# INFSCI 2910 - Independent Study: Foundations

(Fall Term 2012-2013)

# Report

Julio Guerra

---

# Intelligent Linking of Online Textbooks Using Probabilistic Topic Models


## ABSTRACT

With the fast grow of online educational content, the abundance of quality material opens new opportunities to learners. For practically any domain, a learner can easily find tens, hundreds or even thousand of web pages, tutorials and electronic textbooks in the Internet. Modern educational systems can grasp this opportunity offering alternative content to their users by automatic linking similar content parts. This is known as intelligent linking and has been widely explored in the hypertext field. However, well known techniques for intelligent linking reach only relatively medium and low results for fine-grained content matching when applied to hierarchical organized content such as textbooks. In this work we re-visit the problem of intelligent linking of textbooks using modern approaches of probabilistic topic modeling. Working with collections of textbooks in two domains (Algebra and Information Retrieval), we explored options to apply probabilistic topic models to implement fine-grained content matching and demonstrated that such techniques performs much better than the previously used term-based approach.



### General Terms

Algorithms, Measurement, Reliability, Experimentation.

### Keywords

Intelligent linking, probabilistic topic models, textbooks.


## 1. INTRODUCTION

The vast amount of learning content available in the Web opens several opportunities for adaptive educational systems supporting the learner to find the "right content". For practically any domain, a learner can easily find tens, hundreds or even thousand of web pages, tutorials and electronic textbooks in the Internet. The learners can benefit from this abundance of content: those who are not satisfied with the primary content source can switch to alternative resources and find better or more suitable content. However, open-corpus resources are heterogeneous, they may organize the content in very different ways, may cover different amounts of the content and with different levels of details, may use different terms for the same concepts (synonym), and the same terms for different concepts (polysemy). All of these issues make a challenge to educational systems wanting to grasp this content abundance to offer dynamic linking between sections of different content resources that presents similar topics and concepts. Picture the following scenario: imagine a learner reading a section about "Linear Equations in Two Variables" in an textbook about algebra. The learner has problems to understand part of the content and requested her e-learning system to suggest an alternative presentation from several online texbooks on the same subject known to the system. In response, the system returned a ranked list of links to relevant parts 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. Which one is exactly about the topic "linear equations in two variables"? Which one does a better covering of the content of the section the student read?

In the field of hypertext, intelligent and dynamic linking of content (pages) that have no static link between them has been researched extensively and architectures supporting these tasks were suggested and implemented.

The task is in the core of the conception of the Internet as an interlinked collections of content resources. However, well known techniques for intelligent linking reach only relatively medium and low results for fine-grained content matching when applied to hierarchical organized content such as textbooks. We believe that the quality problem of the intelligent linking approach can be addressed by switching from traditional term-based information retrieval approaches that stand behind the majority of known intelligent linking projects a more advanced probabilistic topic modeling technology that has become available over the last few years.

In this paper, we report our work in re-investigate the problem of fine-grained intelligent linking of online textbooks applying two versions of one of the most popular probabilistic topic modeling approach known as Latent Dirichlet Allocation (LDA). LDA [4] is a statistical model that can automatically discover topics from a collection of documents. A variant, Hierarchical LDA (HLDA) [5], is a model that discovers a hierarchy of topics in which lower level topics contains words which are more specific in the domain. One important advantage of such models is that they can deal with synonym and polysemy problems [4]. Over the last few years, these techniques has been applied successfully for discovering semantic structures in large, heterogeneous and unstructured or lightly-structured collections like scientific journal papers or collections of news posts. Our challenge was to explore whether these approaches can be applied successfully within domain-specific collections of hierarchical structured content for the task of document matching.

A study presented in this paper used a collection of textbooks in two domains, algebra and information retrieval, to explore whether intelligent linking based on 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 explored some important parameters associated with the application of topic-based approach in hierarchical textbook context and report the performance results.

## 2. RELATED WORK

### 2.1 Intelligent Linking

Starting from the early work of Mayes and Kibby [16; 22], traditional information retrieval techniques based on keyword-level similarity (see next section) have been applied for intelligent linking in a number of systems and architectures [12]. Unfortunately, the quality of such techniques for intelligent linking is less than perfect, specially when the task is to choose the most similar content alternatives. Keyword-based similarity often links pages that are not really similar. A second-generation research on intelligent linking explores how to improve the content matching. Some researchers focused on keyword-level techniques for generating typed links [3; 10] while others focused on improving the precision of keyword-level linking by incorporating "semantic-oriented" techniques such as latent semantic indexing [20], lexical chaining [14], and ontology-based linking [9]. In parallel and aiming to incorporating the "human in the loop", some researchers explored semi-automatic intelligent linking interfaces, such as map-based navigation with self-organized maps [8] and quotation-based navigation [17]. In this context, our work seeks to expand the "semantic" direction of intelligent linking research by exploring a more recent and more powerful LDA-based probabilistic topic modeling.

### 2.2 IR techniques for document matching

Text document matching has been widely explored by Information Retrieval (IR.) Different techniques such as the boolean model, the vector space model or the probabilistic language models has been used and combined to represent the content of the documents and match them to queries or other documents (see [21] or [2] for a comprehensive description on the field). One of the most used representations is the vector space model (VSM). In the VSM each document is represented with a vector over the space of the whole vocabulary of the collection, using measures like the *term frequency inverse document frequency* (TFIDF). In the VSM, the similarity between two document is often computed using Cosine Similarity, a measure of the angle between the vectors representations. Apache Lucene, a well known and used information retrieval software, uses a modified VSM - TFIDF model as described in [1].

More recently, a considerable amount of research has focused in improving retrieval tasks by indexing documents using dimentionality reduction techniques. Latent Semantic Indexing, or LSI [11] uses singular value decomposition (SVD) for extracting a set of features representing relations between terms in the collection, and

indexes the documents using these features. Probabilistic LSI (pLSI) is an improvement of LSI using better statistical foundations [15]. In pLSI each document is represented as a mixture of semantic features (also called aspects or topics). In [4], Blei argue that one of the limitation of pLSI is the lack of a well defined generative semantics disallowing the model to index documents outside the training collection, and to overcome this proposed Latent Direchlet Allocation (LDA), which is described in details in the next section.

