AUTHOR(S): MBIPOM, B., CRAW, S. and MASSIE, S.

TITLE: Improving e-learning recommendation by using background knowledge.

YEAR: 2018

Publisher citation: MBIPOM, B., CRAW, S. and MASSIE, S. 2018. Improving e-learning recommendation by using background knowledge. Expert systems [online], Early View. Available from: https://doi.org/10.1111/exsy.12265

OpenAIR citation: MBIPOM, B., CRAW, S. and MASSIE, S. 2018. Improving e-learning recommendation by using background knowledge. Expert systems, Early View. Held on OpenAIR [online]. Available from: https://openair.rgu.ac.uk

Improving e-Learning Recommendation by using Background Knowledge

Blessing Mbipom Susan Craw
Stewart Massie
School of Computing Science & Digital Media
Robert Gordon University, Aberdeen, UK
b.e.mbipom/s.craw/s.massie@rgu.ac.uk

Abstract

There is currently a large amount of e-Learning resources available to learners on the Web. However, learners often have difficulty finding and retrieving relevant materials to support their learning goals because they lack the domain knowledge to craft effective queries that convey what they wish to learn. In addition, the unfamiliar vocabulary often used by domain experts makes it difficult to map a learner's query to a relevant learning material. We address these challenges by introducing an innovative method that automatically builds background knowledge for a learning domain. In creating our method, we exploit a structured collection of teaching materials as a guide for identifying the important domain concepts. We enrich the identified concepts with discovered text from an encyclopedia, thereby increasing the richness of our acquired knowledge. We employ the developed background knowledge for influencing the representation and retrieval of learning resources to improve e-Learning recommendation. The effectiveness of our method is evaluated using a collection of Machine Learning and Data Mining papers. Our method outperforms the benchmark, demonstrating the advantage of using background knowledge for improving the representation and recommendation of e-Learning materials.

1 Introduction

Learning-focused content is increasingly available on the Web, thus providing an excellent source of information for building e-Learning systems (Clarà and Barberà, 2013). However, learners often have difficulty finding the right learning materials because they lack the domain knowledge required to formulate effective queries (Chen et al., 2014). In addition, a mismatch in the vocabulary used by learners when crafting their queries and that used by domain experts to describe learning concepts poses a further challenge for systems recommending resources to learners.

Another challenge with e-Learning recommendation is that the learning resources are often unstructured text, and so are not properly indexed for retrieval (Nasraoui and Zhuhadar, 2010). The challenge of dealing with unstructured learning resources creates a difficulty in finding and retrieving relevant learning resources. Hence the need for an effective method of representing learning materials with the aim of improving recommendation.

This paper proposes the automated acquisition of background knowledge about a domain that can then be employed for enhancing e-Learning recommendation. In our method, we create a concept-aware representation that contains a good coverage of relevant topics from the domain. First, we exploit a structured collection of teaching materials as a guide for identifying the important concepts. Next, we enrich the identified concepts with discovered text from an encyclopedia source, thereby increasing the richness of our representation. Our developed method is demonstrated in Machine Learning and Data Mining, although the method we present can be applied to learning materials in other domains.

Other projects such as DeepQA (Ferrucci et al., 2013) and DBpedia (Lehmann et al., 2015) use a range of knowledge-rich representations to enhance retrieval. Such knowledge-rich sources are usually in the form of important topics that describe a domain. While these projects generally rely on handcrafted knowledge sources, they highlight the advantage in exploiting knowledge-rich representations as a basis for improving recommendation.

A good coverage of domain topics is useful for representing learning materials. These domain topics contain rich vocabulary and provide a good knowledge source for mapping learners' queries

to learning materials. Thus allowing us to address the mismatch in the vocabulary used by learners and domain experts. We address this issue by introducing a method that automatically creates custom background knowledge in the form of a rich set of domain topics. Further, we explore building a richer vocabulary to achieve a better coverage of the domain, and this method is employed to improve e-Learning recommendation.

We make several contributions in this work. Firstly, the creation of background knowledge for an e-Learning domain. We describe how we take advantage of the knowledge of experts contained in e-Books to build a knowledge-rich representation that is used to enhance recommendation. Secondly, we present a method that harnesses the developed background knowledge to augment the representation of learning resources in order to improve e-Learning recommendation. Finally, we explore a larger concept vocabulary which provides a better coverage of the domain. We refine our method presented in (Mbipom et al., 2016) to generate a richer and focused set of domain concepts. The results from our evaluation show the improvement in e-Learning recommendation when the richer concept vocabulary is used for representing learning resources.

The rest of this paper is organised as follows. In Section 2 we present related text representation approaches that underpin this work. Section 3 describes the development of our background knowledge using available knowledge sources. Section 4 discusses the representation of learning resources using our methods. Then Section 5 presents the evaluation of the learning resource representation. In Section 6 we present our refined method of generating background knowledge with an evaluation using the richer vocabulary and a larger dataset for recommendation. Finally, Section 7 presents our conclusions.

2 Related Work

E-Learning recommendation is challenging because learning resources are often unstructured text, and so are not properly indexed for retrieval. A possible solution to addressing this challenge is the creation of effective representations that capture the content of learning resources. However, building suitable representations for learning resources in e-Learning environments is not easy (Dietze et al., 2012), as the resources do not have a pre-defined set of features by which they can be indexed.

We propose the creation of a knowledge-rich representation that captures the domain-specific vocabulary contained in learning resources. Figure 1 illustrates two broad approaches often used to address the challenge of text representation. These are corpus-based methods, such as topic models (Blei and McAuliffe, 2007; Chen and Liu, 2014); and structured representations, such as those that take advantage of ontologies (Boyce and Pahl, 2007; Yarandi et al., 2011). In Figure 1, the lower row of items identifies various knowledge sources that can be employed to build a range of knowledge-light to knowledge-rich text representation approaches.

Corpus-based methods usually involve the use of statistical models to identify topics from a corpus. The identified topics are often keywords (Beliga et al., 2015; Matsuo and Ishizuka, 2004) or phrases (Coenen et al., 2007; Witten et al., 1999). Coenen et al. showed that using a combination of keywords and phrases was better than using only keywords (Coenen et al., 2007). These topics can be extracted from different text sources such as: learning resources (Rodrigues et al., 2007; Yang et al., 2016), metadata e.g. Tables of contents (Bousbahi and Chorfi, 2015), and Encyclopedia e.g. Wikipedia (Milne and Witten, 2008; Qureshi et al., 2014). A drawback of the corpus-based methods is that, they normally rely on the coverage of the document collection used, so the topics produced may not be representative of the learning domain.

