Knowledge-based Content Linking for Online Textbooks

Rui Meng, Shuguang Han, Yun Huang, Daqing He and Peter Brusilovsky
School of Information Sciences, University of Pittsburgh
135 N Bellefiled Avenue, Pittsburgh, PA, 15213
Email: {rui.meng,shh69,yuh43,dah44,peterb}@pitt.edu

***Abstract*—Although the volume of online educational resources has dramatically increased in recent years, many of these resources are isolated and distributed in diverse websites and databases. This hinders the discovery and overall usage of online educational resources. By using linking between related subsections of online textbooks as a testbed, this paper explores multiple knowledge-based content linking algorithms for connecting online educational resources. We focus on examining semantic-based methods for identifying important knowledge components in textbooks and their usefulness in linking book subsections. To overcome the data sparsity in representing textbook content, we evaluated the utility of external corpuses, such as more textbooks or other online educational resources in the same domain. Our results show that semantic modeling can be integrated with a term-based approach for additional performance improvement, and that using extra textbooks significantly benefits semantic modeling. Similar results are obtained when we applied the same approach to other domains.**

1. Introduction

The Internet has dramatically increased both the volume and variety of online educational resources, such as online textbooks, online courses, and tutorials. This has prompted the development of online educational repositories to collect and organize high-quality educational resources, as well as the satisfaction of learning-related queries through modern search techniques. However, due to the high variability in different learners' backgrounds, it is still a challenge to find "the right content" to precisely match each learner's individual goals, interests, and background knowledge. Some might be too simple or too advanced for a learner, whereas others might be not novel or highly redundant. This occurs when people may want to obtain alternative resources that not only contain similar knowledge, but that also fulfill a learner's individual background. To connect different educational resources with similar knowledge, we can adopt the techniques developed for the intelligent linking of hypertext [1], [2].

One particularly valuable context for such an intelligent linking technique is in knowledge linking for textbooks. Multiple textbooks are available online for almost any subject. Each textbook is usually well-structured as hypertext, with standard hierarchical and linear navigation; yet, multiple textbooks on the same subject are usually not connected to one another. The research presented in this paper was motivated by the need to provide intelligent connections between similar sections in multiple online textbooks. We want to support a learner who finds that a relevant section in one textbook is either too complicated or too primitive to match her needs with the ability to navigate to an alternate presentation of the same knowledge component in different textbooks.

While this functionality has been explored before [1], [3], past work has mostly focused on connecting related content through term-level similarity. In our work, we wanted to follow a different intelligent linking trend based on semantic similarity. Semantic linking has been explored in the context of connecting Web pages [4], [5], but has not yet been investigated in the context of linking different textbook sections. Learners who have no advanced knowledge on a topic would have a lower ability to distinguish high- and low-quality content, and thus would require more powerful intelligent linking technologies. The key idea behind our approach follows the arguments of semantic linking research: the "aboutness" of a document (in our case, a textbook section) is defined by its concepts, not by the collection of terms that are used to present these concepts. Following this idea, true similarity should be determined on the basis of *knowledge similarity* between two documents, and not just on term overlapping. Meanwhile, a robust semantic representation often requires a large-scale data corpus, whereas textbooks are often highly condensed. We observe that there are plenty of online digital resources for a given domain, such as PubMed for medical topics and ACM Digital Library for computer science. We also include these resources and expect that they can help provide a better semantic representation. To summarize, this paper studies the following two research questions:

- **RQ 1**: Can a semantic representation of book content help perform knowledge-based textbook linking, either by acting alone or by being combined with traditional term-based representation?
- **RQ 2**: Can external resources help perform knowledge-based textbook linking, either by acting alone or by being combined with traditional term-based representation?

2. Related Work

The field of intelligent educational systems has accumulated more than 30 years of research experience in modeling educational content for student modeling and personalization. In this field, educational content modeling has been traditionally achieved by manually decomposing the entire body of knowledge about a particular domain into a set of knowledge components (KCs) [6]. Each KC represents an elementary fragment of knowledge for the domain. In a simple model, the set of KCs has no internal structure. In more advanced domain models, KCs are related to each other,

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.14

which forms a semantic network [7], [8]. While the overall use of KCs has demonstrated its usefulness in delivering different kinds of personalization [9], the KC concept has some critical limitations in scaling up to cover large volumes of online materials, as well as coping with the rapidly changing domains [10], because in these studies, KCs are often manually crafted. Therefore, automatically identifying and linking important KCs in a domain has emerged as an active research topic in adaptive educational hypermedia, and has become part of a broader research stream on "open corpus" adaptive hypermedia [7].

Automatic content modeling has been viewed as the basis for various tasks, including user modeling [6], text classification [11], and ad-hoc retrieval [12]. There are several methods for automatic content modeling. The simplest approach is the term-based method, where the frequencies of terms in the documents and in the collection are counted [13]. However, the term-based method is unable to capture the semantic knowledge that is present behind the content. Therefore, many studies have also explored the automatic representation of educational materials using semantic representation. For example, the Conceptual Open Hypermedia Services Environment (COHSE) [4] has proposed a representation of online educational resources with metadata; through integration with an ontological reasoning service, it can form a conceptual hypermedia educational system. Gauch et al. explored various ontology-based methods for content representation [14], where both the concepts presented in the content and the content of user profiles are connected through ontology.

