Degree of topic $\mathcal{Z}_j$*

Input: Topic $\mathcal{Z}_j$, LCSH Ontology, WordNet Ontology
Output: Matching degree for topic ($\mathcal{Z}_j$)
1. Find matched concepts $\Gamma(\mathcal{MZ}(\mathcal{Z}_j))$ for topic $\mathcal{Z}_j$ from LCSH ontology by using Equation (2)
2. Find unmatched topic words $\mathcal{UZ}(\mathcal{Z}_j)$ from the topic $\mathcal{Z}_j$
3. Find matched concept for unmatched topic words $\Gamma(\mathcal{UZ}(\mathcal{Z}_j))$
4. Combine matched concepts from matched topic words and unmatched topic words into $\Gamma(\mathcal{Z}_j) = \Gamma(\mathcal{MZ}(\mathcal{Z}_j)) \cup \Gamma(\mathcal{UZ}(\mathcal{Z}_j))$
5. Calculate Maximum Matched Pattern $MaxMP(\mathcal{Z}_j)$ by using Equation (3)
6. Calculate Closest Match Concepts $\mathbb{CMC}(\mathcal{Z}_j)$ by using Equation (4)
7. Calculate Matching Degree for topic $\mathcal{Z}_j$ by using Equation (5)

C. Example

This example illustrates a small number of concepts in LCSH about things related to topic 9 in training folder 43 of Reuters dataset. The topic words is represented as $termset(\mathcal{Z}) =$ {"Information", "soliday", "safety", "aviation", "trust", "urged", "air", "gain", "share", "told"}. Some concepts in the Table II below are concepts that match with at least one word in the topic $\mathcal{Z}$. $\Gamma(\mathcal{Z}) = \{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}\}$,$|\Gamma(\mathcal{Z})| = 10$. Following is the matched concept $\Gamma(\mathcal{MZ}(\mathcal{Z}))$ which have at least one word matched with $termset(\mathcal{Z})$.

TABLE II. SOME LCSH CONCEPTS THAT MATCH WITH TOPIC $\mathcal{Z}$

ID	LCSH Concepts
C1	Industrial safety engineers
C2	Medical information science
C3	Common shares
C4	Information theory in biology
C5	Air safety
C6	Profit-sharing trusts
C7	Productivity gain sharing
C8	Reactor Safety Information System
C9	Safety equipment aviation structural mechanics
C10	Time-Shared, Interactive, Computer-Controlled, Information Television

According to Definition 2, the set of unmatched topic words is $U\mathcal{Z}(\mathcal{Z}) =$ {"soliday", "urged", "told"}. We then apply the disambiguation model to that $U\mathcal{Z}(\mathcal{Z})$. We begin by finding concept words that have the highest similarities with the words in $U\mathcal{Z}(\mathcal{Z})$ by applying Jaccard similarity measurement. Among the three words, we find the maximal similarity of "urged" as sim ("urged", "itch")= 1 and sim("urged", "impulse")=1. The words "soliday" and "told" do not have any similarities with the concept words. Because "itch" and "impulse" share the same similarity with the word urged and these words are in the concept words $\mathcal{CW}(Ont)$, we can randomly choose the first highest one which is "itch" to replace for the unmatched topic word "urged". After this step, the set of unmatched topic words becomes $U\mathcal{Z}(\mathcal{Z}) =$ {"soliday", "told"}.

The matched concept for the replaced concept words will be added: C11 = "Itches", i.e. $\Gamma(\mathcal{UZ}(\mathcal{Z})) = \{C_{11}\}$, Hence, the matched concept with disambiguation is $\Gamma(\mathcal{Z}) = \{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}\}$. According to Definition 4, the maximum matched patterns for the given topic is $MaxMP(\mathcal{Z}) =$ {[information, safety], [safety, air], [trust, sharing], [share, gain], [safety, aviation],[itch],[information, share]}. According to Definition 5, the closest matched concept for the given topic is $\mathbb{CMC}(\mathcal{Z}) = \{C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}\}$.

