When One Textbook is not Enough: Linking Multiple Textbooks Using Probabilistic Topic Models

Julio Guerra[1 2], Sergey Sosnovsky[3], and Peter Brusilovsky[1]

[1] University of Pittsburgh, 135 North Bellefield Ave., Pittsburgh, PA. 15260, USA,
jdg60@pitt.edu,peterb@pitt.edu
[2] Universidad Austral de Chile, Independencia 641,Valdivia, Chile
[3] CeLTech, DFKI, Campus D3.2, D-66123 Saarbrücken, Germany
sosnovsky@gmail.com

Abstract. The Web-revolution in publishing and reading is rapidly increasing the volume of online textbooks. Nowadays, for most of the subjects, a selection of online textbooks is available. Such an abundance leads to an interesting opportunity: if a student does not like how a primary textbook presents a particular topic s/he can always access its alternative (e.g. more detailed or advanced) presentation elsewhere. Modern e-learning environments could better support access to different versions of instructional material by generating intelligent links between the textbooks sections that present similar topics and concepts. This paper reports an attempt to investigate the problem of fine-grained intelligent linking of online textbooks based on the probabilistic topic modeling technology. Using collections of textbooks in two domains (Elementary Algebra and Information Retrieval), we have demonstrated that intelligent linking based on probabilistic topic models produces a much better modeling quality than traditional term-based approaches.

Keywords: Hypermedia, Textbooks, LDA, Topic Model, Document Linking

1 Introduction

The vast amount of learning content available on the Web opens several opportunities for adaptive educational systems supporting learners in finding the "right content". Practically, in any domain, a learner can easily find online dozens, hundreds or even thousands of instructional texts, tutorials and electronic textbooks. However, open-corpus resources are heterogeneous, they may organize their content in very different ways: by focusing on different parts of domain knowledge and covering the same parts with different levels of details, by using different terms for the same concepts (synonymy) and the same terms for different concepts (polysemy), by aggregating and structuring domain knowledge with different sets (or hierarchies) of chapters, sections, pages, etc. [9]

The abundance of online textbooks results in interesting opportunities for modern students: those not satisfied with presentation of a particular topic in their primary textbook have a chance to access an alternative presentation from a range of textbooks covering this topic. To turn this opportunity into practice, modern e-learning environments should support their users by generating dynamic links between sections of different books that present similar topics and concepts.

For example, imagine a learner reading a section about "Linear Equations in Two Variables" from an Algebra textbook. The learner has problems understanding some parts of this section and decides to search for other textbooks covering similar content in her/his e-learning system. Since the system has access to multiple educational resources in this domain and understands semantic relations between their parts and sections, it returned a ranked list of links to relevant content of the other textbooks: a chapter titled "Linear equations (part II)" in one book, a section titled "Solving linear systems of equations" in another book, a subsection in the same book titled "Graphical solving of linear equations in two variables", and a chapter titled "Solving equations" in a third book.

The idea of automatic links generation between related fragments of educational content has been pioneered by Mayes and Kibby [7]. The keyword-level similarity-based approach suggested by them has been applied in a number of systems and architectures since[5]. Unfortunately, the practical quality of simple keyword-level similarity linking appeared to be less than perfect, as it can often link pages that were not really meaningfully related, which would result in presenting a wrong page to the user. While similar technologies have been often satisfactory in other domains, low quality of linking is unacceptable for e-Learning systems. As a category of users, students are very susceptible to the system's failures and are rarely able to recover from those on their own. Erroneous instructional material presented to a student can lead to confusion, frustration, lack of motivation and, essentially, poor learning [6].

Another area of research that has addressed the problem of linking documents based on their meaning is Semantic Annotation. Several systems have been built capable of enriching online content with dynamic links to relevant documents (e.g. COHSE [1] and Magpie [4]). Such systems are very good in recognizing special named entities (such as names, places, organizations as well as quantitates, dates, times etc.), and detecting keywords that match concepts names in a source ontology or a thesaurus; unfortunately, they disregard the rest of the content. This is an effective approach for general information access, but it is ill-suited for the educational domain. Instructional content rarely contains specially formatted entities; and one cannot expect that similar documents in two different textbooks from an arbitrary subject will be always using the same keywords.