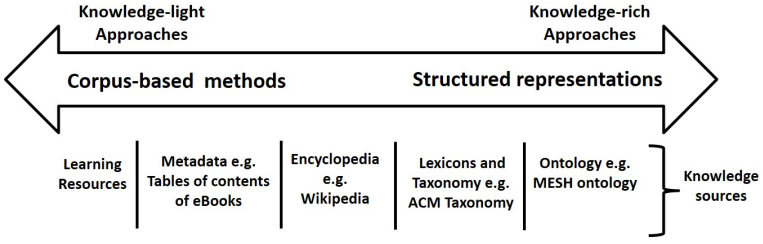


Figure 1: Two broad approaches used for text representation

Structured representations capture relationships between important domain concepts. This often entails using an existing ontology e.g ACM taxonomy (Nasraoui and Zhuhadar, 2010; Ruiz-Iniesta et al., 2014), or creating a new one (Gherasim et al., 2013; Panagiotis et al., 2016). Although ontologies are designed to have a good coverage of their domains, the output is still dependent on the view

of its builders and, because of handcrafting, existing ontologies cannot easily be adapted to new domains. e-Learning is dynamic because new resources are becoming available regularly, and so using fixed ontologies limits the potential to incorporate new content.

The approach adopted in this paper draws insight from both the corpus-based methods and structured representations highlighted in Figure 1. We leverage on a structured corpus of teaching materials such as Tables of contents of e-Books, in order to identify important topics in an e-Learning domain. These topics are a combination of keywords and phrases as recommended in (Coenen et al., 2007). The identified topics are then enriched with discovered text from Wikipedia in order to enhance our representation. In addition, we refine the methods developed in previous work (Mbipom et al., 2016) so that we can generate a richer set of relevant topics that provide a good coverage of the learning domain. Consequently, our approach is employed to influence the representation and retrieval of relevant learning resources.

3 Creation of Background Knowledge

Background knowledge refers to information about a domain that is useful for general understanding and problem-solving (Zhang et al., 2013). We attempt to capture background knowledge as a set of domain concepts, each representing an important topic in the domain. For example, in a learning domain, such as Machine Learning, you would find topics such as Classification, Clustering and Regression. Each of these topics would be represented by a concept, in the form of a concept label and a pseudo-document which describes the concept. The concepts can then be used to underpin the representation of e-Learning resources.

Our knowledge extraction process is shown in Figure 2. The input to this process are domain knowledge sources, and we use a structured collection of teaching materials and an encyclopedia source. Next, ngrams are automatically extracted from our structured collection to generate a set of potential concept labels. Then a domain lexicon is used to validate the extracted ngrams to ensure that the ngrams are also being used in another information source. The encyclopedia provides text descriptions for the identified ngrams. The output from this process is a set of domain concepts, each having a concept label and an associated pseudo-document. We discuss the stages of the background knowledge creation in the following sections.

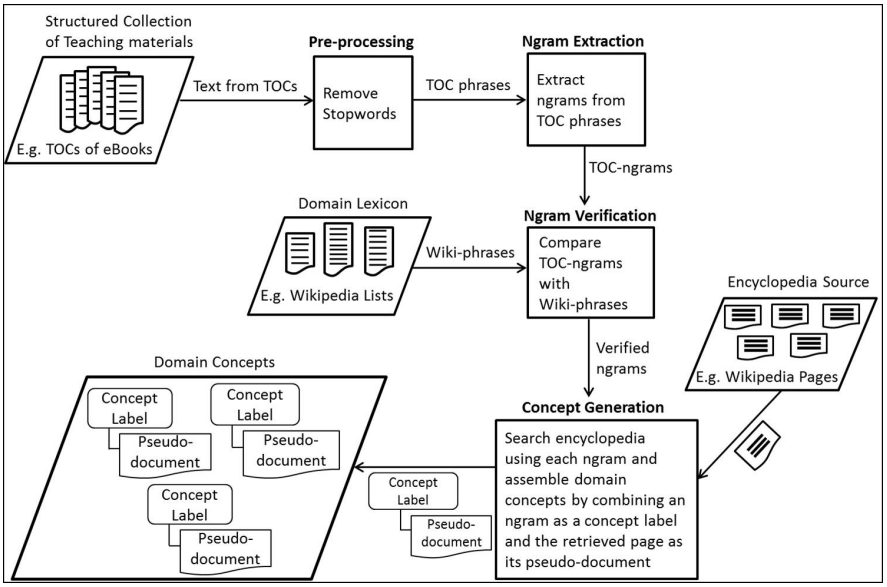


Figure 2: An overview of the background knowledge creation process

3.1 Knowledge Sources

Two knowledge sources are used as initial inputs for discovering concept labels. A structured collection of teaching materials provides a source for extracting important topics identified by teaching experts in the domain, while a domain lexicon provides a broader but more detailed coverage of the relevant topics in the domain. The lexicon is used to verify that the concept labels identified from the

teaching materials are directly relevant. Thereafter, an encyclopedia source, such as Wikipedia pages, is searched and provides the relevant text to form a pseudo-document for each verified concept label. The final output from this process is our set of domain concepts each comprising a concept label and an associated pseudo-document.

Our approach is demonstrated with learning resources from Machine Learning and Data Mining. We use e-Books as our collection of teaching materials; a summary of the books used is shown in Table 1. Two Google Scholar queries: "Introduction to data mining textbook" and "Introduction to machine learning textbook" guided the selection process, and 20 e-Books that met all of the following 3 criteria were chosen. Firstly, the book should be about the domain. Secondly, there should be Google Scholar citations for the book. Thirdly, the book should be accessible. We use the Tables-of-Contents (TOCs) of the books as our structured knowledge source.