The matching degree of the topic $\mathcal{Z}$ is calculated by applying Definition 6 as following: Let's M be the equation like following: $M = \sum_{c \in \mathbb{CMC}(\mathcal{Z}_j)} \frac{|c \cap termSet(\mathcal{Z}_j)| * \underset{w \in c}{Avg}(\mathcal{SW}(w).sim)}{|c|}$

TABLE III. MATCHING DEGREE OF THE TOPIC OVER THE $\mathbb{CMC}(\mathcal{Z})$

Closest matched concepts $\mathbb{CMC}(\mathcal{Z})$	M
[structure, equipment, mechanic, safety, aviation]	2/5 =0.4
[safety, air]	2/2 =1.0
[trust, share, profit]	2/3 =0.667
[itch]	1*1.0 =1.0
[information, system, safety, reactor]	2/4 =0.5
[product, share, gain]	2/3 =0.667
[computer, information, interactive, share, control, time, television]	2/7 = 0.286

$\frac{\left|\bigcup_{c \in \mathbb{CMC}(\mathcal{Z}_j)} c \cap termSet(\mathcal{Z}_j)\right| + |N|}{|termSet(\mathcal{Z}_j)|} = \frac{7+1}{10} = 0.8$; $\mathcal{MD}(\mathcal{Z}) = \frac{4.52}{7} \times \frac{8}{10} \simeq 0.517$

D. Ranking Document Relevance

Information filtering (IF) aims to retrieve information that satisfy users' information needs. Two main parts in IF systems are user's needs representation and filtering task. The user's interest can be represented by terms, phrases or patterns [13, 14]. The user's interest can be represented by terms, phrases or patterns. Some most common methods in modelling user's interests are based on terms such as SVM, BM25, and TF-IDF. The filtering task is the core part in any IF systems. This task matches the user's interest with data extracted from the incoming stream of documents and then filters out the irrelevant documents. In this paper, we use topics generated from training documents by LDA to represent the user's information needs. The topic meaningfulness measurement on the generated topics gave matching degree of each modelled topic. For a new incoming document d, the basic idea is to determine the relevance of the document d to the users' interest by aggregating the significances of all topic words occurring in d. In this section, we proposed a ranking method to rank incoming documents by combining these parameters: topic distributions, matching degrees of the topic models with ontology LCSH and significances of topic models over the examined document. We use these ranking scores for filtering irrelevant documents from incoming document streams.

Topic distribution

Let $\mathcal{V}_{\mathcal{D},j}$ be the average distribution of all topics in the training collection $\mathcal{D}$, $\theta_{\mathcal{D}} = (\mathcal{V}_{\mathcal{D},1}, \mathcal{V}_{\mathcal{D},2}, \dots, \mathcal{V}_{\mathcal{D},v})$, $\sum_{j=1}^{v} \mathcal{V}_{\mathcal{D},j} = 1$ and $\mathcal{V}_{\mathcal{D},j}$ is calculated in this below equation:

$$\mathcal{V}_{\mathcal{D},j} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} P(\mathcal{Z}_j|d) \quad (6)$$

Matching degree of topic $\mathcal{Z}_j$

Matching degree of topic $\mathcal{Z}_j$ with the ontology LCSH, noted as $\mathcal{MD}(\mathcal{Z}_j)$, is to measure the accurateness and interpretation of topic words in the topic. The matching degree is calculated based on the formula (5) in section 3.2.

Significance of a topic over a queried document

Given a topic $\mathcal{Z}_j = \{w_1, w_2, \dots w_N\}$, we assume that the distribution of a topical word w_i in the topic $\mathcal{Z}_j$ somehow affects the final document ranking score of the corpus which infer the topic. If a topical word w_i in the topic $\mathcal{Z}_j$ *with its probability* $Pr(w_i|\mathcal{Z}_j)$ is larger than those of the remained topical words, that high probability topic word can be more important than other words with lower probabilities.

Let call N^+ be the number of topic words w_i where $Pr(w_i|\mathcal{Z}_j) > 0$. Average probability of topic $\mathcal{Z}_j$ is calculated by this formula: $avgPr(\mathcal{Z}_j) = \frac{1}{N^+}\sum_{i=1}^{N^+} Pr(w_i|\mathcal{Z}_j)$

For each topic word w_i whose probability is higher than average probability $avgPr(\mathcal{Z}_j)$ of the topic contributes more to the significance of topic $\mathcal{Z}_j$ than the other remained topic words in the topic $\mathcal{Z}_j$.