This paper reports an attempt to re-investigate the problem of fine-grained intelligent linking of online textbooks using a popular approach to probabilistic topic modeling known as Latent Dirichlet Allocation (LDA). LDA [2] is a statistical model that automatically extracts topics from a collection of doc-

uments. Over the last few years, this technique has been applied successfully for discovering semantic models in large, heterogeneous and unstructured (or lightly-structured) collections, such as scientific journal papers or collections of news posts [2, 3]. However, there is no known attempt of using this technique for modeling hierarchical structured collections, such as textbooks for the task of document linking. Our challenge has been to explore whether LDA can be applied within domain-specific collections of hierarchically structured educational content and can successfully support intelligent linking of online textbooks.

A study presented in this paper uses a collection of textbooks in two domains (Elementary Algebra and Information Retrieval) to explore whether intelligent linking based on probabilistic topic models can achieve a better quality of section-level textbook linking than previously used term-based approaches. To maximize the quality of the new technology, we also explore some important parameters associated with the application of LDA-based approach in hierarchical textbook context and report the performance results.

2 Linking Documents Using LDA

LDA topic models are generative probabilistic models that conceive each document as a mixture of topics and each topic as a mixture of words. LDA represents documents by a probability distribution over the topic space and defines topics as a probability distribution over the vocabulary of the collection. For building the model, a collection of documents is inputed to LDA. The algorithm iteratively assigns a topic to each word in each document and adjust the topic-word probabilities. The process can be seen as an optimization problem, in which the LDA mechanism continuously minimizes the number of "highly probable" topics per document and, at the same time, minimizes the number of "highly-probable" words per topic [2, 10]. This leads to topics incorporating words that often occur together. LDA receives as inputs the number of topics to discover and the two prior (hyper) parameters to control both the sparsity of topic distributions in the documents and the sparsity of word distributions in the topics (the LDA setup for this specific project is explained in more details in Section 4.3). Once the model has been built, each document is represented as a probability vector over all the topics, and each topic as a probability vector over all the words in the vocabulary. Documents are discriminated based on different concentrations of topic probabilities. Taken the topic representations of the documents, the homogeneity between them can be computed using distance or similarity measures [10]. In our case, once we have a vector representation of all chapters, sections and subsections of the textbooks, we can proceed to compute similarity among parts of different books and, for each book part, create a ranking of similar other-books' parts. As we will see in the next section, there are some issues regarding the hierarchical structure of the textbooks to consider in order to obtain meaningful topic distributions of all book parts.

3 Research Questions

The main goal of this work is to explore the use of probabilistic topic models for generating high-accuracy linking of textbook parts (chapters, sections and subsections). The LDA-produced topic models are examined against the standard term-based models. The first research question is:

Q1 *Will probabilistic topic modeling produce a more accurate linking of textbook parts than the common term-based approach?*

The key aspects to consider are the characteristics of textbooks. Textbooks are hierarchically organized and the content of each chapter, section or subsection is the aggregation of its children's contents. The assignment of topic distribution among parts of the textbooks on different levels should consider these characteristics for a correct representation of each document's content. This leads to the second research question:

Q2 *How to incorporate hierarchical structures of textbooks in probabilistic topic models?*

We need a topic model to correctly represent each document (each textbook part) as the aggregation of the content of its sub-parts. One option is to build the topic model from all textbook parts with aggregated content, for example, a section aggregates the content of the subsections and in this form is input to LDA. However, this approach seems to "confuse" LDA as demonstrated by results of our preliminary tests. Other, more reasonable approach is to consider building the model using only documents that have actual (textual) content (usually terminal or leaf nodes, i.e. sub-sections) and further options for indexing the intermediate nodes (i.e computing the topic distribution of chapters and sections). We consider two approaches:

a) Aggregate topic distributions along the hierarchical structure of the textbook while weighting sub-topics' distributions by their sizes. For example, the topic distribution of a section is the aggregation of the topic distributions of it's subsections. The following formula explains the aggregation of topic distributions for a specific book part d.