More recently, various latent topic modeling approaches have been used for automatic content modeling. By representing the content of documents as finite mixtures over a set of latent topics, topic modeling helps move beyond term-based approaches, but simultaneously avoids the rather expensive development cost in some earlier rule-based and ontology-based methods. Particularly, Blei's latent dirichlet allocation (LDA) [15] has also been explored many times in content modeling [16], [17].

Textbooks have been a focus of research on knowledge modeling in adaptive educational hypermedia since the early days of this field [7], [18]. Because of their high quality and good coverage for a given subject domain, they have been identified as handy resources for evaluating automatic knowledge extraction and representation algorithms. Guerra et al. [19] examined the impact of LDA-based topic modeling for supporting semantic representation through fine-grained textbook linking. Their results showed that LDA can be a better representation method than the standard term-based approach at finding similar documents.

3. Methodology

This paper aims to explore different content modeling approaches for textbook linking. To be specific, we attempt to identify similar book subsections from different textbooks, as we think that textbooks are carefully designed by their authors to organize knowledge for a given field, as well as that each book subsection contains certain knowledge components. Therefore, a better content representation method should be able to capture the latent knowledge components in each subsection, which leads to superior book subsection linking performance. Consequently, the overall linking accuracy can be used for assessing knowledge-based modeling algorithms. The evaluation framework for book subsection linking is adopted from Guerra et al. [19].

Since different content linking approaches may generate different performances, we want to explore the performances of fusing multiple modeling approaches. In addition, we also plan to explore the value of external data sources for content modeling. Using these external online resources for better identification of concepts in a specific domain could have potential impacts on the overall quality of modeling.

3.1. Problem Definition

We formally define the research problem as: for a given subsection C_i^1 in $BOOK1$, our goal is to find the corresponding subsections $\mathcal{C}^2$ in $BOOK2$. Note that $\mathcal{C}^2$ can be one or more subsections; therefore, instead of providing the best-matched book subsection, we rank all subsections in $BOOK2$ based on their similarity to C_i^1. This process can be illustrated as shown in Figure 1, where we first represent each book subsection as a vector of knowledge components, and then compute the similarities between sections as similarity between their vectors (using cosine similarity or other approach). Knowledge-based representation of a book subsection is another important element in Figure 1, which we will introduce in Section 3.2.

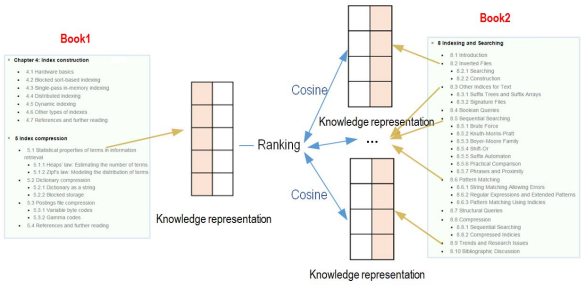

Figure 1. An illustration of knowledge-based linking for textbook subsections. Each book subsection is represented by a vector of knowledge components, and the similarity between two book subsections is measured by cosine similarity.

3.2. Knowledge Component Extraction

The most straightforward approach for KC extraction from textbooks is the **term-based approach**, in which we make a naïve assumption and treat each term (i.e., word) as a knowledge component. This approach serves as our baseline. Commonly, similar or related terms are often used to describe the same knowledge component, such as synonyms and fixed phrases. Thus, we further consider two **semantic-based approaches** for knowledge component extraction and attempt to capture the underlying connections between terms.

In the first approach, we adopt unsupervised topic models to automatically extract latent topics from the given data corpus. This paper uses the basic LDA model proposed by Blei et al. [15], and thus, we name this approach as the *LDA-based approach*. The topic modeling performance is highly related to the data sufficiency. However, two textbooks (i.e., $BOOK1$ and $BOOK2$, as indicated in Figure 1) do not contain adequate information for modeling purposes. To

address this problem, we explore the opportunity of incorporating external data resources. In this paper, we consider two types of external resources: 1) additional textbooks beyond $BOOK1$ and $BOOK2$, but that are still within the same domain; and 2) open data corpora that consist of related conference publications in the same domain. Note that the latter resource might not be readily available in a domain without related academic conferences.

One important problem for topic modeling is any mismatch between the identified topics and human interpretation [20]. Therefore, our second approach explores a way of having humans directly identify knowledge components. In this case, we use author-assigned keywords in scientific publications as candidate knowledge components. Author-assigned keywords delivered the best performance in our recent study of knowledge transfer between systems [21]. In this study, we used the author-assigned keywords in conference publications as domain knowledge components. Specifically, we mined 215,692 ACM conference publications to obtain 168,940 author keywords, which we treated as knowledge components. If one of these knowledge components is mentioned in a book subsection, this component will be used to represent the given book subsection (we also consider the number of occurrences of each knowledge component). Since the goal of this approach is to assign human-understandable concepts for each book subsection, we name it the *concept-based approach*.