Let m_{ij} be a degree of contribution of w_i to the topic $\mathcal{Z}_j$ and m_{ij} is proportional to $Pr(w_i|\mathcal{Z}_j)$. The following formula is applied to topic word with $Pr(w_i|\mathcal{Z}_j) > avgPr(\mathcal{Z}_j)$; $m_{ij} = \frac{Pr(w_i|\mathcal{Z}_j)}{avgPr(\mathcal{Z}_j)}$; $m_{ij} > 1$. Significance of topic word w_i in topic $\mathcal{Z}_j$ is defined as:

$$sig(w_i|\mathcal{Z}_j) = m_{ij} * Pr(w_i|\mathcal{Z}_j) \quad (7)$$

In the filtering phase, let d be a document to be examined and the training corpus $\mathcal{D}$. We would like to determine whether document d is relevant to the topic $\mathcal{Z}_j$ trained by the training corpus $\mathcal{D}$. Explicitly, the significance of the topic $\mathcal{Z}_j$ trained on corpus $\mathcal{D}$ over the queried document d is estimated by aggregating the significance of selected topical word in the document d. The selected topic word is the topic word with its probability higher than the average probability over the examined topic.

$$\boldsymbol{sig}(\mathcal{Z}_j, d) = \sum_{w_i \in d, pr(w_i|\mathcal{Z}_j) > \text{avgPr}(\mathcal{Z}_j)} sig(w_i|\mathcal{Z}_j) \quad (8)$$

Document relevance ranking. For a new incoming document d, the relevance score of d over the training collection $\mathcal{D}$ with ν topic models is measured based on the formula (5), (6) and (8) as following:

$$\boldsymbol{rank}(d|\mathcal{D}) = \sum_{j=1}^{\nu} \boldsymbol{sig}(\mathcal{Z}_j, d) \times \mathcal{V}_{\mathcal{D},j} \times \mathcal{MD}(\mathcal{Z}_j) \quad (9)$$

IV. EXPERIMENTS

A. Dataset

The Library of Congress Subject Headings (LCSH) is originally available for computer processing as MARC. Currently, we can access LCSH through RDF format. In the following experiments, we use subject authority database [15] which contains 440105 topical subject headings. This is a raw RDF file, so we need to parse the RDF file to extract those topical subject headings.

WordNet is a lexical database of English in which nouns, verbs, adjectives and adverbs are organized in terms of synonyms (synsets). There are around 117,000 synsets in WordNet. The relationships among synsets in WordNet are hypernym or hyponym. In the following experiments, WordNet was used in the topic named CRSWN and solved the problem of ambiguity in my proposed model.

The Reuter Corpus Volume 1 (RCV1) dataset [16, 17] was collected by Reuter's journals from the year of 1996 to 1997, covering approximately 806,791 documents about various topics. 100 collections were developed for the TREC filtering track. In TREC track, a collection is referred to as a topic. The dataset is divided into training and testing sets. The training set consists of up to 83,650 documents. The testing set contains the remaining documents, around 723,141 documents. In this study, we use 50 first collections which were evaluated by human assessors.

B. Baseline models

The experiments were extensively conducted to evaluate the effectiveness of the proposed topic evaluation model in information filtering. We divided our experiments into two major categories: document representation and topic evaluation. For document representation, three state-of-the-art methods were implemented as base line models. BM25 represent documents as single terms. LDA_words use single topical words to represent documents. TNG represent documents in terms of phrases. Regarding to topic evaluation, two base line experiments were conducted, correlation score model (CSM) [8] and concept relevance score based on WordNet (CRSWN) in [7].

- Term based representation: BM25 [18] is one of the base line models for representing documents by using terms as user interest.
- Phrase based topic representation TNG: TNG [9, 19] is N-Gram based topic model.
- LDA_words representation**:** User interests are represented by topic words which appear frequently in the training documents [2, 3].
- CSM: Topic evaluation based on Co-occurrence Score Model in IF system:

CSM evaluated topics based on the correlation scores between two any topical words of the examined topic. For more explanation, readers can refer to the work in [8] . In short, the semantic correlation score between a topic word w and remained topical words in $\mathcal{Z}_j$ is denoted as $C(w; \mathcal{Z}_j)$ and calculated as following:

$$C(\text{w}; \mathcal{Z}_j) = \sum_{m=2}^{M}\sum_{l=1}^{m-1} log\frac{D(t_m^{(w)}, t_l^{(w)})+1}{D(t_l^{(w)})} \quad (10)$$