In probabilistic approaches, other similarity measures than the cosine similarity have been proposed [19]. The Kullback Leibler Divergence (KL divergence) [18] is a common measure of difference between two probability distributions and can be used for similarity by a simple inverse or reciprocal transformation. For probabilistic topic models, such as LDA, Steyvers and Griffiths [27] mentioned several approches for compute similarity between topic representations including cosine and symmetric KL-Divergence.

### 2.3 LDA

LDA [4; and other good explanations in 27, 13, 7] is a probabilistic topic model built upon the assumption that every document is generate by mixture of several topics. In LDA, every document is represented as a distribution of probabilities over a set of topics, and every topic is a distribution of probabilities over all the terms in the vocabulary. In the model building process, the co-occurrence of words in the documents is considered in a way to generate topics tending to have high probability on groups of words that often occur together. This simple approach can deal with synonym and polysemy: even when documents differs in the use of some terms for naming particular concepts, all of them will tend to belong to the same topic because they co-occur with other words that surround the meaning. For a detailed explanation of LDA and the statistical foundations supporting it, please refer to Blei's paper [4] or [27, 13].

In the last years, LDA has gain an important attention in the text analyisis and machine learning fields. Many different variation of the model have been proposed to address different scenarios. Labeled LDA [24] incorporates external knowledge in the form of tags to drive the generation of the topics. Author-topic model [25] blends the authorship information to generate topics for a collection of scientific papers. Hierarchical LDA -or HLDA- [5, 6] builds a hierarchy of topics where a parent topics contains more general terms. In HLDA every document is represented as a distribution over the topics in one path in the hierarchical tree of topics created by the model, and each topic is a distribution over the join vocabulary among all topics in the same level. An extense description of structured topic models including HLDA is made in [28].

Altought LDA based models has been applied to a numerous of different collections of documents for discovering structured topics, there is no known attempt of using these techniques for modeling hierarchical structured collections -such as textbooks- for the task of document linking.

## 3. RESEARCH QUESTIONS

The main goal of this work is to explore the use of probabilistic topic models for getting high accuracy on matching textbook parts (chapter, sections and subsections). Two topic models are used: LDA and HLDA. The first research question is:

> ***Q1****: Will probabilistic topic models perform a more accurate document linking than common term-based aproach when applied in collection of textbooks?*

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 the content of it's children. The assignment of topic distribution among parts of the textbooks in different levels of the hierarchy should consider these characteristics for a correct representation of the document content. To this regard, we state a second research question:

> ***Q2****: What would be the best approach to incorporate the hierarchical structure of the textbooks in the topic model?*

We need a topic model that correct represent each document (each textbook part) as the aggregation of the content of its child parts. One option is to build the topic model from all textbook parts with aggregated content. However, this approach seems to "confuse" LDA and showed to perform very poorly in our preliminary tests for

both LDA and HLDA. Other, more reasonable approaches, consider to build the model using only documents that have text (usually leaf nodes) and further options for indexing the intermediate nodes. We consider two options:

a) Aggregate topic distributions along the hierarchical structure by weighting children topic's distributions by their sizes. For simplicity we further call this option *topic aggregation* (TA).

b) Re-indexing the aggregated content of the intermediate documents using the inference mechanism provided by the topic model. For simplicity we call this *document re-indexing* (RI).

Additionally, since the model should reflect the domain regardless of the differences on the terminology that different authors may use, we consider to build the model using several books. We expect that a model built using multiples textbooks will reflect a better understanding of the domain and performs a better document linking than a model build using a single textbook. For simplicity we further call these options *single book* (SB) and *multiple book* (MB), respectively. We state a third research question:

> ***Q3****: Will the model built using multiple textbooks (MB) performs a better document linking compared with a model built using a single textbook (SB)?*

## 4. EXPERIMENTS DESIGN

To address the research questions we conducted several experiments running LDA and HLDA in different conditions within two collection of textbooks: Algebra and Information Retrieval. The conditions are evaluated by the effectivity of the resulting topic model in the task of matching textbook parts (or documents) within two textbooks. We use both *cosine* similarity and a reciprocal *KL-divergence* on the document-topic distributions provided by the model to perform the matching. The effectivity is measured with average NDCG(1), NDCG(3) and NDCG(10) and using a ground truth made by experts that consists of an ideal document mapping between the two books.

### 4.1 Textbooks

Four and five textbooks were downloaded and parsed for algebra and information retrieval, respectively. All text was converted to lowercase and stopwords were removed. Additional frequent words in the domain were also remove (for example: "exercises", "solutions" in algebra.) All textbooks in the algebra domain had free access through the Web at the moment we downloaded them. For Information Retrieval books, four of the five books were given with the consent of the authors to our lab for academic purposes. One book was downloaded from the Web. For further references, the textbooks are labeled as BOOK 1, BOOK2, etc. BOOK 1 is used for building topic models in the Single Book condition (SB). BOOK 3, 4 and 5 are used for building topic models in the Multiple Book condition (MB). BOOK 2 is used for evaluation.

#### *4.1.1 Elementary Algebra textbooks*

- **BOOK 1**: Elementary Algebra, by Wade Ellis, Denny Burzynski. HTML version accessed from http://cnx.org/content/col10614/latest/
- **BOOK 2**: Elementary Algebra, v1, by John Redden. HTML version accessed from http://catalog.flatworldknowledge.com/bookhub/reader/128?e=fwk-redden--ch01_s01
- **BOOK 3**: Understanding Algebra, by James W. Brennan. HTML version accessed from http://www.jamesbrennan.org/algebra/
- **BOOK 4**: Fundamentals of Mathematics, edited by Denny Burzynski and Wade Ellis. HTML version accessed from http://cnx.org/content/col10615/1.4/

#### *4.1.2 Information Retrieval books*

- **BOOK 1**: Introduction to Information Retrieval, by C. D. Manning, P. Raghavan and H. Schütze, Cambridge University Press. 2008. HTML accessed from http://nlp.stanford.edu/IR-book
- **BOOK 2**: Modern Informatio Retrieval, by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.
- **BOOK 3**: Finding Out About, by Richard K. Belew. Text accessed from http://cseweb.ucsd.edu/~rik/foa/l2h/
- **BOOK 4**: Information Storage and Retrieval Systems, by Gerald Kowalski.