$$\Psi'_d = \frac{\sum_{c \in C_d} \left(s_c \Psi_c\right) + s_d \Psi_d}{\sum_{c \in C_d} \left(s_c\right) + s_d} \tag{1}$$

Ψ_d and Ψ'_d are the topic distributions of the node d before and after the aggregation, respectively; C_d is the set of child nodes of d, Ψ_c is the topic distribution of the child c, s_c and s_d are the sizes measured as the number of words in the respective document. For simplicity, hereafter, we refer to this option as *Topic Aggregation* (TA).

b) Re-index the aggregated content of intermediate documents using the inference mechanism provided by the topic model. This means, after the model is built, a version of the collection is created, in which each chapter and section contains the text content of all its children nodes. Then, each aggregated document is inputed to the built LDA model to obtain its new topic distribution. From now on, we refer to this approach as document *Re-Indexing* (RI).

Additionally, since individual textbook models reflect corresponding individual views on the domain stemming from differences in terminology that different authors may use, we are interested to explore the potential added value of building unified topics model using several books. We expect that a model built using multiple textbooks will reflect a better (more complete and objective) understanding of the domain and will support better document linking than a model built using a single textbook. For simplicity, we further call these options *Multiple Book* (MB) and *Single Book* (SB), respectively. Thus, the third research question is:

Q3 *Will the model built using multiple textbooks (MB) produce a better document linking compared to a model built using a single textbook (SB)?*

4 Experiment Design

We have conducted several experiments building LDA from textbooks in two domains: Elementary Algebra and Information Retrieval. Resulting topic models are used for computing similarity between the parts of different books in each domain. The evaluation approach compares the list of similar documents found by the LDA models and the baseline term-based model with an "ideal" document mapping provided by experts.

4.1 Textbooks

Four Algebra and five Information Retrieval textbooks were used in the study. For further references, the textbooks are labeled as BOOK1, BOOK2, etc. For the both domains, topic models in the Single Book condition (SB) are built based on BOOK1, and BOOK2 is used for evaluation. For Algebra, BOOKs 1, 3 and 4 are used for building topic models in the Multiple Book condition (MB); and, in the Information Retrieval domain, BOOK5 is additionally used for this. A sequence of technical procedures was performed on the contents of the textbooks: the text was converted to lowercase, stop-words and additional frequent words in the domain (i.e.: "exercises", "solutions" in Algebra) were removed.

Elementary Algebra Textbooks

1: Elementary Algebra, by W. Ellis & D. Burzynski.

2: Elementary Algebra - v1, by J. Redden.
3: Understanding Algebra, by J. Brennan.
4: Fundamentals of Mathematics, edited by D. Burzynski & W. Ellis.

Information Retrieval Textbooks

1: Introduction to Information Retrieval, by C. Manning, P. Raghavan & H. Schütze.
2: Modern Information Retrieval, by R. Baeza-Yates and B. Ribeiro-Neto.
3: Finding Out About, by R. Belew.
4: Information Storage and Retrieval Systems, by G. Kowalski.
5: Information Retrieval, by C. van Rijsbergen.

4.2 Ground Truth

As the ground truth, we used manual mappings of the two textbooks (BOOK1 and BOOK2) made by groups of experts in both domains. Overall, ten experts contributed: one professor and six PhD students from the School of Information Sciences at the University of Pittsburgh; three researchers from the Centre for e-Learning Technologies at DFKI. In the Algebra domain, five chapters of BOOK1 were mapped to BOOK2. In the Information Retrieval domain, four chapters of BOOK1 were mapped to BOOK2. To obtain more objective judgements, for each chapter, two experts were providing the mapping. A dedicated Web-interface was developed to facilitate the mapping task. Specific instructions were given in order to have a consistent mapping results: i) every chapter, section and subsection of BOOK1 had to be mapped to zero or more parts from BOOK2; ii) mapping had to be as accurate as possible; iii) mapping could relate book parts from different levels (for example, sections could be mapped to chapters, sections, or subsections); iv) it had to be taken into account that the content of a textbook part is the aggregation of the content of its subparts. Additionally, experts had to assign to every mapping a level of relevance and a level of confidence, both ranging from 1 to 3 (low, medium, high). A score for each match was computed by multiplying the relevance and confidence levels. The final score of each match was computed as the sum of the scores provided by both experts. Non-matched parts of BOOK2 were assigned with the zero score. Finally, the ground truth was represented as the compilation of all mappings blended into a single list: each element is a part of the BOOK1 and the respective ranked list of all matches from BOOK2 with their final scores.