We use Wikipedia to create our domain lexicon because it contains articles for many learning domains (Völkel et al., 2006; Zheng et al., 2010), and the contributions of many people (Yang and Lai, 2010), so this provides the coverage we need in our lexicon. The lexicon is generated from 2 Wikipedia sources. First, the phrases in the *contents* and *overview* sections of the chosen domain are extracted to form a topic list. Then, a list with the titles of articles related to the domain is added to the topic list to assemble our lexicon. Overall, our domain lexicon contains a set of 664 Wiki-phrases.

Table 1: Summary of e-Books used

Book Title & Author	**Cites**
Machine learning; Mitchell	264
Introduction to machine learning; Alpaydin	2621
Machine learning a probabilistic perspective; Murphy	1059
Introduction to machine learning; Kodratoff	159
Gaussian processes for machine learning; Rasmussen & Williams	5365
Introduction to machine learning; Smola & Vishwanathan	38
Machine learning, neural and statistical classification; Michie, Spiegelhalter, & Taylor	2899
Introduction to machine learning; Nilsson	155
A First Encounter with Machine Learning; Welling	7
Bayesian reasoning and machine learning; Barber	271
Foundations of machine learning; Mohri, Rostamizadeh, & Talwalkar	197
Data mining-practical machine learning tools and techniques; Witten & Frank	27098
Data mining concepts models and techniques; Gorunescu	244
Web data mining; Liu	1596
An introduction to data mining; Larose	1371
Data mining concepts and techniques; Han & Kamber	22856
Introduction to data mining; Tan, Steinbach, & Kumar	6887
Principles of data mining; Bramer	402
Introduction to data mining for the life sciences; Sullivan	15
Data mining concepts methods and applications; Yin, Kaku, Tang, & Zhu	23

3.2 Generating Potential Domain Concepts

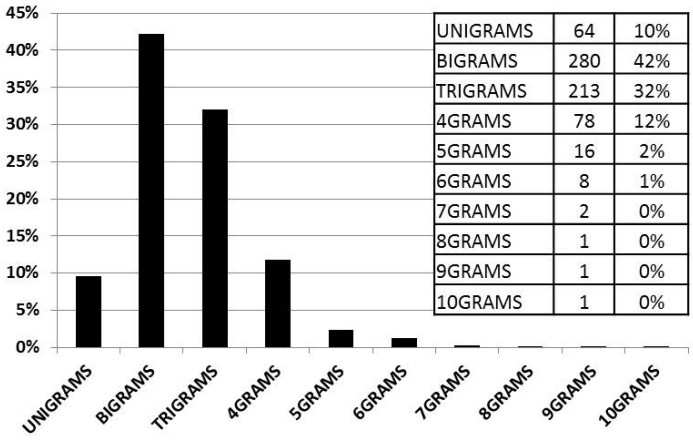
In the first stage of the process, the text from the TOCs is pre-processed. We remove punctuations, symbols, and numbers from the TOCs, so that only words are used for generating concept labels. After this, we remove 2 sets of stopwords. First, a standard English stopwords list, which allows us to remove common words and still retain a good set of words for generating our concept labels. Second, an additional set of words which we refer to as TOC-stopwords are removed. It contains: structural words, such as *chapter* and *appendix*, which relate to the structure of the TOCs; roman numerals, such as *xxiv* and *xxxv*, which are used to indicate the sections in a TOC; and words, such as *introduction* and *conclusion*, which describe parts of a learning material and are generic across domains. In addition, words referring directly to the name of the domain used for demonstration are removed, as we wish to generate concepts that describe the domain.

We do not use stemming because we found it harmful during pre-processing. When searching an encyclopedia source with the stemmed form of words, relevant results would not be returned. The output from pre-processing is a set of TOC phrases. In the next stage, we apply ngram extraction to the TOC phrases to generate all 1-3 grams from the entire set of TOC phrases. The output from this process are TOC-ngrams containing a set of 2038 unigrams, 5405 bigrams and 6133 trigrams, which

are used as the potential domain concept labels. Many irrelevant ngrams are generated from the TOCs because we have simply selected all 1-3 grams.

3.3 Verifying Concept Labels using Domain Lexicon

A domain lexicon is used to verify the generated TOC-ngrams to confirm which of the ngrams are relevant for the domain. Our domain lexicon contains a set of 664 Wiki-phrases, each of which is pre-processed by removing non-alphanumeric characters. The distribution of Wiki-phrases is shown in Figure 3. The 84% of the Wiki-phrases that are 1-3 grams are used for verification. The comparison of TOC-ngrams with the domain lexicon identifies the potential domain concept labels that are actually being used to describe aspects of the chosen domain in Wikipedia. During verification, ngrams referring directly to the title of the domain, e.g. *machine learning* and *data mining*, are not included in the Wiki-phrases because our aim is to generate concept labels that describe specific topics within the domain. Overall, a set of 17 unigrams, 58 bigrams and 15 trigrams are verified as potential concept labels. Bigrams yield the highest number of ngrams, which indicates that bigrams are particularly useful for describing topics in this domain.



UNIGRAMS	64	10%
BIGRAMS	280	42%
TRIGRAMS	213	32%
4GRAMS	78	12%
5GRAMS	16	2%
6GRAMS	8	1%
7GRAMS	2	0%
8GRAMS	1	0%
9GRAMS	1	0%
10GRAMS	1	0%

Figure 3: Distribution of Wiki-phrases used for verifying concept labels

3.4 Domain Concept Generation

Our domain concepts are generated after a second verification step is applied to the ngrams returned from the previous stage. Each ngram is retained as a concept label if all of 3 criteria are met. Firstly, if a Wikipedia page describing the ngram exists. Secondly, if the text describing the ngram is not contained as part of the page describing another ngram. Thirdly, if the ngram is not a synonym of another ngram. For the third criteria, if two ngrams are synonyms, the ngram with the higher frequency is retained as a concept label while its synonym is retained as part of the extracted text. For example, 2 ngrams *cluster analysis* and *clustering* are regarded as synonyms in Wikipedia, so the text associated with them is the same. The label *clustering* is retained as the concept label because it occurs more frequently in the TOCs, and its synonym, *cluster analysis* is contained as part of the discovered text.

The concept labels are used to search Wikipedia pages in order to generate a description for the identified concept label. The search returns discovered text that forms a pseudo-document which includes the concept label. So, the concept label and pseudo-document pair make up a domain concept. Overall, 73 domain concepts are generated. Each pseudo-document is pre-processed using standard techniques of English stopwords removal and Porter stemming (Porter, 1980). The pseudo-document terms form the concept vocabulary that can be used to represent resources.