- **BOOK 5**: Information Retrieval, by C.J. van Rijsbergen.

## 4.2 Ground Truth

The ground truth is a manual mapping of 2 textbooks (BOOK1 and BOOK 2) made by experts in both domains. 10 experts contributed: 1 professor and 6 PhD student from the Information Science School at the Univrsity of Pittsburgh, 3 researchers from CelTech research lab at DFKI (Universitaat der Saarland.) In the Algebra domain 5 chapters of BOOK 1 were mapped to BOOK 2. In the Information Retrieval domain, 4 chapters of BOOK 1 were mapped to BOOK 2. We ensured to have 2 expert mapping each chapter. We developed a web interface for facilitate the mapping task. Directions were given in order to have an accurate content matching: i) every chapter, section and subsection in the BOOK 1 can be mapped to 0 or more parts in the BOOK 2; ii) match as accurately as possible, i.e try to find the parts that better fit the content: each match could relate book parts on different levels, for example, a chapter matching a section; iii) consider that the content of a textbook part is the aggregation of the content of its sub-parts, e.g., the content of a section is the sum of all the content of its sub-sections.

Additionally, a level or relevance and confidence, both ranging from 1 to 3 (low, medium, high) were asked for each match.

Finally, the ground truth was the compitation of all experts mappings blended in a single list. For each mapped part of the BOOK 1, all matches were ranked with a score computed using relevance and confidence. High scores were placed to match when both experts agreed.

## 4.3 Conditions and Hypotheses

Q1 stresses the idea of comparing LDA and HLDA with a baseline. The baseline is the effectivity of document matching using the term-based approach implemented in Apache Lucene (http://lucene.apache.org) and is further described in the next section. Q2 stresses the idea of comparing different approaches to incorporate the hierarchy in the topic model. For simplicity, we further refer to this factor as the *aggregation* strategy and have two levels: *topic aggregation* (TA) and *re-indexing* (RI), as described in section 3. Q3 stresses the comparison between a model built using a single book with a model built using multiple books from the collection. For simplicity, we further call this factor *books* and it have two levels: *single book* (SB) and *multiple books* (MB). Additionally, similar documents are matched using cosine similarity and reciprocal symmetric KL-Divergence. For simplicity we further call this factor *similarity* with two levels: *Cosine* and *KL-Divergence* (note that we simply refer as KL-Divergence to a reciprocal symmetric KL-Divergence). The 8 model conditions are shown in Table 1. Each of them is applied to both LDA and HLDA in the 2 domains.

**Table 1:** The eight conditions.

| | | Aggregation | | | |
|---|---|---|---|---|---|
| | | Topic Aggregation (TA) | | Re-Indexing (RI) | |
| | | KL-Divergence | Cosine | KL-Divergence | Cosine |
| Books | Single Book (SB) | TA-SB-KLDiv | TA-SB-Cos | RI-SB-KLDiv | RI-SB-Cos |
| | Multiple Books (MB) | TA-MB-KLDiv | TA-MB-Cos | RI-MB-KLDiv | RI-MB-Cos |

For addressing the research question 1 (Q1), the eight conditions are compared to the baseline using NDCG(1), NDCG(3) and NDCG(10). We define th following hypotheses:

**H1.1** A model built using LDA performs a better document matching than the baseline.

**H1.2** A model built using HLDA performs a better document matching than the baseline.

Since the baseline accuracy score is a fixed value, we use *one-sample t-test* to test the hypotheses H1.1 and H1.2.

For addressing the research question Q2 and Q3, we state the following hypotheses:

**H2** A probabilistic topic model that aggregates the topic probabilities among the book hierarchies (TA) will perform a better document matching than a model using re-indexing of the aggregated content (RI).

**H3** A probabilistic topic model built using several textbooks (MB) will perform a better document matching than a model built using a single textbook (SB).

For testing H2 and H3, the different condition groups are compared between them using a three-way ANOVA. Interactions, multigroups comparison and marginal means are reported.

## 4.4 Measuring the Effectivity for Document linking

We evaluate the effectivity of a model condition in finding similar documents using average NDCG at 1, 3 and 10 as follows: i) for each part in the BOOK 1, all the parts in the BOOK 2 are ranked computing cosine similarity and reciprocal symmetric KL-divergence on the respective topic distributions; ii) each ranked list is evaluated to the respective rank list in the ground truth (ideal rank with scores) using NDCG at 1, 3 and 10; iii) an average NDCG at 1, 3 and 10 is computed among all the parts in BOOK 1. Since both LDA and HLDA use random seeds in the sampling process for analyzing the collections, every run produce different topic sets. Considering this, each model condition is run 30 times (N=30), and the average and standard deviation are reported in the statistical tests (this is not the number of iterations of the sampling process for buildng the topic model. Each model iterates 2000 or 4000 times as explained in the next section). NDCG(1) represents the ability of the finding the top similar document in the first position. NDCG(3) and NDCG(10) relax the score letting the ranking to have relevant documents up to positions 3 and 10 respetively. However, NDCG penalizes the score when relevant documents are fund in lower positions respecting the position they occupe in the ideal rank.

### *4.4.1 Baseline*

The baseline is the effectivity (average NDCG at 1, 3 and 10) of the ranked lists resulting of querying an index built using Apache Lucene (http://lucene.apache.org). For each part in the BOOK 1, a query is performed to the index. The index is built using BOOK 2. Lucene performs similarity between query and documents using a variant of the TFIDF model described in [1].