4.3 LDA Setup

We used implementation of LDA provided by the MALLET Toolkit [8]. In MALLET, LDA setup depends on the number of topics, the number of sampling iterations, the smoothing over document-topic distribution hyper-parameter α, and the smoothing over topic-word distribution hyper-parameter β (a good explanation of the LDA hyper-parameters can be found in [10]). We set the number

of iterations to 2000, taking into account the size of the documents and the collections. For selecting the number of topics, we followed a simple approach: since we expect the topics to represent semantic units of textbook content, we estimated that the number of topics should be between the number of sections and the number of subsections in an average book (sometimes, sections covers several topics, and, sometimes, subsections cover examples, exercises or different views of the same topic). In the Algebra domain, BOOK1 has 74 sections and 228 subsections; thus, we chose the number of topics equal to 150. This number also gave us the best results in the preliminary tests. In the Information Retrieval domain, the BOOK1 has 120 sections and 178 terminal nodes (some sections do not have children and are counted also as subsections). Here we also chose 150 as the number of topics. With regards to LDA hyper-parameters, we set up initial values of $\alpha = 0.01$ and $\beta = 0.01$, and then used the fixed-point optimization for hyper-parameters [11] implemented in MALLET.

4.4 Measuring the Effectiveness of Document Linking

We evaluate the effectiveness of all compared models on the task of finding documents similar to a target document using average NDCG@1, @3 and @10[4] as follows:

1. for every part of BOOK1, each part of BOOK2 is ranked according to the similarity of their topic distributions. As a similarity measure, we used the reciprocal symmetric KL-divergence[5];
2. for every part of BOOK1, the ranked lists generated by evaluated models are compared against the respective rank list from the ground truth using NDCG@1, @3 and @10;
3. for every LDA condition, the average NDCG@1, @3 and @10 are computed over all parts of BOOK1;
4. the LDA sampling process uses random seeds and produces different topic sets on each run; for every LDA condition, we run the model 30 times, thus, obtaining N=30 data points.

Baseline. As the baseline we use the effectiveness (measured by NDCG@1, @3 and @10) of the ranked lists obtained as a result of querying an index built using Apache Lucene[6]. Lucene computes similarity between a query and the

[4] NDCG (normalized discounted cumulative gain) measures the quality of ranking documents by relevance. It compares the target ranking to the positions that documents occupy in the ideal list and penalizes the mismatches. NDCG@1 measures the effectiveness of the model to find the top relevant document. In the same way, NDCG@3 and NDCG@10 measure the quality of ranking the first three and ten items, respectively.

[5] Kullback-Leibler (KL-) divergence computes the difference between two probability distributions. The reciprocal symmetric KL-divergence is a modification of the original formula that can be used as a similarity measure [10].

[6] http://lucene.apache.org

documents in the collection using the standard TF-IDF model. For each part of the BOOK1, we toss a query to the index built from BOOK2. The query returns a ranked list of the parts in the BOOK2, which is used to compute the baseline average values for NDCG@1, @3 and @10.