3.3. Knowledge-based Linking for Textbooks

After obtaining knowledge components for each book subsection, we then compute the relative importance of each knowledge component. For both term-based and concept-based approaches, we employ the TF-IDF weighting for each knowledge component (i.e., term or concept). In the LDA-based method, we use the probabilistic distributions over all topics to measure the importance of each knowledge component (i.e., the latent topic). This is the knowledge representation process, as shown in Figure 1. When determining the corresponding book subsections of $BOOK2$ for a given subsection C_i^1 in $BOOK1$, we compute all similarity scores between C_i^1 and any book subsection C_j^2 in $BOOK2$. The ranking is then based on the similarity scores, and the top-ranked book subsections are treated as the best-matched book subsections for intelligent linking.

4. Experimental Setup

4.1. Data Collection

To link the knowledge components among book subsections, we consider two widely used textbooks in the information retrieval field - one is Manning et al. (Introduction to Information Retrieval; in short, *IIR*) and the other is Baeza-Yates et al. (Modern Information Retrieval; in short, *MIR*). We chose the information retrieval field because, first, the authors have been working on information retrieval research for many years, which makes it easier to understand and assess the knowledge structure for different textbooks; and, second, as information retrieval is a relatively well-developed field, there are various external data resources to enhance our knowledge linking.

Specifically, we consider two types of external corpora. The first one contains other information retrieval textbooks, excluding the MIR and IIR books. In this study, we worked on three textbooks, which included *Finding Out About* by R. Belew, *Information Storage and Retrieval Systems* by Kowalski and Maybury and *Information Retrieval* by C. van Rijsbergen. The second corpus contains conference publications in the information retrieval field. For this corpus, we select all ACM SIGIR conference papers (a total of 3,150 papers) ranging from 1971 to 2013. These papers are available in ACM Digital Library, and only abstracts were extracted for building the corpus.

When extracting the content of each subsection from five textbooks, we remove stopwords, exclude non-letter symbols (such as brackets and punctuation) as well as numbers and terms which appear only once in the text, but do not perform stemming. The original average length of each subsection is 579 terms. After the above pre-processing, only 286 terms (161 unique terms after removing duplicates) are left. As for the concept extraction, we apply greedy matching on each subsection, to see which ACM concepts appear in the given subsection. After that, 164 unique concepts are extracted on average, and these contain both single-word and multi-word concepts.

4.2. Ground-truth

Instead of manually creating a new data corpus, we directly employed the ground-truth about mappings in book subsections in the information retrieval textbooks obtained from Guerra et al. [19]. There are two reasons for using this test collection and the ground-truth. First, Guerra et al. [19] provided the mapping between book subsections on MIR and IIR. Second, this method also included book subsection mapping for Algebra, which can be used for testing the ability of our algorithms to be generalized.

Specifically, the ground-truth dataset for information retrieval contains four chapters with 47 subsections for mapping IIR to MIR. Two experts were asked to provide mapping and to rate the relevance of each subsection-to-subsection mapping pair on a 3-point Likert scale (1 being lowest and 3 being highest). The final score of each book subsection pair was computed as the average of the scores provided by both experts. Non-matched parts of MIR were assigned zero relevance. In total, the 47 subsection from IIR are mapped to 88 subsections in MIR. The rest of the MIR subsections all have zero relevance.

4.3. Evaluation Metrics

As shown in the ground-truth discussion in Section 4.2, one subsection in IIR may map to more than one subsection in MIR. In fact, 55.3% of the ground-truth are one-to-one mapping relationships, 21.3% are one-to-two mapping relationships, 14.9% are one-to-three mapping relationships, and the rest are one-to-N (N>3) mapping relationships. Therefore, we adopted well-known ranking-based evaluation metrics for evaluation – NDCG@N. Here, we set N to be 1 and 3, since more than half of the mappings are either one-to-one relationships (55.3%), and the vast majority (91.5%) are one-to-N ($1 \leq N \leq 3$) relationships.

4.4. Experiment Design

To answer the research questions mentioned in Section 1, we designed the following two experiments. First, we exam-

ined the values of semantic-based approaches (both concept-based and LDA-based) by comparing their performances against the baseline – term-based approach. At the same time, since term- and semantic-based approaches try to capture a book's content at different levels, we also explored the ways of fusing them. In this paper, we merge different methods through a simple linear interpolation, as indicated in Equation 1.

$$S_{c_i,c_j}^{merged} = \lambda \cdot S_{c_i,c_j}^{term} + (1-\lambda) \cdot S_{c_i,c_j}^{semantic} \quad (1)$$

Our second experiment aimed at examining the values of external data resources for intelligent book content linking, where we considered an open data corpus containing thousands of academic conference publications and extra textbooks from the same domain. Because the term-based method only relies on the terms occurring in a book's subsection, and the concept-based approach only employs concepts from a predefined concept list, they both cannot explicitly incorporate external data resources, although we acknowledge that the knowledge components in the concept-based approach were obtained through a large-scale external ACM conference publications. Thus, the effectiveness of external resources was examined in the LDA-based approach.

There are two important parameters that our approaches needed to examine. They are the interpolation coefficient λ and the number of topics for LDA. Both of these parameters were determined based on their optimal performance in the training dataset. When choosing the best parameters, we varied nine different values for λ (from 0.1 to 0.9, with step 0.1) and five different values for the number of topics (10, 50, 100, 150, and 200). The training performance was based on the average of 5-fold cross-validation. To avoid sampling bias for train/test division, we randomly sampled the dataset 100 times. The reported performance in Section 5 is based on the average performance of 100 testing runs. The whole procedure is shown in Figure 2.