## 4.5 LDA and HLDA set up

We use MALLET Toolkit implementation of LDA and HLDA [23]. In MALLET, LDA set up depends on the number of topics, the number of iterations for the sampling process, the smoothing over topic distribution hyperparameter $\alpha$, and the smoothing over topic-word distribution hyperparameter $\beta$ (a good explanation of the LDA hyperparameters can be found in [27]). 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 the textbooks' content, we estimated that the number of topics should be a number between the number of sections and the number of subsections in the average book (sometimes sections covers several topics, and sometimes subsections cover examples, exercises or different views of the same topic). In the algebra domain BOOK 1 have 74 sections and 228 subsections, thus we choose 150 topics. This number also gave us the best results in the preliminary tests. In Information retrieval domain the BOOK 1 has 120 sections and 178 terminal nodes (some sections does not open in subsections and are counted as subsections). Here we also chose 150 topics. About hyperparameters, we set up initial values of $\alpha = 0.01$ and $\beta = 0.01$, and then used the fixed-point optimization for hyperparameters [28] implemented in MALLET.

In HLDA, the number of topics is a result of the algorithm. Instead, the number of levels (L) of the topic hierarchy should be provided in MALLET implementation. We choose four levels (L=4) because it copy the domain structure (domain, chapter, section, subsection). Also, in HLDA three hyperparameters controls the shape of the tree and the topic generation. The hyperparameter $\gamma$ controls the tendency of creating new branches (new topics) or following existig paths when the sampling method assigns words to topics. Small values of $\gamma$ ($\gamma << 1$) tend to produce trees with fewer and umbalanced branches. The hyperparameter $\alpha$ is a smoothing over the level distribution in a document. Higher values of $\alpha$ smooths the differences of the probability of the topics in the path of each document (remember that in HLDA each document is represented as a distribution over the topics in one path in the topic hierarchy). The hyperparameter $\eta$ smooths the topic-word distributions. Small values of $\eta$ ($\eta << 1$) tend to concentrate topic probabilities in small number of words and thus produce more specific topics. The

**Table 2**. The four conditions of LDA using KL-Divergence in the Algebra domain.

| | NDCG(1) | | | NDCG(3) | | | NDCG(10) | | |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 0.3662 | | | 0.5807 | | | 0.6582 | | |
| | Mean | Std. Dev. | Sig. (p) | Mean | Std. Dev. | Sig. (p) | Mean | Std. Dev. | Sig. (p) |
| LDA-SB-TA | 0.547 | 0.025 | <.001 | 0.647 | 0.018 | <.001 | 0.691 | 0.015 | <.001 |
| LDA-MB-TA | 0.532 | 0.036 | <.001 | 0.62 | 0.021 | <.001 | 0.663 | 0.017 | 0.165 |
| LDA-SB-RI | 0.456 | 0.027 | <.001 | 0.601 | 0.027 | <.001 | 0.675 | 0.019 | <.001 |
| LDA-MB-RI | 0.414 | 0.04 | <.001 | 0.572 | 0.032 | 0.132 | 0.647 | 0.026 | 0.022 |

number of topics and the shape of the tree are highly sensitive to the combination of γ and η [6]. While small γ can constraint the tendency of creating new branches, small values of η (more concentrate topics) will tend to generate a sparser tree.

In our preliminary tests we found that sparse trees produce better results in the document matching task (more specific topics are more discriminative), and that this sparsity is much more dependent on lower values of η than higher values of γ. This is consistent with [26] who states that HLDA is highly sensitive to η. Also, the number of topics in the tree depends highly in the number of documents in the collection used during the sampling process. Since we test a single book (SB) and a multiple book (MB) while generating the models and we want to generate similar amount of topics in each condition, we choose either η=0.01 and η=0.001 for single book and multiple books conditions, respectively. In preliminary tests we found that these values give similar number of topics, respectively. About γ, we found out that small values (γ<1) tend to produce highly unbalanced trees. After several preliminary tests, we then chose γ=3.5. Good results were less dependant of α and in general, the model performs better with relatively high values. We chose α=5.

## 5. RESULTS

### 5.1 Do LDA and HLDA perform better than baseline?

For testing H1.1 and H1.2, a one-sample t-test was performed to compare the four conditions (SB-TA, MB-TA, SB-RI, MB-RI) in each model (LDA, HLDA) with the scores of the baseline in 3 levels of NDCG (1, 3, 10) using both Cosine similarity and symmetric KL-Divergence.

#### *5.1.1 Algebra domain*

In Algebra, LDA conditions using KL-Divergence gave higher scores than Cosine in all NDCG levels, thus KL-Divergence scores are reported in table 2. Figures 1 and Figure 2 show NDCG(1) and NDCG(10) scores for all conditions compared with the baseline (dotted line). Further analysis comparing the different conditions, including KL-Divergence and Cosine as a factor are performed in the next section.

Examining **Table 2**, all results were significantly higher than the baseline except for LDA-MB-RI in NDCG(3) ($M$ = .572, $SD$ = .032, $p$ = .132), LDA-MB-TA in NDCG(10) ($M$ = .663, $SD$ = .017, $p$ = .165) and LDA-MB-RI in NDCG(10) which was significantly lower than the respective baseline ($M$ = .647, $SD$ = .026, $p$ = .022). Among the significantly higher scores, the higher effect sizes were always presented by LDA-SB-TA (NDCG(1): $M$ = .546, $SD$ = .025, $p$ < .001, Cohen's $d$ = 7.232; NDCG(3): $M$ = .647, $SD$ = .018, $p$ < .001, Cohen's $d$ = 3.683; NDCG(10): $M$ = .691, $SD$ = .015, $p$ < .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 **H1.1** in the Algebra domain.