4 Representing Learning Resources Using Background Knowledge

Our background knowledge contains a rich representation of the learning domain and by harnessing this knowledge for representing learning resources, we expect to retrieve documents based on the do-

main concepts that they contain. These concepts are designed to be effective for e-Learning, because they are assembled from TOCs of teaching materials (Agrawal et al., 2012). We present two approaches which have been developed by employing our background knowledge in the representation of learning resources.

4.1 The CONCEPTBASED Document Representation approach

Representing documents with the concept vocabulary allows retrieval to focus on the concepts contained in the documents. Figures 4 & 5 illustrate the CONCEPTBASED method. Firstly, in Figure 4, the concept vocabulary, $t_1 \ldots t_c$, from the pseudo-documents of concepts, $C_1 \ldots C_m$, is used to create a term-concept matrix and a term-document matrix using TF-IDF weighting (Salton and Buckley, 1988). In Figure 4a, c_{ij} is the TF-IDF of term t_i in concept C_j, while Figure 4b shows d_{ik} which is the TF-IDF of t_i in D_k.

	C_1	...	C_j	...	C_m
t_1					
...					
t_i			c_{ij}		
...					
t_c					

(a) Term-concept matrix

	D_1	...	D_k	...	D_n
t_1					
...					
t_i			d_{ik}		
...					
t_c					

(b) Term-document matrix

Figure 4: Term matrices for concepts and documents

Next, documents $D_1 \ldots D_n$ are represented with respect to concepts by computing the cosine similarity of the term vectors for concepts and documents. The output is the concept-document matrix shown in Figure 5a, where y_{jk} is the cosine similarity of the vertical shaded term vectors for C_j and D_k from Figures 4a and 4b respectively. Finally, the document similarity is generated by computing the cosine similarity of concept-vectors for documents. Figure 5b shows z_{km}, which is the cosine similarity of the concept-vectors for D_k and D_m from Figure 5a. So, the CONCEPTBASED approach uses the document representation and similarity in Figure 5 to influence retrieval. We expect to retrieve documents that are similar based on the domain concepts that they contain.

	D_1	...	D_k	...	D_m	...	D_n
C_1							
...							
C_j			y_{jk}		y_{jm}		
...							
C_m							

(a) Concept-document matrix representation

	D_1	...	D_m	...	D_n
D_1	1				
...					
D_k			z_{km}		
...					
D_n					1

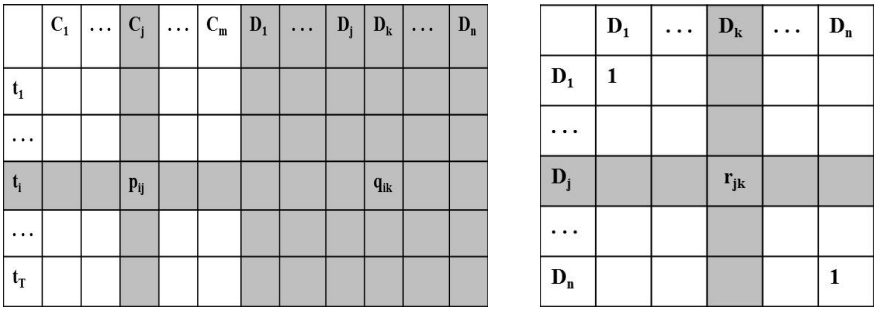
(b) Document-document similarity

Figure 5: Document representation and similarity using the CONCEPTBASED approach

4.2 The HYBRID Document Representation Approach

The HYBRID approach exploits the relative distribution of the vocabulary in the concept and document spaces to augment the representation of learning resources with a bigger, but focused, vocabulary as shown in Figure 6. So the TF-IDF weight of a term changes depending on its relative frequency in both spaces. First, our 73 domain concepts, $C_1 \ldots C_m$ from section 3.4, and the documents we wish to represent, $D_1 \ldots D_n$, are merged to form a corpus. Next, a term-document matrix with TF-IDF weighting is created using all the terms, $t_1 \ldots t_T$ from the vocabulary of the merged corpus as shown in Figure 6a. Entry q_{ik} is the TF-IDF weight of term t_i in D_k. If t_i has a lower relative frequency in the

concept space compared to the document space, then the weight q_{ik} is boosted. So, distinctive terms from the concept space will get boosted. Although the overlap of terms from both spaces are useful for altering the term weights, it is valuable to keep all the terms from the document space because this gives us a richer vocabulary. The shaded term vectors for $D_1 \ \ldots \ D_n$ in Figure 6a form a term-document matrix for documents whose term weights have been influenced by the presence of terms from the concept vocabulary.



(a) Hybrid term-document matrix representation

(b) Hybrid document similarity

Figure 6: Representation and similarity of documents using the HYBRID approach

Finally, the document similarity in Figure 6b, is generated by computing the cosine similarity between the augmented term vectors for $D_1 \ \ldots \ D_n$. Entry r_{jk} is the cosine similarity of the term vectors for documents, D_j and D_k from Figure 6a. The HYBRID method exploits the vocabulary in the concept and document spaces to influence the retrieval of documents.

5 Evaluating Learning Resource Representation

Our methods are evaluated on a collection of topic-labelled learning resources by simulating an e-Learning recommendation task. We use a collection from Microsoft Academic Search (MAS)(Hands, 2012), in which the author-defined keywords associated with each paper identifies the topics they contain. The keywords represent what relevance would mean in an e-Learning domain and we exploit them for judging document relevance. The papers from MAS act as our e-Learning resources, and using a query-by-example scenario, we evaluate the relevance of a retrieved document by considering the overlap of keywords with the query. This evaluation approach allows us to measure the ability of the methods to identify relevant learning resources.

We compare the performance of our CONCEPTBASED and HYBRID methods against that of Bag of Words (BOW). The BOW is a standard Information Retrieval method where documents are represented using terms from the document space only with TF-IDF weighting. For each of the 3 methods, the documents are first pre-processed by removing English stopwords and applying Porter stemming. Then, after representation, a similarity-based retrieval is employed using cosine similarity.