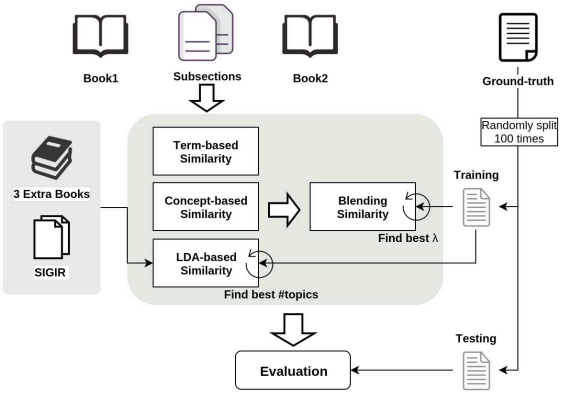

Figure 2. An illustration of our experimental procedure. Training and testing datasets are based on an 80/20 division. Training datasets are used for obtaining best-model parameters and testing datasets are used for reporting evaluation performance.

5. Experimental Results

This section presents the results of two sets of experiments: studying the value of semantic approaches, including a concept-based approach and a topic modeling approach (see Section 3.2); and examining how external resources can help model latent topics and further enhance the performance of book subsection mapping. Our resulting analysis includes extensive statistical tests using a non-parametric Wilcoxon signed-rank test, since the data are not distributed normally.

5.1. The Value of Semantic Approaches

5.1.1. Term- VS. Semantic-based Methods. The effectiveness of the term-based approach and the usefulness of the LDA-based approach for knowledge-based book subsection linking has been demonstrated in the literature [19]. However, it remains unclear whether the other semantic approaches that are based on knowledge concepts (i.e., author-assigned keywords for scientific publications) will work in this task.

Table 1 shows the book subsection mapping performance for the term-based approach and two semantic-based approaches, and presents two important results. First, the concept-based approach performs the worst among all three methods. One potential reason for this poor performance is the problem of data sparsity. For example, the average number of terms for each book subsection is 161 in the term-based approach, while there are only 72 concepts for each book subsection. Retaining only the concept-level content may exclude many useful textbook terms that have not been picked up by the authors from research papers (knowledge concepts used in our paper are determined by author-assigned keywords in research papers). Another potential reason is the mismatch between educational knowledge concepts in textbooks and research knowledge concepts in academic publications (which were used to generate candidate knowledge concepts). Educational knowledge concepts might be more well-established and generic, while research knowledge concepts assigned by the authors of a conference paper could be related to emerging research ideas, as well as being specific to that paper. We will further examine this hypothesis in Section 5.2.3.

Second, the LDA-based approach achieves the best performance at NDCG@1, which indicates that LDA does pick up the most important knowledge components for linking two book subsections that have the same content. However, it does not perform better than (or is comparable) to the term-based approach at NDCG@3. One possible explanation for this finding is a lack of sufficient content in a book subsection. LDA attempts to discover latent semantic relations among terms, based on their co-occurrences. Insufficient information on co-occurrences may not affect the modeling of the most important knowledge components, but can influence the identification of marginal important knowledge components. To address this issue, we will explore the effectiveness of involving an external data corpus for better topic modeling in Section 5.2.

TABLE 1. KNOWLEDGE-BASED BOOK SUBSECTION LINKING PERFORMANCE FOR TERM- AND SEMANTIC-BASED APPROACHES. ∗/† DENOTES A SIGNIFICANT CHANGE OVER THE TERM/CONCEPT-BASED METHODS. NUMBERS IN BOLD INDICATE THE BEST PERFORMANCE

	NDCG@1	**NDCG@3**
Term-based	0.3148(0.1184)	**0.5498(0.1150)**
Concept-based	0.1554(0.1027) ∗	0.3538(0.1115) ∗
LDA-based	**0.3807(0.1294)** ∗†	0.4681(0.1179) ∗†

5.1.2. Fusing Term- & Semantic-based Methods. Term-based and semantic-based approaches capture book content at different levels. Latent topics project book contents into semantic dimensions, but the transformation will cause information to be lost. On the other hand, the term-based approach retains the full depth of information, but does not measure term-term relations. An intuitive line of thinking is to combine these approaches into a unified framework, as shown in Formula (1). Table 2, which shows the results of different fusion approaches, reveals that fusions are better than the pure semantic approaches for both LDA-based and concept-based approaches. In particular, combining term and LDA-based topics achieves the best performance at NDCG@1, which indicates that LDA-based topics and terms capture the book content in complementary aspects, and they both contribute to generating better mapping performance.

TABLE 2. KNOWLEDGE-BASED BOOK SUBSECTION LINKING PERFORMANCE WITH DIFFERENT FUSION METHODS. ∗/†/◇ DENOTES A SIGNIFICANT PERFORMANCE CHANGE OVER A TERM-BASED/CONCEPT-BASED/LDA-BASED APPROACH. NUMBERS IN BOLD INDICATE THE BEST PERFORMANCE.