All conditions using HLDA were significantly lower than the baseline. The extreme scores were HLDA-SB-TA ($M$ = .242, $SD$ = .045, $p$ < .001, Cohen's $d$ = 2.76) and HLDA-MB-RI ($M$ = .124, $SD$ = .03, $p$ < .001, Cohen's $d$ = 13.6). **H1.2** was not supported in Algebra domain.

**Table 3**. The four conditions of LDA using KL-Divergence in the Information Retrieval domain.

| | NDCG(1) | | | NDCG(3) | | | NDCG(10) | | |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 0.057 | | | 0.186 | | | 0.258 | | |
| | Mean | Std. Dev. | Sig. (p) | Mean | Std. Dev. | Sig. (p) | Mean | Std. Dev. | Sig. (p) |
| LDA-SB-TA | 0.345 | 0.051 | <.001 | 0.461 | 0.042 | <.001 | 0.536 | 0.033 | <.001 |
| LDA-MB-TA | 0.309 | 0.063 | <.001 | 0.418 | 0.045 | <.001 | 0.52 | 0.039 | <.001 |
| LDA-SB-RI | 0.36 | 0.066 | <.001 | 0.484 | 0.053 | <.001 | 0.556 | 0.045 | <.001 |
| LDA-MB-RI | 0.336 | 0.05 | <.001 | 0.456 | 0.054 | <.001 | 0.534 | 0.041 | <.001 |

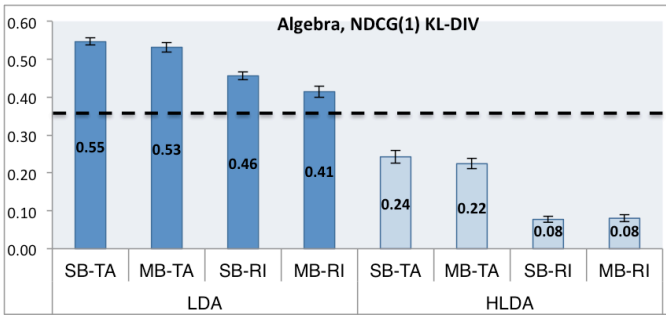


**Figure 1**: LDA and HLDA conditions compared with the baseline for NDCG(1) in Algebra domain. Baseline is indicated with a dotted line.

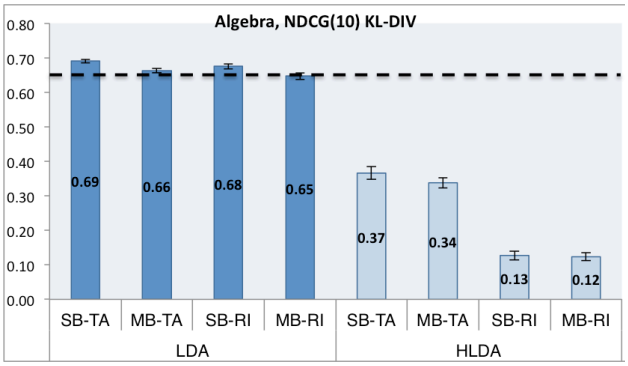


**Figure 2**: LDA and HLDA conditions compared with the baseline for NDCG(10) in Algebra domain. Baseline is indicated with a dotted line.

*5.1.2 Information Retrieval*

Using the Information Retrieval collection, all LDA conditions performs better than the baseline using both cosine and KL-Divergence. KL-Divergence values are reported in Table 3. Figure 3 and Figure 4 show NDCG(1) and NDCG(10) scores for all conditions compared with the baseline (dotted line). These results support H1.1.

There were only two HLDA conditions that present significantly higher scores than the baseline, both of them for NDCG(1): HLDA-SB-TA ($M$ = .145, $SD$ = .060, $p$ < .001,Cohen's $d$ = 1.478) and HLDA-MB-TA ($M$ = .153, $SD$ = .057, $p$ < .001, Cohen's $d$ = 1.697). For NDCG(3) and NDCG(10) HLDA conditions performed poorly than the baseline. These results partially supports **H1.2**.

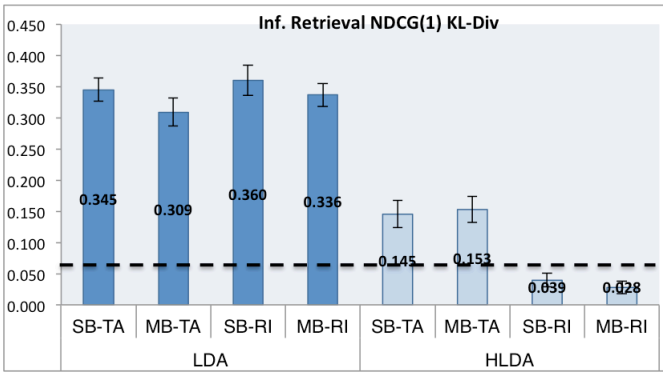


**Figure 3**: LDA and HLDA conditions compared with the baseline for NDCG(1) in Information Retrieval domain. Baseline is indicated with a dotted line.

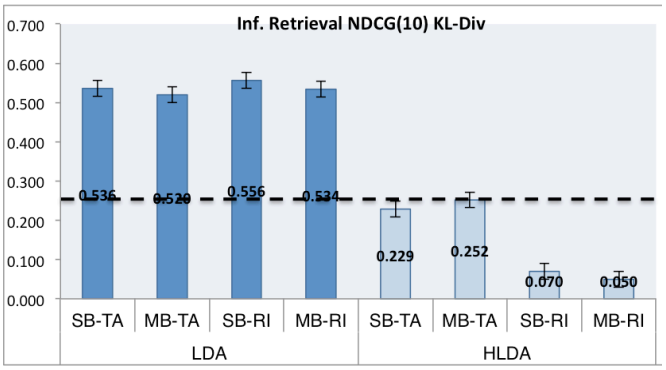


**Figure 4**: LDA and HLDA conditions compared with the baseline for NDCG(10) in Information Retrieval domain. Baseline is indicated with a dotted line.