Where $D(t, t')$ is co-document frequency of words t and t'. $\mathcal{D}(t)$ is document frequency of word t. Average of co-occurrence scores over all topics in the modelled collection $\mathcal{D}$ is denoted as $\mathcal{T}_D$, where m is number of topics trained from the collection $\mathcal{D}$. $\mathcal{T}_D$ is defined as:

$$\mathcal{T}_D = \frac{1}{m}\sum_{j=1}^{m}\sum_{\text{w} \in \mathcal{Z}_j} C(w; \mathcal{Z}_j^w) \quad (11)$$

Originally, the study in [8] used $\mathcal{T}_D$ to compare the quality of the examined topics. Then, they had human experts to evaluate the assessments. In this base line, we evaluated the performance of topic evaluation differently by applying the assessed topics to information filtering system. Specifically, the significance of topic $\mathcal{Z}_j$ over the query document d is estimated by aggregating the frequency of the co-occurrence of topic words appearing in the document d:

$$\boldsymbol{sig}(\mathcal{Z}_j, d) = \sum_{w_i, w_t \in d, w_i, w_t \in termSet(\mathcal{Z}_j)} C(w_i, w_t) \quad (12)$$

From (11) and (12), relevance ranking model for document d is defined as following formula:

$$\boldsymbol{rank}(d,\mathcal{D}) = \sum_{j=1}^{v} \boldsymbol{sig}(\mathcal{Z}_j, d) \times \mathcal{T}_D) \quad (13)$$

- CRSWN: Topic evaluation based on Conceptual Relevance Score in WordNet ontology.

The work CRSWN calculated the relative relevance score of the examined topic based on the distances between matched concepts in the WordNet ontology. Readers can refer to the study in [7] for detail. In summary, the relevance score of a concept c over topical word t_i, denoted as $\phi(c, t_i)$ and measured as following:

$$\phi(c, t_i) = w_{cov} \,.\, cov(c, t_i) + w_{spec} \,.\, spec\,(c, t_i) \quad (14)$$

Where: $cov(c, t_i)$ is coverage of a concept c over the topical word $t_i \in \mathcal{Z}_j$, $cov(c, t_i) = \frac{|\delta(t_i) \cap \delta(t_i, c_c)|}{|\delta(t_i)|}$; and $spec\,(c, t_i)$ is the specificity of the concept c over topical word t_i ; $spec(c, t_i) = w_h\, h(c) + w_p\, depth(c, t_i)$ where $w_h = 0.5$ and $w_p = 0.5$ are weights set a priori ; $h(c)$ is the height of concept c and $depth(c, t_i)$ is the distance from concept c to topic word t_i. We also set $w_{cov} = 0.5$ and $w_{spec} = 0.5$.

Similar to the CSM model, CRSWN used WordNet for measuring the correlations between topic words by applying formula (14). Then, for evaluating their method, they used human judgments to evaluate the proposed method. In this base line experiment, we evaluated the performance of topic evaluation differently by using information filtering system instead of human judgements as in [7]. For an incoming document d. Significance of topic $\mathcal{Z}_j$ over the document d is calculated by:

$$\boldsymbol{sig}(\mathcal{Z}_j, d) = \sum_{w_i \in d, w_i \in termSet(\mathcal{Z}_j)} Termfreq\,(w_i) \quad (15)$$

The average relevance scores of all matched concepts with topical words in $\mathcal{Z}_j$ is calculated as in formula (16) where δ all matched concepts in WordNet leading to all topic words is $\mathcal{Z}_j$, $\phi(c_t; \mathcal{Z})$ is in equation (14).

$$avgRS(\mathcal{Z}_j) = \frac{1}{|\delta|} \sum_{c_t \in \delta} \sum_{t_k \in \mathcal{Z}_j} \phi(c_t; \mathrm{t}_k) \quad (16)$$

From (15) and (16), relevance ranking for an incoming document d over the training collection $\mathcal{D}$ is defined as:

$$\boldsymbol{rank}(d,\mathcal{D}) = \sum_{j=1}^{v} \boldsymbol{sig}(\mathcal{Z}_j, d) \times avgRS(\mathcal{Z}_j)$$