5.1 Evaluation Method and Dataset

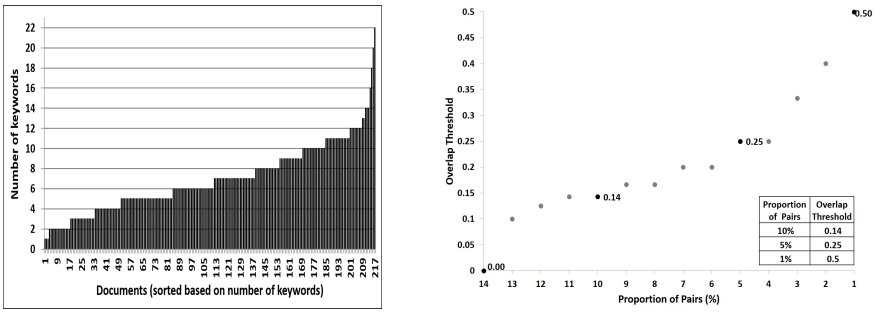
Evaluations using human evaluators are expensive, so we take advantage of the author-defined keywords for judging the relevance of a document. The keywords are used to define an overlap metric. Given a query document Q with a set of keywords K_Q, and a retrieved document R with its set of keywords K_R, the relevance of R to Q is based on the overlap of K_R with K_Q. The overlap is computed as:

$$Overlap(K_Q, K_R) = \frac{|K_Q \cap K_R|}{min\left(|K_Q|, |K_R|\right)} \tag{1}$$

We decide if a retrieval is relevant by setting an overlap threshold, and if the overlap between K_Q and K_R meets the threshold, then K_R is considered to be relevant.

Figure 7 shows the number of keywords per document and the overlap of document pairs for the first dataset used. Our first dataset which we refer to as dataset 1 contains 217 Machine Learning and Data Mining papers. A distribution of the keywords per document is shown in Figure 7a, where the documents are sorted based on the number of keywords they contain. There are 903 unique keywords, and 1,497 keywords in total. A summary of the overlap scores for all document pairs is shown in Figure 7b. There are $23,436$ entries for the 217 document pairs, and $20,251$ are zero, meaning that there is no overlap in 86% of the data. So only 14% of the data have an overlap of keywords, indicating that the distribution of keyword overlap is skewed. There are 10% of document

pairs with overlap scores ≥ 0.14, and 5% are ≥ 0.25. For experiments with this dataset we use 0.14 and 0.25 as thresholds, thus avoiding extreme values that would allow either very many or few of the documents to be considered as relevant.



(a) # of keywords per Document (b) Overlap of document pairs

Figure 7: Number of keywords per document and overlap profile of document pairs in dataset 1

Our interest is in the topmost documents retrieved, because we want our top recommendations to be relevant. We use precision@n to determine the proportion of relevant documents retrieved:

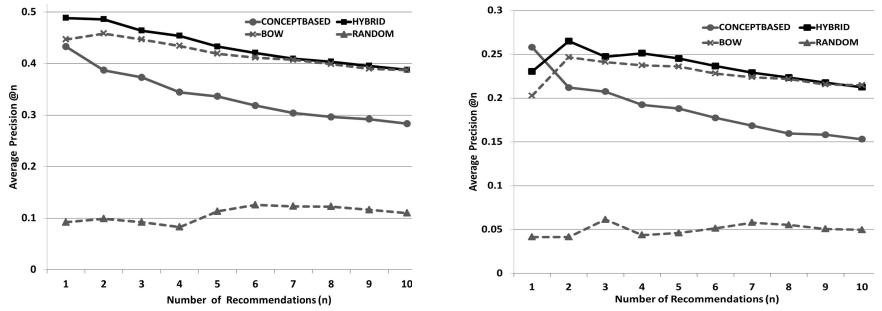
$$Precision@n = \frac{|retrievedDocuments \cap relevantDocuments|}{n} \tag{2}$$

where, n is the number of documents retrieved each time, *retrievedDocuments* is the set of documents retrieved, and *relevantDocuments* are those documents that are considered to be relevant i.e. have an overlap that is greater than the threshold.

5.2 Evaluation Results

The methods are evaluated using a leave-one-out retrieval. In Figure 8, the number of recommendations (n) is shown on the x-axis and the average precision@n is shown on the y-axis. RANDOM(▲) has been included to give an idea of the relationship between the threshold and the precision values. RANDOM results are consistent with the relationship between the threshold and the proportion of data in Figure 7b.

Overall, HYBRID(■) performs better than BOW(×) and CONCEPTBASED(●), showing that augmenting the representation of documents with a bigger, but focused vocabulary, as done in HYBRID, is a better way of harnessing our background knowledge. BOW also performs well because the document vocabulary is large, but the vocabulary used in CONCEPTBASED may be too limited. The complexity of the representation method in HYBRID overcomes the limitation of CONCEPTBASED. All the graphs fall as the number of recommendations, *n* increases. This is expected because the earlier retrievals are more likely to be relevant. However, the overlap of HYBRID and BOW at higher values of n may be because the documents retrieved by both methods are drawn from the same neighbourhoods.



(a) Results at Threshold of 0.14 (b) Results at Threshold of 0.25

Figure 8: Precision of the methods at overlap thresholds of 0.14 and 0.25 on dataset 1

The relative performance at a threshold of 0.25 in Figure 8b, is similar to the performance at 0.14. However, at this more challenging threshold, HYBRID and BOW do not perform well on the first retrieval. Generally, the results show that the HYBRID method is able to identify relevant learning

resources by highlighting the domain concepts they contain, and this is important in e-Learning. The graphs show that augmenting the representation of learning resources with our background knowledge is beneficial for e-Learning recommendation.

6 Refined Background Knowledge

One issue with the previous concept generation method is that the concept vocabulary produced was limited. A suitable representation for e-Learning resources should have a good coverage of relevant domain topics. In this section, we discuss the steps taken to refine our method used for generating domain concepts in order to improve our background knowledge and increase the coverage of our concept vocabulary.