H1.1 is supported: LDA performs significantly better than the baseline in both domains.

H1.2 is partially supported: HLDA performs significantly lower than the baseline in Algebra domain and significantly higher than the baseline only for NDCG(1) in the conditions including Topic Aggregation in the Information Retrieval domain. A possible explanation grounds in the nature of HLDA, where each document is represented by a single path in the topic hierarchy. For example, in a hierarchy of 4 levels, each document is represented with 4 topics. We may expect the model having less ability than LDA to discriminate between documents because it represents each document with just few topics.

## 5.2 Do TA performs better than RI?

For testing H2 in each domain, a three-way between subjects ANOVA was performed on scores for NDCG(1), NDCG(3), and NDCG(10) as a function of the *aggregation* strategy (TA, RI), the *books* strategy (SB, MB) and the *similarity* strategy (Cosine, KL-Divergence). HLDA was not included in this analysis due to the poor performance compared with the baseline.

### *5.2.1 Algebra domain*

The assumption of homogeneity of the variance was met for NDCG(1), $F(7,232) = 1.173$, $p = .319$ and NDCG(10), $F(7,232) = 1.717$, $p = .106$. Homogeneity of the variance was not met for NDCG(3), $F(7,232) = 2.344$, $p = .025$, thus it is discarded of further analysis. Assumption of normality was met for all conditions in NDCG(1) and NDCG(10). NDCG(3) did not met this assumption in conditions TA-MB-Cosine (*Shapiro-Wilk* = .928, $p = .043$), RI-SB-Cosine (*Shapiro-Wilk* = .928, $p = .043$) and RI-MB-Cosine (*Shapiro-Wilk* = .925, $p = .037$). All other assumptions were met.

The patterns of difference on NDCG(1) scores among strategies of similarity were significantly different between TA and RI (**Figure 5**), $F(1,239) = 23.862$, $p<.001$, $\eta^2 = .093$. No other significative interactions were observed among NDCG levels. About main effect of the aggregation, marginal means for NDCG(1) and NDCG(10) are shown in Table 4. Results showed that TA results in significantly higher scores than RI for all NDCG levels averaged across books strategy and similarity strategy. This support H2. For NDCG(1), TA ($M$ = .507, $SE$ = .003) was significantly higher than RI ($M$ = .425, $SE$ = .003), $F(1,239) = 312.417$, $p < .001$, $\eta^2 = .574$. For NDCG(10), TA ($M$ = .665, $SE$ = .002) was significantly higher than RI ($M$ = .651, $SE$ = .002), $F(1,239) = 26.894$, $p < .001$, $\eta^2 = .104$.

In order to find the pattern of differences on the NDCG(1) scores among *similarity* strategies (Cosine, KL-Divergence) for TA and RI *aggregation* strategies, a post hoc test was performed to compare the four conditions: TA-Cos, TA-KLDiv, RI-Cos, and RI-KLDiv. In NDCG(1), TA-KLDiv ($M$ = .539) performed significantly better than the other three conditions: TA-Cos ($M$ = .475), $p < .001$, RI-KLDiv ($M$ = .435), $p < .001$, and RI-KLCos ($M$ = .416), $p < .001$. TA-Cos ($M$ = .475) performs significantly higher than RI-Cos ($M$ = .416), $p < .001$, and RI-KLDiv ($M$ = .435), $p < .001$. However, there was no significant difference between RI-KLDiv and RI-Cos. These results support **H2** for NDCG(1): TA is always better than RI.

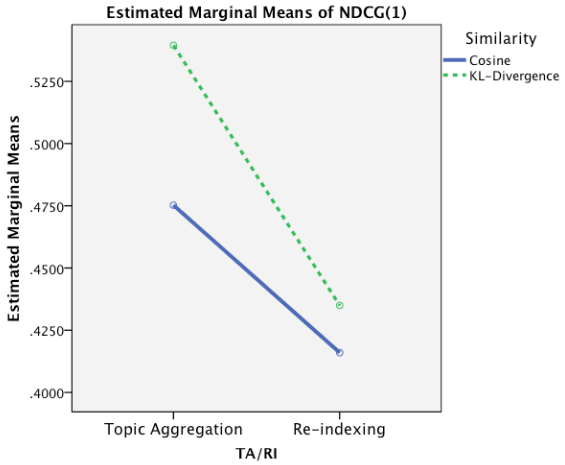


**Figure 5**: Interaction between aggregation strategy and similarity strategy for NDCG(1) in the Algebra domain.

*5.2.2 Information Retrieval*

The assumption of homogeneity of the variance was met for NDCG(1), $F(7,232) = 1.629$, $p = .128$ and NDCG(3), $F(7,232) = .941$, $p = .475$, and NDCG(10), $F(7,232) = 1.205$, $p = .301$. Assumption of normality was met for all conditions in NDCG(1). NDCG(3) did not meet this assumption in condition RI-MB-KLDiv (*Shapiro-Wilk* = .929, $p = .047$). NDCG(10) did not meet normality in condition RI-MB-Cosine (*Shapiro-Wilk* = .922, $p = .031$). All other assumptions were met.

**Table 4** (bottom part) shows marginal means for all factor levels in NDCG(1), NDCG(3) and NDCG(10) for Information Retrieval domain. The patterns of difference on NDCG(3) and NDCG(10) scores among strategies of similarity were significantly different between TA and RI. These interactions are: in NDCG(3), $F(1,239) = 10.457$, $p < .001$, $\eta^2 = .043$; in NDCG(10), $F(1,239) = 6.825$, $p = .01$, $\eta^2 = .029$. No significant interaction were observed in NDCG(1) involving aggregation strategy. There was no significant main effect of the aggregation strategy in all levels of NDCG. These results do not support **H2**.