4.5 Conditions and Hypotheses

Q1 examines the idea of contrasting LDA-based models with the baseline. Q2 is focused on comparing two different strategies helping to incorporating textbook hierarchies in the topic model: *topic aggregation* (TA), and *re-indexing* (RI). We further refer to this as the *aggregation* factor. Q3 stresses the comparison between a model built using a *single book* (SB) with a model built using *multiple books* (MB) from the collection. We further refer to these options as the *books* factor. Thus, four conditions are defined as can be seen in Table 1. KL-Divergence is used as the similarity measure; it helps to produce ranked lists of links from the nodes of BOOK1 to the documents of BOOK2. Ranked list are evaluated against the ground truth using averaged values of NDCG@1, @3 and @10. Finally, the LDA models are sampled 30 times to produce 30 data points for every condition. The experiment is executed separately in both domains.

Table 1. The 4 model conditions.

		Aggregation	
		Topic Aggregation (TA)	Re-Indexing (RI)
Books	Single Book (SB)	SB-TA	SB-RI
	Multiple Books (MB)	MB-TA	MB-RI

To address Q1, the four conditions are compared to the baseline values of NDCG@1, @3 and @10. Here, we define the following hypothesis:

H1 *At least one of the models built using LDA performs better document linking than the baseline.*

Since the baseline accuracy score is a fixed value, we use one-sample t-tests to verify H1. To answer Q2 and Q3, we state the following hypotheses:

H2 *The LDA model aggregating topic probabilities based on the book hierarchy (TA) will perform better document linking than the model using re-indexing of content-aggregated intermediate documents (RI).*

H3 *The LDA model built using several textbooks (MB) will perform better document linking than the model built using a single textbook (SB).*

For testing H2 and H3, the four groups are compared based on the 2x2 between-subjects ANOVA design (see Table 1). Interactions, main effect and marginal means are reported.

5 Results

5.1 LDA vs. Baseline

Mean values and standard deviations of NDCG levels for all LDA conditions together with the results of t-tests comparing them against the baseline are reported in Table 2. Figure 1 presents these data graphically.

Table 2. Effectivity of the four conditions in the Algebra domain (top) and Information Retrieval domain (bottom).

Algebra

	NDCG@1			NDCG@3			NDCG@10		
Baseline	*.3662*			*.5807*			*.6582*		
	Mean	Std. Dev.	Sig. (p)	Mean	Std. Dev.	Sig. (p)	Mean	Std. Dev.	Sig. (p)
SB-TA	.547	.025	<.001	.647	.018	<.001	.691	.015	<.001
MB-TA	.532	.036	<.001	.620	.021	<.001	.663	.017	.165
SB-RI	.456	.027	<.001	.601	.027	<.001	.675	.019	<.001
MB-RI	.414	.040	<.001	.572	.032	.132	.647	.026	.022

Information Retrieval

	NDCG@1			NDCG@3			NDCG@10		
Baseline	*.057*			*.186*			*.258*		
	Mean	Std. Dev.	Sig. (p)	Mean	Std. Dev.	Sig. (p)	Mean	Std. Dev.	Sig. (p)
SB-TA	.345	.051	<.001	.461	.042	<.001	.536	.033	<.001
MB-TA	.309	.063	<.001	.418	.045	<.001	.520	.039	<.001
SB-RI	.360	.066	<.001	.484	.053	<.001	.556	.045	<.001
MB-RI	.336	.050	<.001	.456	.054	<.001	.534	.041	<.001

Algebra domain. As it is seen from Table 2, for Algebra domain, all LDA models for all NDCG levels produce significantly better results than the baseline except for NDCG@3 of MB-RI, NDCG@10 of MB-TA and NDCG@10 of MB-RI (which is even significantly lower than the respective baseline value). Among the significantly better LDA scores, the highest effect sizes were always presented by the SB-TA model outperforming the baseline by several standard deviations (NDCG@1: $p < .001$, Cohen's $d = 7.232$; NDCG@3: $p < .001$, Cohen's $d = 3.683$; NDCG@10: $p < .001$, Cohen's $d = 2.187$). Since for all NDCG levels there was at least one condition that performs significantly better than the baseline, these results support H1 in the Algebra domain.

Information Retrieval domain. Within the Information Retrieval collection, all LDA conditions perform significantly better than the baseline (see the bottom part in Table 2).