6.1 Enriched Domain Concepts

In developing this method, we go through the phases described in sections 3.2 - 3.4. First, in addition to the TOC stopwords, the SMART stopwords (Salton, 1971) are also removed during pre-processing. This allows us to remove words that do not contribute to learning terms, and still retain a good set of words for generating our concepts. Second, words referring to the name of the domain used for demonstration such as: *machine*, *learning*, *data*, and *mining* are not removed during pre-processing, as we observed that removing these words before ngram generation prevents other relevant ngrams such as *instance based learning* or *reinforcement learning*, that contain any of these words from being identified. Third, we increase our ngram extraction to generate 1-5 grams from our TOC-phrases because, a distribution of the Wiki-phrases in Figure 3 showed that 99% of phrases are 1-5grams; this allows us to increase the number of concepts we can generate.

We apply ngram extraction to the TOC-phrases to produce the following TOC-ngrams: 2467 Unigrams; 5387 Bigrams; 3625 Trigrams; 1668 Fourgrams; and 576 Fivegrams. The TOC-ngrams are verified as described in Section 3.3 using the Wiki-phrases to produce a set of potential concept labels containing 24 Unigrams; 96 Bigrams; 38 Trigrams; 6 Fourgrams; and no Fivegrams. A second verification step as described in Section 3.4 is applied to the potential concept labels. This entails using the verified ngrams to search Wikipedia pages in order to generate a domain concept. The search returns discovered text that forms a pseudo-document and a concept label. Overall, our refined method has 150 domain concepts that pass the second verification, each having a concept label and pseudo-document pair. The pseudo-document terms are pre-processed using standard techniques of English stopword removal and Porter Stemming. These terms now form the concept vocabulary of our refined background knowledge which we refer to as the CONCEPTBASED+ method.

6.2 Recommendation using the CONCEPTBASED+ approach

The CONCEPTBASED+ method employs the richer concept vocabulary of our refined background knowledge for representing documents. We expect the representation created using the CONCEPTBASED+ method to contain a better coverage of the learning domain because of the richer concepts it contains. Our aim is to address the issue of the limited concepts contained in the CONCEPTBASED method. For recommendation using CONCEPTBASED+, we use the same representation and document similarity as the CONCEPTBASED method illustrated in Figures 4 & 5, but with a richer concept vocabulary. So documents are represented with respect to concepts by computing the cosine similarity of term vectors for concepts and documents to produce a concept document matrix. Then, the similarity between documents can be generated by computing the similarity between respective concept vectors for documents.

By using the CONCEPTBASED+ method for representation, we expect to retrieve documents that are similar based on the concepts they contain, and this is obtained from a document-document similarity matrix as shown in Figure 5b. A standard approach of representing documents would be to define the document similarity based on the term document matrix illustrated in Figure 4b, but this exploits the concept vocabulary only. In our approach, we put more emphasis on the domain concepts, so we use the concept document matrix illustrated in Figure 5a, to underpin the similarity between documents. The CONCEPTBASED+ method combines the focus with breadth of a richer set of domain concepts when representing documents.

6.3 Evaluating the Refined Representation

This section investigates whether the domain concepts generated using a refined approach i.e. CONCEPTBASED+ are better for representing documents than concepts generated with a standard method

i.e.CONCEPTBASED. The same evaluation method and dataset 1 presented in Section 5.1 is adopted here, and a leave-one-out retrieval is applied for evaluating the methods. In Figure 9, the number of recommendations is shown on the x-axis while the average precision@n is shown on the y-axis. An overlap threshold of 0.14 is used because there are 10% of document pairs in this dataset with overlap scores ≥ 0.14.

The performance of CONCEPTBASED+(♦) is shown by the darker line, and CONCEPTBASED(●) by the gray line. BOW(×) is included as the benchmark and RANDOM(▲) gives an idea of the relationship between the threshold used and the precision values. The graphs of all the methods fall as the number of recommendations, n increases. This is expected as earlier retrievals are more likely to be relevant. Overall, CONCEPTBASED+ outperforms CONCEPTBASED, BOW, and RANDOM, by producing better recommendations for all values of n. This performance shows the advantage of using the richer concept vocabulary for representing learning materials. The results confirm that CONCEPTBASED+ contains concepts that have a better coverage of the learning domain than CONCEPTBASED which has a limited set of concepts. So we adopt CONCEPTBASED+ as a background knowledge representation for learning materials in this domain.

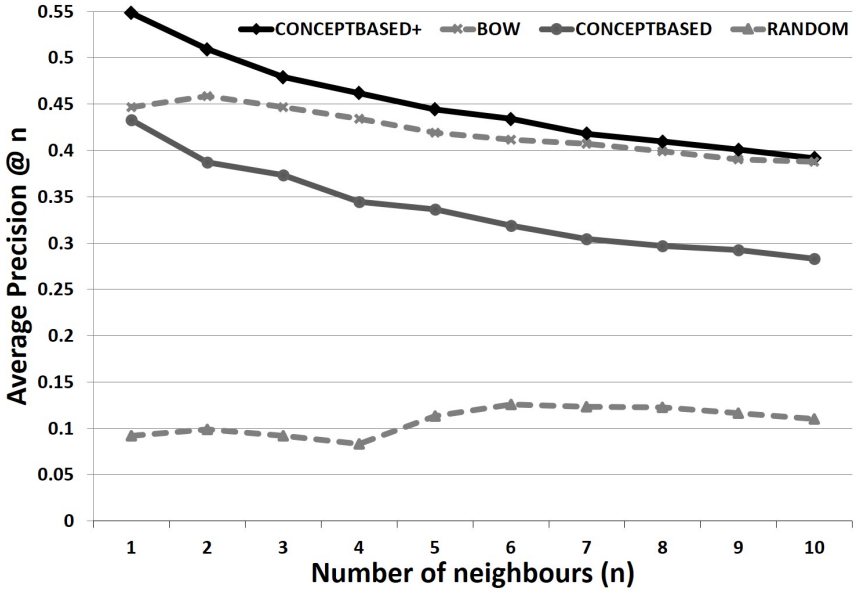
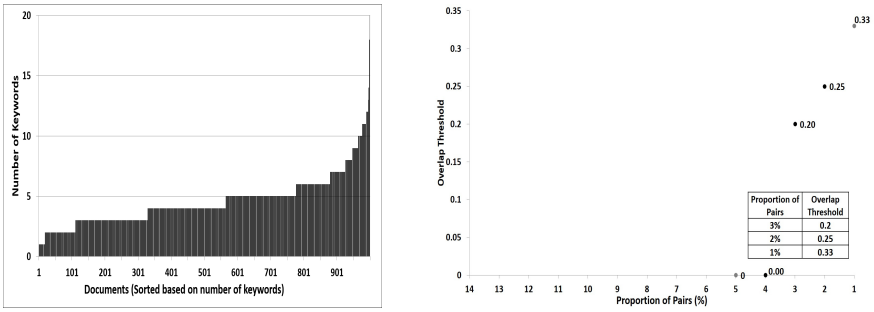