In order to find the pattern of differences on the NDCG(3) and NDCG(10) scores among similarity strategies (Cosine, KL-Divergence) for aggregation strategies (TA, RI), a post hoc test was performed to compare the four conditions: TA-Cos, TA-KLDiv, RI-Cos, and RI-KLDiv. In NDCG(3), TA-KLDiv ($M = .439$) was significantly lower than the other three conditions: TA-Cos ($M = .492$), $p < .001$; RI-KLDiv ($M = .47$), $p = .007$; RI-Cos ($M = .484$), $p < .001$. There was no other significant difference among conditions. This suggest that TA strategy is highly dependent of the *similarity* strategy used: TA performs lower when KL-Div is used. In NDCG(10), TA-Cos ($M = .571$) was significantly better than TA-KLDiv ($M = .528$), $p < .001$, and RI-KLDiv ($M = .545$), $p = .003$. RI-Cos ($M = .564$) was significantly better than TA-KLDiv ($M = .528$), $p < .001$. No other significant difference among conditions was found in NDCG(10). These results suggest that there is no difference in choosing between TA and RI when either Cosine or KL-Divergence is used. However, Cosine performs better when TA is chosen.

H2 is partially supported: In Algebra domain, Topic Aggregation (TA) strategy performs better than Re-Indexing (RI). Interactions with similarity strategies shows that when TA is used, KL-Divergence is the best similarity strategy. However, the results are different in Information Retrieval domain, where there is no difference between TA and RI and Cosine performs better than KL-Divergence.

One possible explanation is the different nature of the two domains and the difference in the collections. For example, we look at the vocabulary size that LDA is using (after stopword removal) in each domain and found out that in Algebra, a single book produce a vocabulary of 2,220 terms, and in Information Retrieval, a single book model worked with a vocabulary of 13,405 terms.

## 5.3 Do MB perform better than SB?

The ANOVA analysis described in the previous section is also used to test H3.

### *5.3.1 Algebra domain*

In Algebra domain, there were no significant interactions among books and similarity strategies and/or aggregation strategies. About main effect of books strategies, results showed that SB produce significantly higher scores than MB for all NDCG levels averaged across aggregation strategies and similarity strategies (Table 4, row labeled *Books*). For NDCG(1), SB ($M = .482$, $SE = .003$) was significantly higher than MB ($M = .451$, $SE = .003$), $F(1,239) = 46.306$, $p < .001$, $\eta^2 = .166$. For NDCG(3), SB ($M = .60$, $SE = .003$) was significantly higher than MB

**Table 4**: Marginal Means

| | | NDCG(1) | | NDCG(3) | | NDCG(10) | |
|---|---|---|---|---|---|---|---|
| **Algebra** | | **M** | **SE** | **M** | **SE** | **M** | **SE** |
| Aggregation | Topic Aggregation (TA) | .507 | .003 | .602 | .003 | .665 | .002 |
| | Re-Indexing (RI) | .425 | .003 | .568 | .003 | .651 | .002 |
| Books | Single Book (SB) | .482 | .003 | .600 | .003 | .672 | .002 |
| | Multiple Books (MB) | .451 | .003 | .570 | .003 | .644 | .002 |
| Similarity | Cosine | .446 | .003 | .561 | .003 | .647 | .002 |
| | KL-Divergence | .487 | .003 | .610 | .003 | .669 | .002 |
| | | **NDCG(1)** | | **NDCG(3)** | | **NDCG(10)** | |
| **Information Retrieval** | | **M** | **SE** | **M** | **SE** | **M** | **SE** |
| Aggregation | Topic Aggregation (TA) | .339 | .006 | .465 | .004 | .550 | .003 |
| | Re-Indexing (RI) | .349 | .006 | .477 | .004 | .555 | .003 |
| Books | Single Book (SB) | .355 | .006 | .485 | .004 | .559 | .003 |
| | Multiple Books (MB) | .333 | .006 | .458 | .004 | .545 | .003 |
| Similarity | Cosine | .351 | .006 | .488 | .004 | .568 | .003 |
| | KL-Divergence | .338 | .006 | .455 | .004 | .537 | .003 |

($M$ = .57, $SE$ = .003), $F(1,239) = 69.409$, $p < .001$, $\eta^2 = .230$. For NDCG(10), SB ($M$ = .672, $SE$ = .002) was significantly higher than MB ($M$ = .644, $SE$ = .002), $F(1,239) = 98.257$, $p < .001$, $\eta^2 = .298$. These results do not support **H3**.

*5.3.2 Information Retrieval*

Similar result were obtained in Information Retrieval domain regarding *books* strategies. There were no significant interactions with the other factors (*similarity*, *aggregation*) and the main effect of *books* were significant in all levels of NDCG averaged across *aggregation* strategies and *similarity* strategies (see Table 5, bottom part). For NDCG(1), SB ($M$ = .355, $SE$ = .006) was significantly higher than MB ($M$ = .333, $SE$ = .006), $F(1,239) = 7.981$, $p = .005$, $\eta^2 = .033$. For NDCG(3), SB ($M$ = .485, $SE$ = .004) was significantly higher than MB ($M$ = .458, $SE$ = .004), $F(1,239) = 20.361$, $p < .001$, $\eta^2 = .081$. For NDCG(10), SB ($M$ = .559, $SE$ = .003) was significantly higher than MB ($M$ = .545, $SE$ = .003), $F(1,239) = 7.919$, $p = .005$, $\eta^2 = .033$. These results do not support **H3**.

H3 is not supported in either Algebra and Information Retrieval domains. Results showed that the model built using a single book (SB) will give significantly better results than the model built using multiple books (MB), regardless of the use of different strategies for aggregation and similarity.

## 6. CONCLUSIONS

In this work we explored the use of LDA and HLDA topic models within collections of textbooks for the task of document linking. We applied different approaches for building the topic model considering the hierarchical structure of the textbooks. We showed that LDA is a valuable alternative and performs much better than term-based approaches, specially, for finding the top similar documents (NDCG(1), NDCG(3)). We discovered that a simple mechanism of aggregating the weighed topic distributions along the hierarchical structure of the textbooks works the best, and that the topic model build using one textbook makes a better document matching than a model built using multiple books. The results also showed that different apporaches for computing similarity among topic distributions worked differently in different domains.