C. Experimental settings

The experiments were extensively conducted to evaluate the effectiveness of the proposed topic evaluation model when applied to information filtering systems. For topic models used in the experiments, we trained LDA topic models with 50 collections in the Reuters corpus in MALLET toolkit [20]. The initial parameter setting for LDA training is $\alpha = \frac{50}{v}$; $\beta = 0.01$; number of topics per one collection is $v = 10$; maximal number of representative topical words per topic is 10. In the experiment, the number of topic words depends on the occurrences of topic words each training collection. The representative topic words in our proposed model are words with their probabilities higher than average and the number of words per topic must less than 10. Specifically, the *ith* topic word in *jth* topic is selected if $pr(w_i|\mathcal{Z}_j) > \text{avgPr}(\mathcal{Z}_j)$, where $\text{avgPr}(\mathcal{Z}_j)$ is the average distribution of assigned topic words in topic $\mathcal{Z}_j$.

Statistically, there are about 1830 topic words in 500 topics from 50 first training collections in Reuters Dataset. Statistically, nearly 10% of the topic words are unmatched topic words when we examined the topic words with LCSH ontology. In the experiment, we also find that there are around 145000 concept words.

D. Results and discussion

Precision and Recall are the two most effectiveness measures of text retrieval applications. Recall measures how well the information filtering is finding the relevant documents to a specific query. Precision measures how well the information filtering system rejects the non-relevant documents. In these experiments, we used four main evaluation metrics to evaluate the models.

The scores of Top-20 indicates the relevance proportion out of top 20 retrieved documents. In other words, the Top-20 score evaluates the precision and recall for the first 20 retrieved documents.

Mean Average Precision (MAP) measures precision at each relevant document first, and averaging precision over all topics afterwards. MAP measurement 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 measure indicates the effectiveness of the system. The higher this value of b/p is, the better the implemented system.

F1 scores reflect the harmonic average of the precision and recall. F1 emphasizes the effectiveness of retrieved sets.

The experimental results are provided in Table IV below.

TABLE IV. COMPARISON AMONG METHODS

Methods	Top-20	b/p	MAP	F1
STRbTCM	**0.503**	**0.438**	**0.463**	**0.453**
CRSWN	0.469	0.426	0.439	0.438
CSM	0.445	0.395	0.408	0.416
improvement %	*+7.25%*	*+2.82%*	*+5.47%*	*+4.42%*
LDA_words	0.466	0.424	0.439	0.438
TNG	0.444	0.367	0.371	0.386
improvement %	*+7.94%*	*+3.30%*	*+5.47%*	*+3.42%*
BM25	0.434	0.339	0.401	0.410
improvement %	*+15.90%*	*+29.2%*	*+15.46%*	*+10.49%*

Comparison with topic evaluation models: As displayed in the Table IV, our method showed higher performances than two methods of topic evaluation: CRSWN, CSM. Specifically, the maximum change is more than 7.00% in Top-20 score, where it is 0.503 in STRbTCM and 0.469 in CRSWM. Similarly, the score of MAP is 0.463 in our model and 0.439 in CRSWN, which makes an improvement change nearly 5.50%.

Comparison with term-based models and topic-based models: The new proposed model provides higher performance than BM25 in all four evaluation metric. Specifically, Top-20 score is 0.503 in our proposed model while it is only 0.434 in BM25. The improvement change is nearly 16%. Our proposed model performed better than TNG and LDA_words methods in all four criteria. While STRbTCM gained 0.438 in b/p score, it was 0.424 in LDA_words. The improvement change in b/p score is more than 3.00%. Similarly, the difference in MAP score was 0.463 and 0.439 in STRbTCM and LDA_words respectively, which

make the improvement change reaching approximately 5.50%.

The new proposed model outperformed the document representations based on terms like BM25 and topic such as LDA_words and TNG. Regarding to topic evaluation, this new model STRbTCM provides a new automatic topic evaluation based on term-concept matching with the external resource named LCSH. The matching degree measurement can accurately estimate the quality of topics better than other methods based on WordNet and word-correlation score as in [8] and in [7] respectively. In addition, the automatic approaches of assessment obviously solve the efficiency problems of time – consuming, human efforts and biases in humans' assessment methods in [7, 8]. Hence, this proposed model of assessment can be applied to large scale applications where the number of topics is large.