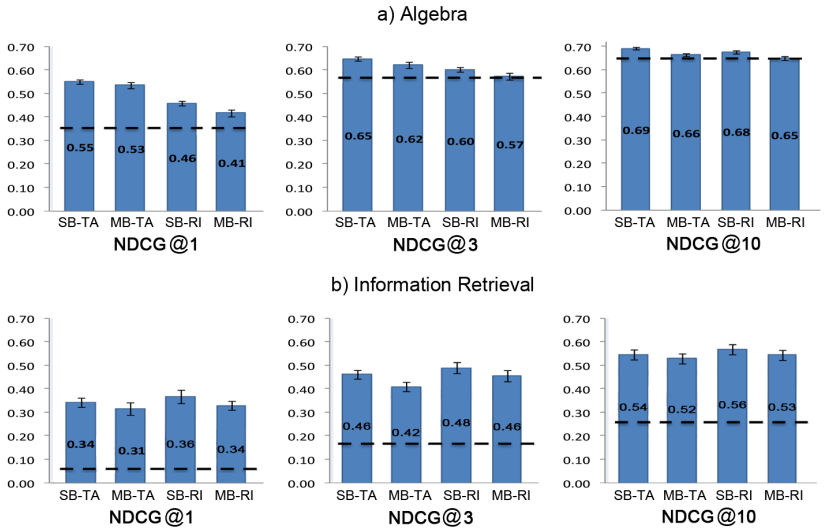

Fig. 1. Four conditions individually compared with the baseline in Algebra and Information Retrieval domains. Baseline NDCG@1,@3 and @10 values are indicated with dotted lines.

H1 is supported. Overall, LDA produces significantly better document linking than the baseline term-based model in both domains. The highest results are always obtained for NDCG@1, which means that the LDA-based approach performs much better in finding the best matches, i.e the most similar documents. It is also very interesting to note the difference between the two domains. Information Retrieval is a newer domain than Algebra and, thus, has a less standardized vocabulary. Therefore, LDA is especially effective compared to the standard term-based approach. This result confirms the power of probabilistic topic modeling in domains with sparse and non-standardized vocabularies.

5.2 TA vs. RI

For testing H2, in each domain, a 2x2 between-subjects ANOVA was performed on the scores of NDCG@1, NDCG@3, and NDCG@10 as a function of *aggregation* strategy (TA, RI) and *books* strategy (SB, MB). Table 3 presents the marginal means and standard error values for all conditions and domains.

Algebra domain. All ANOVA assumptions have been satisfied for all conditions and all NDCG levels.

The patterns of differences among *aggregation* strategies are significantly different between SB and MB only for NDCG@1, $F(1,116) = 5.274$, $p = .023$, $\eta_p^2 = .043$. No other significant interactions have been observed. For the main effect of *aggregation*, the results show that TA produces significantly higher scores than

Table 3. Marginal Means.

		NDCG@1		NDCG@3)		NDCG@10	
Algebra		M	SE	M	SE	M	SE
Aggregation	TA	.539	.004	.633	.003	.677	.003
	RI	.435	.004	.587	.003	.661	.003
Books	SB	.502	.004	.624	.003	.683	.003
	MB	.473	.004	.596	.003	.655	.003
Information Retrieval		M	SE	M	SE	M	SE
Aggregation	TA	.327	.007	.439	.006	.528	.005
	RI	.348	.007	.470	.006	.545	.005
Books	SB	.352	.007	.473	.006	.546	.005
	MB	.323	.007	.437	.006	.527	.005

RI for all NDCG levels averaged across *books* strategy (NDCG@1: $F(1,116) = 309.496$, $p < .001$, $\eta_p^2 = .727$; NDCG@3: $F(1,116) = 103.181$, $p < .001$, $\eta_p^2 = .471$; NDCG@10: $F(1,116) = 19.332$, $p < .001$, $\eta_p^2 = .143$). These results support H2 in the Algebra domain. Additionally, it can be seen that the effect decreases with every next NDCG level. This underlines the important of choosing the right *aggregation* strategy when the goal is to find the most relevant documents from another textbook.