We believe that using LDA is a promising approach for addresing the problem of horizontal navigation using open corpus structured collections. The ability of the topic-based model to accurately find the very first top similar documents is a clear advantage over traditional methods and can be used to implement better recommender support in adaptive educational hypermedia systems. It is in our research agenda to incorporate this technology in the e-learning environments as long as to keep investigating mechanisms to improve the quality of the models and extend the use to tasks other than document matching. Our future work includes to apply techniques of topics models evaluation, to combine topics and textbooks structure to discover semantic relations among the topics, to combine topics models with keyword and concept extraction techniques, and to further investigate the application in other domains.

## 7. ACKNOWLEDGMENTS

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.

Thanks to my adisor, Peter Brusilovsky, and to Sergey Sosnovsky for their valuable directions and feedback during this work. Special thanks to all the people in both PAWs and CelTech that helped in the textbooks mapping task.

## 8. REFERENCES

[1] Apache Software Foundation. n.d. Lucene TF-IDF similarity class API documentation. Retrieved November 2, 2012, from http://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

[2] Baeza-Yates, R., Ribeiro-Neto, B. 1999. Modern Information Retrieval, *Addison-Wesley Longman Publishing Co.,* Inc., Boston, MA, 1999

[3] Bareiss, R. and Osgood, R. 1993. Applying AI models to the design of exploratory hypermedia systems. In *Proceedings of Fifth ACM Conference on Hypertext*, Seattle, WA, 1993, ACM Press, pp. 94-105.

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

[5] Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. 2004. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems* 16. Cambridge, MA: MIT Press.

[6] Blei, D. M., Griffiths, T. L., and Jordan, M. I. 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM 57, 2, Article 7*, (January 2010), 30 pages. DOI = 10.1145/1667053.1667056 http://doi.acm.org/10.1145/1667053.1667056

[7] Blei, D.M. 2012. Probabilistic Topic Models, *Communications of the ACM*, 55 (4), April, pp 77–84.

[8] Brusilovsky, P. and Rizzo, R. 2002. Map-based horizontal navigation in educational hypertext. *Journal of Digital Information* 3 (1), http://jodi.ecs.soton.ac.uk/Articles/v03/i01/Brusilovsky/.

[9] Carr, L., Hall, W., Bechhofer, S., and Goble, C. 2001. Conceptual linking: Ontology-based open hypermedia. In *Proceedings of 10th International World Wide Web Conference*, Hong Kong, May 1-5, 2001, ACM Press, pp. 334-342.

[10] Cleary, C. and Bareiss, R. 1996. Practical methods for automatically generating typed links. In *Proceedings of 7th ACM Conference on Hypertext*, Washington, DC, March 16-20, 1996, ACM, pp. 31-41.

[11] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407, 1990.

[12] Fountain, A., Hall, W., Heath, I., and Davis, H. 1992. MICROCOSM: an open model for hypermedia with dynamic linking. In: *N. Streitz, A. Rizk and J. André (eds.): Hypertext: concepts, systems and applications*. Cambridge University Press, pp. 298-311.

[13] Heinrich, G. 2005. Parameter estimation for text analysis. Technical report, 2005.

[14] Green, S. 1999. Building hypertext links by computing semantic similarity. *IEEE Transactions on Knowledge and Data Engineering* 11 (5), 713-730.

[15] Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of SIGIR '99*, Berkeley, CA, USA, 1999.

[16] Kibby, M. R. and Hayes, J. T. 1989. Towards intelligent hypertext. In: *R. McAleese (ed.) Hypertext: Theory into practice*. Oxford: Intellect, pp. 164-171.

[17] Kolak, O. and Schilit, B. N. 2008. Generating Links by Mining Quotations. In: *Proceedings of The 19th ACM Conference on Hypertext & Hypermedia*, Pittsburgh, Pennsylvania, USA, June 19-21, 2008, pp. 117-126.

[18] Kullback, S., Leibler, R.A. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics* 22 (1): 79–86. doi:10.1214/aoms/1177729694

[19] Lee, L. 1999. Measures of distributional similarity. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p.25-32, June 20-26, 1999, College Park, Maryland. doi 10.3115/1034678.1034693

[20] Macedo, A. A., Pimentel, M. d. G. C., and Camacho-Guerrero, J. A. 2002. An infrastructture for open latent semantic linking. In: *K. M. Anderson, S. Moulthrop and J. Blustein (eds.) Proceedings of 13th ACM Conference on Hypertext and Hypermedia* (Hypertext 2002), College Park, MD, June 11-15, 2002, ACM, pp. 107-115.

[21] Manning, C. D., Raghavan, P., Schütze, H. 2008. *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, 2008

[22] Mayes, J. T., Kibby, M. R., and Watson, H. 1988. StrathTutor: The development and evaluation of a learning-by-browsing on the Macintosh. *Computers and Education* 12 (1), 221-229.

[23] McCallum, Andrew Kachites. 2002. *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu. 2002.

[24] Ramage, D., Hall, D., Nallapati, R., and Manning C. D. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*: Volume 1 - Volume 1 (EMNLP '09), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 248-256.

[25] Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P. 2004. The author-topic model for authors and documents. In: *The 20th conference on Uncertainty in artificial intelligence*, pp. 487–494 (2004)

[26] Sivic, J., Russell, B.C., Zisserman, A., Freeman, W.T., Efros, A.A. 2008. Unsupervised discovery of visual object class hierarchies. *Computer Vision and Pattern Recognition*, 2008. CVPR 2008. IEEE Conference on , vol., no., pp.1-8, 23-28 June 2008. doi: 10.1109/CVPR.2008.4587622

[27] Steyvers, M. and Griffiths, T. 2007. Probabilistic topic models. In T. Landauer, D.S.McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*. Erlbaum, 2007.

[28] Wallach, H. M. 2008. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.