	NDCG@1	**NDCG@3**
Term-based	0.3148(0.1184)	0.5498(0.1150)
Concept-based	0.1554(0.1027)	0.3538(0.1115)
LDA-based	0.3807(0.1294)	0.4681(0.1179)
Term + Concept	0.2776(0.1248)∗†	0.5480(0.1220)∗†
Term + LDA	**0.3986(0.1327)**∗ † ◇	**0.5559(0.1176)**†◇

However, fusing terms and concepts still does not outperform the purely term-based approach, which might be due to a lack of data and the mismatch between research knowledge concepts and educational knowledge concepts, as mentioned in Section 5.1.1. Clearly, the LDA-based approach captures better semantic information about a book subsection than the research concepts assigned by authors (i.e., the concept-based approach). We find similar insights when plotting the performance change over different values of λ (which refers to the importance of the term-based approach) for the fusion methods in the training datasets[1] (see Figure 3). According to Figure 3, we find that Term + LDA has a concave shape, with $\lambda = 0.3$ achieving the best NDCG@1 and $\lambda = 0.6$ obtaining the best NDCG@3, while Term + Concept keeps increasing with a higher weight on the term-based method.

In terms of NDCG@3, although the fusion methods have significant improvements over purely semantic-based approaches, they still do not significantly outperform or are comparable to the term-based approach (though Term + LDA has a slightly better performance). Meanwhile, according to Figure 3, Term + LDA reaches the best performance when setting λ to 0.3 for NDCG@1 and 0.6 for NDCG@3. This suggests that we should include more information from a book subsection if we want to better model marginally important topics. In the next section, we would like to explore the ways of involving an external data corpus for better modeling of latent semantic topics in book subsections.

1. Note that this is different from the performance in the testing dataset; in this case, we varied different λ and chose the optimal one when generating the testing performance.

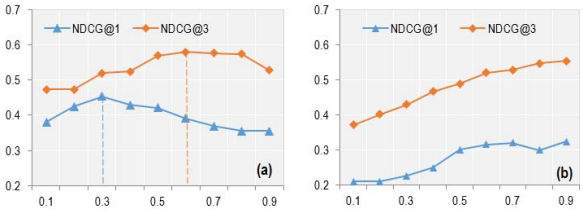

Figure 3. Knowledge-based book subsection linking performance in the *training datasets* for different fusion methods: (a) Term + LDA and (b) Term + Concept, which vary with different values of λ

5.2. The Value of the External Corpus

The experiments presented in Section 5.1 were all based on two information retrieval books (i.e., IIR and MIThereforeR), whose data are at a relatively small scale. Therefore, this section explores what and how external data resources can be used to better model latent semantic topics to further increase the book content mapping performance. When building the external data corpus, we encountered two different options: extra textbooks for the same domain, or an open data corpus that was related to the given field. Textbooks are written for educational purposes, which means textbooks on the same domain would be expected to have similar knowledge concepts. However, unless they are in a very popular domain, the number of textbooks available online is usually small. In contrast, an open data corpus, such as the ACM digital library, may contain a large number of conference and journal papers, but the knowledge concepts extracted from the conference/journal papers corpus might have different characteristics from those presented in textbooks. Therefore, it's worth spending time to carefully determine which of the two extra data resources is more helpful. We will discuss them separately in the next two subsections and compare their performances in the third subsection.

5.2.1. Textbook-based External Corpus. This section examines the value of using three additional textbooks to information retrieval, as listed in Section 4.1. Since the Term-based and Term + LDA approaches achieved the best performances in the book subsection linking experiments in previous sections, we include them as two baselines. According to Table 2, the Term + LDA outperforms the pure LDA-based approach; therefore, our experiments presented here were all performed on top of the Term + LDA approach. The NDCG values for incorporating external textbooks are provided in Table 3. We find that Term + LDA (+3 extra textbooks) significantly outperforms the two baselines on NDCG@3, beats the Term-based method, and does not differ significantly from Term + LDA for NDCG@1. This result clearly indicates the value of a textbook-based external data corpus.

5.2.2. Academic Publication-based External Corpus. This section examines the value of using conference publications as an external data corpus. To build such a corpus, we downloaded about 3,150 publications from the ACM SIGIR conference. Again, we keep the Term-based and Term + LDA approaches as two baselines, and run our new approach on top of the Term + LDA method. Experimental results with this new external data corpus are provided in Table 4, where we find, to our surprise, that

TABLE 3. KNOWLEDGE-BASED BOOK SUBSECTION LINKING PERFORMANCE FOR TEXTBOOK-BASED EXTERNAL DATA. ∗/† DENOTES A SIGNIFICANT PERFORMANCE CHANGE OVER THE TERM-BASED/TERM + LDA PERFORMANCE. NUMBERS IN BOLD INDICATE THE BEST PERFORMANCE.

	NDCG@1	**NDCG@3**
Term-based	0.3148(0.1184)	0.5498(0.1150)
Term + LDA	**0.3986(0.1327)**	0.5559(0.1176)
Term + LDA (+ 3 extra textbooks)	0.3878(0.1160)∗	**0.5985(0.1115)** ∗†

adding SIGIR publications does not improve the mapping performance. On the contrary, adding these publications even reduces the NDCG@3 performance. We think that this finding is probably due to the topic mismatching problem between textbooks and academic publications. Textbooks may contain only well-established and important knowledge concepts, whereas conference publications generally have a relatively wider range of topic interests and tend to explore novel research areas. We think that the publication date and number of citations can be used for measuring maturity and importance of a research topic, from which we can generate the following two hypotheses.

- H1: publications from earlier years provide a greater utility for textbook knowledge modeling.
- H2: publications with a larger number of citations provide a greater utility for textbook knowledge modeling.