In order to find the pattern of differences in NDCG@1 scores among *books* and *aggregation* strategies, a simple main effect of *aggregation* has been performed for SB and MB strategies. When LDA models are build using a single book, TA (M=.547, SE=.006) is significantly better than RI (M=.456, SE=.006): $F(1,116) = 116.983$, $p < .001$, $\eta_p^2 = .502$. When multiple books are used, TA (M=.532, SE=.006) also significantly outperforms RI (M=.414, SE=.006): $F(1,116) = 197.787$, $p < .001$, $\eta_p^2 = .630$. These results showed that the interaction between *aggregation* and *books* factors for NDCG@1 does not change the fact that TA works better than RI in the Algebra domain; it only means that, when using TA, there is no difference between SB and MB conditions.

Information Retrieval Domain. All ANOVA assumptions have been satisfied for all conditions and all NDCG levels, except for a slight deviation from normality of NDCG@3 under MB-RI condition (Shapiro-Wilk = .929, p = .047).

There is no significant difference in the patterns of differences among *aggregation* strategies between *book* strategies for all NDCG levels. The main effect of *aggregation* has been observed for all NDCG levels. However, it is the opposite of what we have seen in the Algebra domain. For Information Retrieval, TA produces significantly lower scores than RI (NDCG@1: $F(1,116) = 4.017$, $p = .047$, $\eta_p^2 = .033$; NDCG@3: $F(1,116) = 11.903$, $p = .001$, $\eta_p^2 = .093$; NDCG@10: $F(1,116) = 5.667$, $p = .019$, $\eta_p^2 = .047$). These results do not support H2 in the Information Retrieval domain.

H2 is partially supported. In the Algebra domain, Topic Aggregation strategy performs better than Re-Indexing. However, the results are opposite in the Information Retrieval domain, where RI performs better than TA. This is interesting, as domain differences seem to matter when choosing the best *aggregation* strategy. One possible explanation can be inferred from the different characteristics of the two domains and the differences between the used textbooks. For example, the vocabularies that LDA models are processing in these domains are very different in size. In Algebra, a single book produces a vocabulary of 2,220 terms, while in Information Retrieval, a single book model works with the vocabulary of 13,405 terms. Further research is needed in order to better understand the differences among the domains and how they influence the topic modeling process.

5.3 MB vs. SB

In this section, we compare the performance of LDA models built from a Single Book (SB) with models built from Multiple Books (MB) in order to test H3. The ANOVA design described in previous sections is also used here as well.

Algebra domain. In the Algebra domain, there are no significant interactions among *books* and *aggregation* factors except for NDCG@1. The simple main effect test reported in the previous section has shown that it does not matter, whether the LDA model is built from a single book or multiple books, when it uses TA as the *aggregation* strategy. For the main effect of *books*, the results show that, in the Algebra domain, SB produces significantly higher scores than MB averaged across *aggregation* strategies for all NDCG levels (NDCG@1: $F(1,116) = 24.180$, $p < .001$, $\eta_p^2 = .172$; NDCG@3: $F(1,116) = 38.361$, $p < .001$, $\eta_p^2 = .249$; NDCG@10: $F(1,116) = 62.095$, $p < .001$, $\eta_p^2 = .349$). Thus, H3 is not supported in the Algebra domain.

Information Retrieval domain. Similar results are obtained in the Information Retrieval domain. There is significant difference between SB and MB strategies averaged across *aggregation* for all NDCG levels (NDCG@1: $F(1,116) = 7.849$, $p = .006$, $\eta_p^2 = .063$; NDCG@3: $F(1,116) = 16.099$, $p < .001$, $\eta_p^2 = .122$; NDCG@10: $F(1,116) = 6.871$, $p = .010$, $\eta_p^2 = .056$). All these effects strongly favor using the Single Book strategy over Multiple Books and do not support H3 in Information Retrieval domain as well.