Figure 9: Comparing CONCEPTBASED and CONCEPTBASED+ at a threshold of 0.14 on dataset 1

6.4 Evaluation Using a Larger Dataset

We compare the performance of our HYBRID and CONCEPTBASED+ methods against that of the standard BOW approach on a larger dataset, in order to confirm our findings from the previous experiments. Figure 10 contains the number of keywords per document and the overlap of document pairs for the second dataset used. Our second dataset which we refer to as dataset 2 contains 1000 Machine Learning and Data Mining papers also from Microsoft Academic Research. Figure 10a contains a distribution of the keywords per document, where the documents are sorted based on the number of keywords they contain. There are 3063 unique keywords, and 4551 keywords in total. We take advantage of these author-defined keywords for judging relevance. A summary of the overlap profile of document pairs for dataset 2 is shown in Figure 10b. There are 499,500 entries for the 1000 document pairs, and 480,129 entries are zero, meaning that there is no overlap in 96% of the data. So only 4% of the data have an overlap of keywords, indicating that the distribution of keyword overlap is skewed. There are 3% of document pairs with overlap scores ≥ 0.2. The same evaluation method presented in 5.1 is used here. Then a leave-one-out retrieval method is applied, and precision@n as given in Equation 2 is used to determine the proportion of relevant documents retrieved. With dataset 2, we use a threshold of 0.2 thus preventing values that allow either too many or few documents to be considered as relevant. In Figure 11, the number of recommendations is shown on the x-axis and the average precision@n is on the y-axis. The average precision values are based on the overlap of keywords between document pairs and the threshold value chosen for the experiment. RANDOM(▲) gives an idea of the relationship between the threshold and the precision values, and the results are consistent with the overlap profile in Figure 10b.



(a) # of keywords per document

(b) Overlap of document pairs

Figure 10: Number of keywords per document and overlap profile of document pairs in dataset 2

On this bigger dataset, CONCEPTBASED+(♦) method outperforms HYBRID(■), BOW(×), and CONCEPTBASED(●), confirming that using a richer and focused vocabulary to represent documents is useful for e-Learning recommendation. The results also show HYBRID performing better than BOW, again confirming that augmenting the representation of learning resources with domain concepts is better than using the content only for e-Learning recommendation. Experiments were also run at thresholds of 0.25 and 0.33 and the relative performance at these thresholds is similar to the performance at 0.2, so the graphs are not shown. Our results show that we are able to leverage on the vocabulary from CONCEPTBASED+ which is not only a larger vocabulary, but one focused on domain concepts, thus allowing our method to influence the retrieval and recommendation of relevant learning resources.

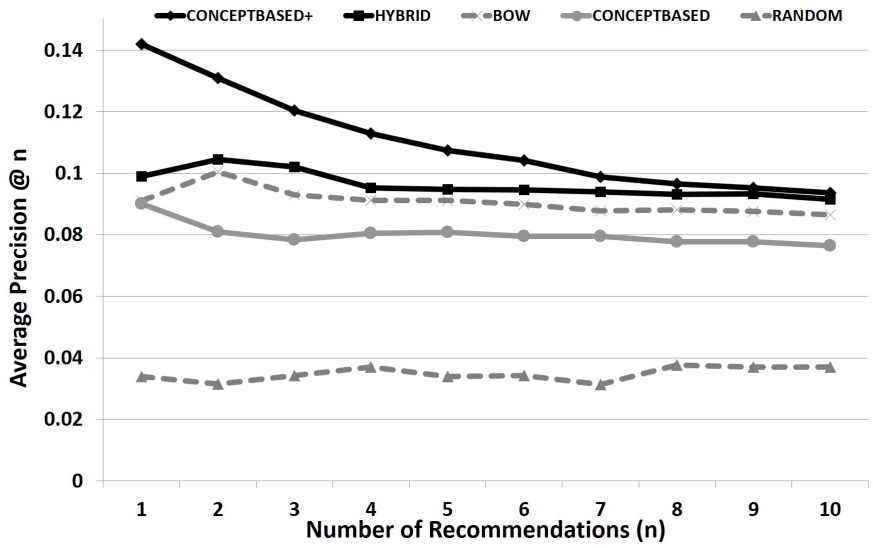


Figure 11: Precision of the methods at overlap threshold of 0.2 on dataset 2

7 Conclusions

The growing availability of e-Learning materials on the Web provides opportunities for learners to easily access new and valuable information. However, finding good materials is difficult because retrieval has to overcome the challenge of ineffective queries often input by learners. e-Learning recommendation offers a possible solution to this difficulty. Though, recommendation in e-Learning environments is challenging because the learning materials are often unstructured text, and so are not properly indexed for retrieval. We address this challenge by creating a method that automatically acquires background knowledge in the form of a rich set of concepts related to the selected learning domain. In building our method, we take advantage of the knowledge of experts contained in the TOCs of e-Books to identify relevant domain topics. By using e-Books we benefit from the provenance associated with these teaching materials. The identified topics are enriched with discovered

text from Wikipedia, and this extends the coverage and richness of our representation.

CONCEPTBASED method takes advantage of similar distributions of concept terms in the concept and document spaces to define a concept-term driven representation. Although the concept vocabulary in CONCEPTBASED is limited, HYBRID exploits the relative distribution of the vocabulary in the concept and document spaces to augment the representation of learning resources with a larger vocabulary influenced by domain concepts. CONCEPTBASED+ provides a richer concept vocabulary that allows concept-based distinctiveness to be helpful in the representation and retrieval of documents. This refined method allows us to generate a richer and focused set of domain concepts, which provides a better coverage of the domain. The performance of CONCEPTBASED+ in our evaluation shows the advantage of using the richer concept vocabulary for representing learning materials. Our results confirm an improvement in e-Learning recommendation when a rich concept vocabulary is used for representing learning resources.