To examine these two hypotheses, we conducted the following two experiments. The first experiment divides the whole SIGIR data corpus into three time periods (year 1971 - 2001, 2002 - 2008 and 2009 - 2015) to make sure that each period contains around the same number of publications (around 1,000 publications). In the second experiment, we divide the SIGIR data corpus into three groups, based on the number of citations. In total, we obtain three groups (with #citations 0 - 2 as low, 3 - 12 as medium, and more than 12 as high) to make each consist of around 1,000 publications. The results for each experiment are provided in Table 4, where we use italics to highlight the best performances within each experiment group.

Table 4 shows that incorporating SIGIR publications from 1971 to 2001 generates the best performance among all three subgroups in the temporal-based data corpus division – it significantly outperforms 2009 - 2015 at NDCG@1 and 2002 - 2008 at NDCG@3. We also find that incorporating publications with medium and high numbers of citations (the subgroups of medium or high have no significant difference from one another) achieve the best performance in the citation-based publication division. These results are consistent with our two hypotheses. The above experiments demonstrate that, after an intuitive and simple refinement, the external information from the open data corpus is beneficial for improving the overall mapping performance.

5.2.3. Comparing Two External Corpora. In Section 5.1.1, we mentioned that educational knowledge concepts might be different from research knowledge concepts; therefore, we think that a textbook-based data corpus may perform better in book subsection linking than an academic conference data corpus. To examine this hypothesis, we selected and compared the best-performing approaches of both two data corpora in Table 3 and 4. The new results are provided in Table 5, where we find that the textbook-based external data corpus achieves the best result (though most of them are not statistically significant) among all three methods, particularly for NDCG@3. To some extent, this finding validates our proposed assumption. Although the academic publication-based external corpus does not show a significantly worse performance in most cases, achieving the best performance often requires extensive data selection and filtering. This process may not be easy to generalize to different domains and disciplines.

5.3. Qualitative Analysis of LDA Topics

In previous sections, we mentioned that the relatively worse performance of the concept-based approach and academic publication-based approaches could be due to the mismatch between educational knowledge components (mined from textbooks) and research knowledge components (extracted from research publications). To provide a further understanding of their differences, we conducted a qualitative analysis of the latent topics. Specifically, we applied the LDA to two data corpora - one compiled from five information retrieval textbooks and another that consisted of all SIGIR publications. We set the number of topics as 200 for both corpora because of its relatively better performance.

After manually examining the identified latent topics, we selected several aligned and misaligned knowledge components (see Figure 4). As the figure shows, there are many aligned knowledge components in both data corpora. For example, *Vector Space Model* is identified as a knowledge component in both data corpora with almost identical term distributions. However, more frequently, we encounter a mismatch between knowledge components in the two corpora. For instance, *Stemming* is an important knowledge concept in information retrieval, which is covered in all textbooks; however, the techniques related to stemming have been well developed for years, so little research is still being conducted on stemming in academia. Therefore, it is difficult to obtain such knowledge components from research publications. Similarly, several novel information retrieval research topics, such as *Expert Search*, have been developed in recent decades and are not yet included in textbooks. All these examples suggest the necessity of filtering misaligned knowledge components for textbook content linking.

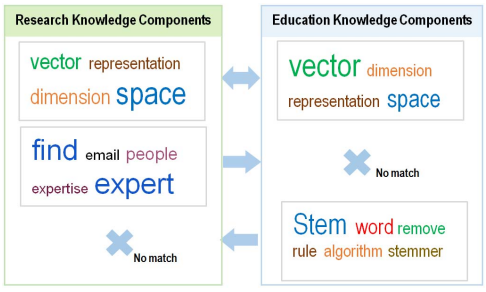

Figure 4. A comparison of the identified knowledge concepts from five textbooks (educational knowledge concepts; see the right figure) and all SIGIR publications (research knowledge concepts; see the left figure).

TABLE 4. KNOWLEDGE-BASED BOOK SUBSECTION LINKING PERFORMANCE WITH AN ACADEMIC PUBLICATION-BASED EXTERNAL DATA CORPUS. $*/\dagger$ DENOTES A SIGNIFICANT CHANGE OVER THE TERM-BASED/TERM+LDA) PERFORMANCE. NUMBERS IN BOLD (ITALIC) INDICATE THE BEST PERFORMANCE AMONG ALL METHODS (UNDER EACH CATEGORY). $\diamond$ INDICATES A SIGNIFICANT PERFORMANCE CHANGE OVER THE BEST-PERFORMING SUBGROUP IN THE TEMPORAL-BASED/CITATION-BASED DIVISION GROUPS.