H3 is not supported. In both, Algebra and Information Retrieval domains, the results have shown that LDA models built using a Single Book (SB) will produce significantly better document linking than the models built from Multiple Books (MB) regardless of the hierarchical content *aggregation* method.

6 Conclusions

In this work, we have explored the use of LDA-based topic models within collections of textbooks for the task of fine-grained cross-collection document linking. We have shown that LDA is a valuable alternative to the standard term-based approach and outperforms it, especially, for finding the most similar documents (NDCG@1, NDCG@3) and in less standardized domains. We have applied two different approaches for building topic models with regards to the hierarchical structure of textbooks and discovered that, in different domains (and, perhaps, content collections), different aggregation strategies can be better. We have also observed that the topic models built using one textbook produce a better document linking than the model built using multiple books, contrary to our expectations. Further research is needed to improve our understanding of these differences and their impact on the LDA model building.

We believe, that using LDA is a promising approach for addressing the problem of automatic and fine-grained textbook linking and it can be further applied to facilitate content modeling and adaptation in the open corpus settings. The ability of probabilistic topic-based models to help finding the top similar documents is a clear advantage over the traditional term-based methods and can be used to implement effective recommendation and navigation support in adaptive educational hypermedia systems. We plan to incorporate this technology in an e-learning environment, as well as to keep investigating mechanisms for improving the quality of LDA models. Our future work will include experimenting with the techniques for topic models evaluation, utilizing topics and textbooks structures to discover semantic relations among the topics, ensembling topic models with the models based on keyword and concept extraction techniques, and further investigating the application of this approach in other domains.

Acknowledgements. Julio Guerra is supported by Chilean Scholarship (Becas Chile) from the National Commission for Science Research and Technology (CONICYT, Chile), and the Universidad Austral de Chile.

References

1. Bechhofer, S., Goble, C., Carr, L., Kampa, S., Hall, W., & De Roure, D. (2003). COHSE: Conceptual Open Hypermedia Service. In S. Handschuh & S. Staab (Eds.), *Annotation for the Semantic Web* (pp. 193-211). IOS Press.
2. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.
3. Blei, D M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning, June 25-29, Pittsburgh, PA* (pp. 113-120).
4. Domingue, J., Dzbor, M., & Motta, E. (2004). Magpie: supporting browsing and navigation on the semantic web. In C. Rich & N. J. Nunes (Eds.), *Proceedings of 9th International Conference on Intelligent User Interfaces (IUI '04), Funchal, Madeira, Portugal* (pp. 191-197). New York, NY, USA: ACM Press.

5. Fountain, A., Hall, W., Heath, I., & Davis, H. (1992). MICROCOSM: an open model for hypermedia with dynamic linking. In: N. Streitz, A. Rizk & J. André (Eds.), *Hypertext: concepts, systems and applications* (pp. 298-311). Cambridge University Press.
6. Jednoralski, D., Melis, E., Sosnovsky, S., & Ullrich C. (2010). Gap Detection in Web-based Adaptive Educational Systems, In *Proceedings of the 9th International Conference on Web-based Learning (ICWL2010), Shanghai, China, Dec. 8-10, 2010* (pp. 111-120). Berlin/Heidelberg: Springer-Verlag.
7. Mayes, J. T., Kibby, M. R., & Watson, H. (1988). StrathTutor: The development and evaluation of a learning-by-browsing on the Macintosh. *Computers and Education 12 (1)*, 221-229.
8. McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. Available from http://mallet.cs.umass.edu
9. Sosnovsky, S., Hsiao, I-H., & Brusilovsky, P. (2012). Adaptation "in the Wild": Ontology-based Personalization of Open-Corpus Learning Material. In A. Ravenscroft, S. Lindstaedt, C. Delgado Kloos, & D. Hernández-Leo (Eds.), *Proceedings of EC-TEL 2012: Seventh European Conference on Technology Enhanced Learning* (pp. 425 - 431). Berlin/Heidelberg: Springer-Verlag.
10. Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Laurence Erlbaum.
11. Wallach, H. M. (2008). Structured Topic Models for Language. (Doctoral dissertation). University of Cambridge, UK.