V. CONCLUSION

In topic modelling, it has been said that the meaningless and ambiguous words in topic models negatively affect the interpretation of topics from the modelled collection. In the aspect of the applications, topics with meaningless or ambiguous words prevent the performance of text based applications from gaining higher performance results. A new topic evaluation method to automatically assess the quality of the topics is necessary. This paper has proposed a new method called STRbTCM to evaluate topics based on a large external knowledge resource LCSH. Particularly, the proposed Matching Degree scores can fully measure the interpretations of topics. The disambiguation task was proposed to increase the topic interpretation by finding similar words in the LCSH for unmatched topic words. We applied evaluated topics to information filtering systems over Reuters dataset. The results of this new model were compared to two methods of topic evaluation which are CSM and CRSWN, two models of topic-based representation and term- based representation. Through extensive number of experiments, we found that our model currently outperformed all the baseline experiments. In summary, this new model STRbTCM has proven an innovative method for topic evaluation.

ACKNOWLEDGMENT

This work is supported by the VIED/QUT Research Doctoral Scholarship.

REFERENCES

[1] Wei, X. and W.B. Croft, *LDA-based document models for ad-hoc retrieval*, in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006, ACM: Seattle, Washington, USA. p. 178-185.

[2] Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent Dirichlet Allocation.* Journal of Machine Learning Research 3, 2003: p. 993-1022.

[3] Blei, D.M., *Probabilistic topic models.* Communications of the ACM, 2012. **55**(4): p. 77.

[4] Newman, D., et al. *Evaluating Topic Models for Digital Libraries*. in *JCDL '10 Proceedings of the 10th annual joint conference on Digital libraries* 2010. New York, NY, USA ACM.

[5] Chang, J., et al. *Reading tea leaves: How humans interpret topic models*. in *Advances in neural information processing systems*. 2009.

[6] Yan, X., et al., *A biterm topic model for short texts*, in *Proceedings of the 22nd international conference on World Wide Web*. 2013, ACM: Rio de Janeiro, Brazil. p. 1445-1456.

[7] Musat, C., et al., *Improving Topic Evaluation Using Conceptual Knowledge*, in *22nd International Joint Conference on Artificial Intelligence (IJCAI)*. Jul 2011, hal-00616245: Barcelona, Spain. p. pp 1866-1871.

[8] Mimno, D., et al., *Optimizing semantic coherence in topic models*, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2011, Association for Computational Linguistics: Edinburgh, United Kingdom. p. 262-272.

[9] Wang, X., A. McCallum, and X. Wei. *Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval*. in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. 2007.

[10] Newman, D., et al. *Automatic evaluation of topic coherence*. in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2010. Association for Computational Linguistics.

[11] Ahsaee, M.G., M. Naghibzadeh, and S.E. Yasrebi. *Using WordNet to determine semantic similarity of words*. in *Telecommunications (IST), 2010 5th International Symposium on*. 2010. IEEE.

[12] Wikipedia. *Jaccard index*. 18-May-2018]; Available from: https://en.wikipedia.org/wiki/Jaccard_index.

[13] Hanani, U., B. Shapira, and P. Shoval, *Information Filtering: Overview of Issues, Research and Systems.* User Modeling and User-Adapted Interaction, 2001. **11**(3): p. 203-259.

[14] Belkin, N.J. and W.B. Croft, *Information filtering and information retrieval: two sides of the same coin?* Communications of the ACM, 1992. **35**(12): p. 29-38.

[15] Congress.gov. *Libray of Congress*. 2017 [cited 2016 September]; Available from: https://www.loc.gov/.

[16] Rose, T., M. Stevenson, and M. Whitehead. *The Reuters Corpus Volume 1-from Yesterday's News to Tomorrow's Language Resources*. in *LREC*. 2002. Las Palmas.

[17] Lewis, D.D., et al., *Rcv1: A new benchmark collection for text categorization research.* Journal of machine learning research, 2004. **5**(Apr): p. 361-397.

[18] Robertson, S., H. Zaragoza, and M. Taylor, *Simple BM25 extension to multiple weighted fields*, in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. 2004, ACM: Washington, D.C., USA. p. 42-49.

[19] Shirakawa, M., T. Hara, and S. Nishio, *IDF for Word N-grams.* ACM Trans. Inf. Syst., 2017. **36**(1): p. 1-38.

[20] McCallum, A. *MALLET: A machine learning for language toolkit*. 2002; Available from: http://mallat.cs.umass.edu.