	NDCG@1	**NDCG@3**
Term-based	0.3148(0.1184)	0.5498(0.1150)
Term + LDA	**0.3986(0.1327)**	0.5559(0.1176)
Term + LDA (+ all publications in SIGIR)	0.3818(0.1350) $*$	0.5299(0.1165) $\dagger$
Term + LDA (+ publications in SIGIR from year 1971 to 2001)	*0.3783(0.1285)* $*$	*0.5556(0.1212)*
Term + LDA (+ publications in SIGIR from year 2002 to 2008)	0.3583(0.1328) $*\dagger$	0.5242(0.1169) $*\dagger\diamond$
Term + LDA (+ publications in SIGIR from year 2009 to 2015)	0.3571(0.1204)$*\dagger\diamond$	0.5434(0.1210)
Term + LDA (+ publications in SIGIR with high #citations)	0.3396(0.1381) $\dagger$	0.5640(0.1206) $*$
Term + LDA (+ publications in SIGIR with medium #citations)	*0.3612(0.1306)* $*\dagger$	***0.5813(0.1127)*** $*\dagger$
Term + LDA (+ publications in SIGIR with low #citations)	0.2881(0.1102) $*\dagger\diamond$	0.4747(0.1140) $*\dagger\diamond$

TABLE 5. KNOWLEDGE-BASED SUBSECTION LINKING PERFORMANCE FOR EXTERNAL DATA CORPORA. $*$ DENOTES A SIGNIFICANT CHANGE OVER TERM + LDA (+ 3 EXTRA TEXTBOOKS). NUMBERS IN BOLD INDICATE THE BEST PERFORMANCE.

	NDCG@1	**NDCG@3**
Term + LDA (+ 3 extra textbooks)	**0.3878(0.1160)**	**0.5985(0.1115)**
Term + LDA (+ SIGIR 1971-2001)	0.3783(0.1285)	0.5556(0.1212) $*$
Term + LDA (+ SIGIR medium #citations)	0.3612(0.1306)	0.5813(0.1127)

5.4. Applying to a New Domain

The above experimental results are exclusively based on the data corpus obtained from the information retrieval domain. To demonstrate the ability of our approach to be generalized, we repeat the above experiments in a new domain - Elementary Algebra. We chose this domain because it greatly differs from information retrieval – knowledge concepts of this domain are more stable, as they have been well-established for more than two hundred years, and there are few modern academic publications in this domain. Specifically, we are interested in comparing the following book content linking approaches.

Compared methods. In this section, we want to understand the utility of the LDA-based semantic approach (i.e., LDA-based), the fusion of LDA and term-based methods (i.e., Term + LDA), and the incorporation of a textbook-based external corpus (i.e., Term + LDA with extra textbooks). Note that we do not include the concept-related approaches (i.e., concept-based and Term + Concept) and do not examine the utility of an academic publication-based external data corpus, because such information is not easily accessible in the Elementary Algebra domain. We do think that the concept information and external data corpus can be obtained through alternative resources, such as Wikipedia entries and articles, which is one of our future directions. To perform these experiments, we first need to build a data corpus with a ground-truth mapping of book subsections.

Textbooks. Similar to the information-retrieval domain, we consider five books for *Elementary Algebra*, which includes:

- *Elementary Algebra* by Ellis & Burzynski
- *Elementary Algebra* - v1 by Redden
- *Understanding Algebra* by Brennan
- *Fundamentals of Mathematics* by Burzynski & Ellis
- *Elementary Algebra Textbook*, second edition by Department of Mathematics, College of the Redwoods

The first two books were used for building a ground-truth, while the remaining three were treated as the textbook-based external data corpus. Similar to the pre-processing for information retrieval books, we also performed a removal of stop words and an exclusion of numbers, but did not perform stemming. However, we did not remove the non-letter symbols, such as $\cdot$, $\div$, $\neq$, $\geq$,$\sqrt{b^2-4ac}$, because they are also important basic concepts for elementary algebra.

Ground-truth. The ground-truth for book subsection linking is constructed in the same manner as in information retrieval books. We asked experts to manually label subsection mappings for the first two books. In total, we acquired five chapters with 92 subsections mapping from $BOOK1$ to 134 subsections in $BOOK2$.

The experimental results from the compared methods are provided in Table 6. The results are consistent with the findings in the information retrieval domain. Specifically, we found that the pure LDA-based approach performs equivalently or better than the term-based approach, particularly for NDCG@3, while the fusion of terms and LDA-based semantic topics outperform each single approach. Again, incorporating three extra books achieves the best performance among all four approaches. These findings clearly suggest the validity and ability of our ideas to be generalized across at least two domains.

TABLE 6. KNOWLEDGE-BASED BOOK SUBSECTION LINKING PERFORMANCE FOR ALGEBRA. $*/\dagger/\diamond$ DENOTES A SIGNIFICANT PERFORMANCE CHANGE OVER THE TERM-BASED/LDA-BASED/TERM + LDA APPROACH. NUMBERS IN BOLD INDICATE THE BEST PERFORMANCE.

	NDCG@1	**NDCG@3**
Term-based	0.4326(0.0913)	0.5402(0.0849)
LDA	0.4274(0.0997)	0.5674(0.0842)$*$
Term+LDA	0.4805(0.0931) $*\dagger$	0.5707(0.0783)$*\dagger$
Term+LDA (+3 extra textbooks)	**0.5558(0.0988)**$*\dagger\diamond$	**0.6328(0.0820)**$*\dagger\diamond$

6. Conclusion and Future Work

The rapid growth of online educational resources encourages researchers to focus on better modeling and linking approaches for a wide variety of educational content, which can be helpful in recommending education materials, recognizing important concepts, and other learning challenges. These challenges motivate us to examine semantic-based approaches for educational content modeling and linking, as well as ways of combining these approaches with term-based methods. To demonstrate the utility of different content modeling approaches, we employ the evaluation framework proposed by Guerra et al. [19], in which the modeling performance is examined by its contribution on linking subsections among different textbooks.