References

Agrawal, R., Chakraborty, S., Gollapudi, S., Kannan, A., and Kenthapadi, K. (2012). Quality of textbooks: An empirical study. In *ACM Symposium on Computing for Development*, pages 16:1–16:1. doi: 10.1145/2160601.2160623.

Beliga, S., Meštrović, A., and Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences*, 39(1):1–20.

Blei, D. M. and McAuliffe, J. D. (2007). Supervised topic models. In *Neural Information Processing Systems*, pages 121–128.

Bousbahi, F. and Chorfi, H. (2015). MOOC-Rec: A case based recommender system for MOOCs. *Procedia - Social and Behavioral Sciences*, 195:1813 – 1822. doi: 10.1016/j.sbspro.2015.06.395.

Boyce, S. and Pahl, C. (2007). Developing domain ontologies for course content. *Journal of Educational Technology & Society*, 10(3):275–288.

Chen, W., Niu, Z., Zhao, X., and Li, Y. (2014). A hybrid recommendation algorithm adapted in e-learning environments. *World Wide Web*, 17(2):271–284. doi: 10.1007/s11280-012-0187-z.

Chen, Z. and Liu, B. (2014). Topic modeling using topics from many domains, lifelong learning and big data. In *31st International Conference on Machine Learning*, pages 703–711.

Clarà, M. and Barberà, E. (2013). Learning online: Massive Open Online Courses (MOOCs), connectivism, and cultural psychology. *Distance Education*, 34(1):129–136. doi: 10.1080/01587919.2013.770428.

Coenen, F., Leng, P., Sanderson, R., and Wang, Y. J. (2007). Statistical identification of key phrases for text classification. In *Machine Learning and Data Mining in Pattern Recognition*, pages 838–853. Springer. doi: 10.1007/978-3-540-73499-4_63.

Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N., and Taibi, D. (2012). Linked education: Interlinking educational resources and the Web of data. In *27th Annual ACM Symposium on Applied Computing*, pages 366–371. 10.1145/2245276.2245347.

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., and Mueller, E. T. (2013). Watson: Beyond Jeopardy! *Artificial Intelligence*, 199:93–105. doi: https://doi.org/10.1016/j.artint.2012.06.009.

Gherasim, T., Harzallah, M., Berio, G., and Kuntz, P. (2013). Methods and tools for automatic construction of ontologies from textual resources: A framework for comparison and its application. In *Advances in Knowledge Discovery and Management*, pages 177–201. Springer. doi: 10.1007/978-3-642-35855-5_9.

Hands, A. (2012). Microsoft academic search. *Technical Services Quarterly*, 29(3):251–252. doi: 10.1080/07317131.2012.682026.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195. doi: 10.3233/SW-140134.

Matsuo, Y. and Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169. doi: 10.1142/S0218213004001466.

Mbipom, B., Craw, S., and Massie, S. (2016). Harnessing background knowledge for e-learning recommendation. In *Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV*, pages 3–17. Springer. doi: 10.1007/978-3-319-47175-4_1.

Milne, D. and Witten, I. H. (2008). Learning to link with Wikipedia. In *17th ACM Conference on Information and Knowledge Management*, pages 509–518. doi: 10.1145/1458082.1458150.

Nasraoui, O. and Zhuhadar, L. (2010). Improving recall and precision of a personalized semantic search engine for e-learning. In *4th IEEE International Conference on Digital Society*, pages 216–221. doi: 10.1109/ICDS.2010.63.

Panagiotis, S., Ioannis, P., Christos, G., and Achilles, K. (2016). APLe: Agents for personalized learning in distance learning. In *7th International Conference on Computer Supported Education*, pages 37–56. Springer. doi: 10.1007/978-3-319-29585-5_3.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Qureshi, M. A., O'Riordan, C., and Pasi, G. (2014). Exploiting Wikipedia to identify domain-specific key terms/phrases from a short-text collection. In *Proceedings of the 5th Italian Information Retrieval Workshop*, pages 63–74.

Rodrigues, L., Antunes, B., Gomes, P., Santos, A., Barbeira, J., and Carvalho, R. (2007). Using textual CBR for e-learning content categorization and retrieval. In *4th International Conference on Case-Based Reasoning workshop on textual case-based reasoning*.

Ruiz-Iniesta, A., Jimenez-Diaz, G., and Gomez-Albarran, M. (2014). A semantically enriched context-aware OER recommendation strategy and its application to a computer science OER repository. *IEEE Transactions on Education*, 57(4):255–260. doi: 10.1109/TE.2014.2309554.

Salton, G. (1971). *The SMART Retrieval System-Experiments in Automatic Document Processing*. Prentice-Hall, Upper Saddle River, NJ, USA. doi: 10.1109/TPC.1972.6591971.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523. doi: 10.1016/0306-4573(88)90021-0.

Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., and Studer, R. (2006). Semantic Wikipedia. In *Proceedings of the 15th International Conference on World Wide Web*, pages 585–594. ACM. doi: 10.1145/1135777.1135863.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. (1999). KEA: Practical automatic keyphrase extraction. In *4th ACM Conference on Digital libraries*, pages 254–255. doi: 10.1145/313238.313437.

Yang, H.-L. and Lai, C.-Y. (2010). Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26(6):1377 – 1383. doi: 10.1016/j.chb.2010.04.011.

Yang, K., Chen, Z., Cai, Y., Huang, D., and Leung, H. F. (2016). Improved automatic keyword extraction given more semantic knowledge. In *International Conference on Database Systems for Advanced Applications*, pages 112–125. Springer. doi: 10.1007/978-3-319-32055-7_10.

Yarandi, M., Tawil, A.-R., and Jahankhani, H. (2011). Adaptive e-learning system using ontology. In *In Proceedings of the 22nd International Workshop on Database and Expert Systems Applications*, pages 511–516. doi: 10.1109/DEXA.2011.9.

Zhang, X., Liu, J., and Cole, M. (2013). Task topic knowledge vs. background domain knowledge: Impact of two types of knowledge on user search performance. In *Advances in Information Systems and Technologies*, pages 179–191. Springer. doi: 10.1007/978-3-642-36981-0_17.

Zheng, Z., Li, F., Huang, M., and Zhu, X. (2010). Learning to link entities with knowledge base. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 483–491.