Experimental results, evaluated on an expert-crafted ground-truth data on *Information Retrieval* books, demonstrate the utilities of semantic approaches for knowledge content modeling. Both the purely semantic approach and the fusion of semantic- and term-based approaches show their value toward improving book subsection linking performance. Also, the semantic approaches can be further improved through incorporating appropriate external data resources. We find that textbooks are often used for educational purposes and focus on educational knowledge concepts, which may provide better support for textbook linking. This is demonstrated by a relatively better performance of textbook-based external resources over the academic publication-based external resources. A later study with ground-truth data from the *Elementary Algebra* domain confirmed our findings and pointed to its overall ability to be generalized to other domains.

In the future, we will explore the following topics. First, this paper focuses only on modeling and connecting knowledge components across different textbooks and ignores the hierarchical relations among knowledge components. For example, relationships between two subsections often suggest the relevance between parent and sibling sections. Similarly, the sequential order of textbook subsections often indicates prerequisite relations among them. However, such information is not properly modeled in the current study.

Second, this study demonstrates the utility of a semantic representation of educational content and the value of its fusion with term-based knowledge representation for linking online textbooks. However, their effectiveness may be limited by the amount of available textbook sources. Some recent semantic modeling methods [22] may provide a better representation of educational knowledge and lead to further improvement of knowledge-linking performance.

Third, the Link-the-Wiki Track [23] produces a benchmark that focuses on link discovery between Wikipedia documents. This similar task gives us another opportunity to evaluate the effectiveness and general applicability of our approaches. Further integration of Wikipedia concepts and educational knowledge can be explored as well.

References

[1] M. R. Kibby and J. T. Hayes, *Towards intelligent hypertext*. Oxford: Intellect, 1989, pp. 164–171.

[2] D. Tudhope and C. Taylor, "Navigation via similarity: automatic linking based on semantic closeness," *Information Processing & Management*, vol. 33, no. 2, pp. 233–242, 1997.

[3] O. Kolak and B. N. Schilit, "Generating links by mining quotations," in *The 19th ACM Conference on Hypertext & Hypermedia*, 2008, pp. 117–126.

[4] L. Carr, W. Hall, S. Bechhofer, and C. Goble, "Conceptual linking: Ontology-based open hypermedia," in *10th International World Wide Web Conference*. ACM Press, pp. 334–342.

[5] D. Milne and I. H. Witten, "Learning to link with wikipedia," in *The 17th ACM conference on Conference on information and knowledge management*. ACM Press, pp. 509–518.

[6] P. Brusilovsky and E. Millán, "User models for adaptive hypermedia and adaptive educational systems," in *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Neidl, Eds. Springer-Verlag, 2007, pp. 3–53.

[7] N. Henze and W. Nejdl, "Adaptation in open corpus hypermedia," *International Journal of Artificial Intelligence in Education*, vol. 12, no. 4, pp. 325–350.

[8] K. A. Papanikolaou, M. Grigoriadou, H. Kornilakis, and G. D. Magoulas, "Personalising the interaction in a web-based educational hypermedia system: the case of inspire," *User Modeling and User Adapted Interaction*, vol. 13, no. 3, pp. 213–267, 2003.

[9] P. Brusilovsky and C. Peylo, "Adaptive and intelligent web-based educational systems," *International Journal of Artificial Intelligence in Education*, vol. 13, no. 2-4, pp. 159–172, 2003.

[10] P. Brusilovsky and N. Henze, "Open corpus adaptive educational hypermedia," in *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Neidl, Eds. Berlin Heidelberg New York: Springer-Verlag, 2007, pp. 671–696.

[11] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[12] X. Wei 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*. ACM, 2006, pp. 178–185.

[13] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[14] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-based personalized search and browsing," *Web Intelligence and Agent Systems: An international Journal*, vol. 1, no. 3, 4, pp. 219–234, 2003.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[16] K. R. Canini, L. Shi, and T. L. Griffiths, "Online inference of topics with latent dirichlet allocation," in *International conference on artificial intelligence and statistics*, 2009, pp. 65–72.

[17] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 487–494.

[18] P. Brusilovsky, E. Schwarz, and G. Weber, "Electronic textbooks on the world wide web: From static hypertext to interactivity and adaptivity," *Web based instruction*, pp. 255–261, 1997.

[19] J. Guerra, S. Sosnovsky, and P. Brusilovsky, "When one textbook is not enough: Linking multiple textbooks using probabilistic topic models," in *Proc. of 8th European Conference on Technology Enhanced Learning (EC-TEL 2013)*, D. Hernandez-Leo, T. Ley, R. Klamma, and A. Harrer, Eds., 2013, pp. 125–138.

[20] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine learning*, vol. 95, no. 3, pp. 423–469, 2014.

[21] C. Wongchokprasitti, J. Peltonen, T. Ruotsalo, P. Bandyopadhyay, G. Jacucci, and P. Brusilovsky, "User model in a box: Cross-system user model transfer for resolving cold start problems," in *Proceedings of the 23nd Conference on User Modeling, Adaptation and Personalization*. Springer Verlag, 2015, pp. 289–301.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[23] A. Trotman, D. Alexander, and S. Geva, "Overview of the inex 2010 link the wiki track," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, 2010, pp. 241–249.