# Analysis of Computer Science Related Curriculum on LDA and Isomap

Takayuki Sekiya Information Technology Center, the University of Tokyo 3-8-1 Komaba, Meguro, Tokyo, Japan sekiya@ecc.u-tokyo.ac.jp Yoshitatsu Matsuda Department of Integrated Information Technology, Aoyama Gakuin University, 5-10-1 Fuchinobe, Sagamihara, Kanagawa, Japan matsuda@it.aoyama.ac.jp

Kazunori Yamaguchi Graduate School of Arts and Sciences, the University of Tokyo 3-8-1 Komaba, Meguro, Tokyo, Japan yamaguch@graco.c.utokyo.ac.jp

## ABSTRACT

A good curriculum is crucial for a successful university education. When developing a curriculum, topics, such as natural science, informatics, and so on are set first, course syllabi are written accordingly. However, the topics actually covered by the courses are not guaranteed to be identical to the initially set topics. To find out if the actual topics are covered by the developed course syllabi, we developed a method of systematically analyzing syllabi that uses latent Dirichlet allocation (LDA) and Isomap. We applied this method to the syllabi of MIT and those of the Open University, and verified that the method is effective.

## **Categories and Subject Descriptors**

K.3.2 [**Computers and Education**]: Computer and Information Science Education—*curriculum* 

## **General Terms**

Experimentation

#### Keywords

Syllabus, Curriculum, Curriculum Analysis, LDA, Isomap

#### 1. INTRODUCTION

A curriculum should represent characteristic educational activity that each university offers to students. To design such an original curriculum, faculties have to analyze current curricula, however, it is not an easy task to grasp characteristics of a curriculum because the analysis of a curriculum requires professional knowledge in various fields.

In this paper, we propose a method to generate a map of the syllabi from which we can understand the whole structure of a curriculum represented by its syllabi. To generate such a map, there are two problems. First, the distribution

ITiCSE'10, June 26–30, 2010, Bilkent, Ankara, Turkey.

Copyright 2010 ACM 978-1-60558-820-9/10/06 ...\$10.00.

of terms is sparse and the overall structure of the syllabi cannot be determined by terms themselves. To overcome the first problem, we use topics instead of term sets. The topics are extracted from syllabi using latent Dirichlet allocation (LDA)[4]. Second, the distribution of syllabi is distortedly located in high-dimensional topic space. To overcome the second problem, we use Isomap[12].

Using these methods, we conducted experiments. First, in order to see the reliability and informativeness of our method, we applied our method to CS2008[13]<sup>1</sup>, which is the curricular guidance of computer science. Then, we applied our method for comparison of two curricula. We first extracted model topics from CS2008, and created a map of the computer science curriculum of MIT as a reference curriculum. Then, we plotted the course syllabi of the Open University (OU) curriculum into the map of the MIT curriculum. From this comparison, we can see what range of the standard computer science topics is covered by the computer science courses of MIT and OU.

Related works are explained in Section 2. In Section 3, basic theories are explained. Experimental results are detailed in Section 4. Section 5 concludes this paper.

#### 2. RELATED WORK

Since a curriculum is one of the most important assets of higher education, some faculties developed curriculum design tools and made it public[15, 6]. In many cases, such tools require teachers to define courses with units of knowledge[9], which takes a lot of time and efforts. Tungare et al. created a repository system for computer science syllabi[14]. They developed tools such as Syllabus-Maker for creating and comparing syllabi. They have not developed a technique to grasp characteristics of a whole curriculum. With our method, faculties have to prepare only course syllabi.

For university students and teachers, only syllabi give the fundamental information about courses. Therefore, there are some researches on syllabi analysis. Ronchetti et al. tried to compare syllabi using Computing Curricula[10] in a similar fashion with ours, though it is still in a preliminary stage.

Mima developed the MIMA search that uses automatic recognition techniques on technical words and clustering words[8]. It generates a graph with words and syllabi as nodes and word-syllabus matrix as arcs. It is useful for browsing local relationships between words and syllabi; however, it lacks

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

<sup>&</sup>lt;sup>1</sup>CS2008 is the interim revision of CS2001.

the global structure visualization function that our approach offers. Ida et al. developed a course classification system using syllabus data[7]. With their tool, users can analyze curricula interactively from various viewpoints. However, it does not automatically give a holistic view of the entire curriculum.

There is a project to compare syllabi of schools among different countries[5]. In this project, the responses of questionnaire from teachers and students are analysed with MCA (Multiple Components Analysis). Our method analyzes course syllabi automatically.

# 3. CURRICULUM STRUCTURE ANALYSIS

# 3.1 Criteria of Curriculum Analysis Method

Improper visualization causes misunderstanding and is harmful. To avoid this problem, we first propose the following criteria which a proper visualization should satisfy.

**Criterion 1:** In order to create a map of a curriculum, we need bases on which courses are represented. Because there are no universal bases, we have to be able to generate bases from a collection of courses.

**Criterion 2:** The relative position in a map should be correct. In other words, if a course is related to two topics A and B, the course is located between these topics in the map.

**Criterion 3:** In order to compare one curriculum with other curriculum, we need to put the two curricula on the same map. It is desirable that we can use one curriculum as a reference and can map course syllabi of other curricula to the reference so that other curricula can be compared on the same reference.

### 3.2 LDA

We use LDA to extract latent topics from course syllabi, and use them as bases required by Criterion 1. LDA is a method used to extract latent topics based on a generative probabilistic model of collections of discrete data, such as text corpora. Among a variety of LDA models, we use the model proposed by Blei[4] because we can adjust how strongly courses are related to topics by some parameters such as number of topics and Dirichlet parameters so that it satisfies Criterion 2.

Given a document-word matrix of text corpora, LDA estimates a set of topics, where each topic is characterized by a distribution over words. In LDA, a *T*-topic LDA model assumes the following generative process of an *N*-word document:

- 1. Choose  $\theta \sim Dirichlet(\boldsymbol{\alpha})$
- 2. For each of the N words  $w_n$ 
  - (a) Choose a topic  $z_n \sim Multinomial(\theta)$
  - (b) Choose a word  $w_n$  from  $p(w_n|z_n,\beta)$ , a multinomial probability conditioned on the topic  $z_n$ .

*T*: number of topics.  $\alpha$ : parameter of Dirichlet distribution.  $\theta$ : topic mixture.  $\beta_{ij} = p(w_n = i|z_n = j)$ : probability distributions of words over topic  $z_n$  (topic-word matrix).  $\boldsymbol{w} = (w_1, \dots, w_N)$ : a document (a sequence of words).  $p(\boldsymbol{w}|z_n)$ : probability distributions of a document over topic  $z_n$  (document-topic matrix).

Using the EM algorithm<sup>2</sup>, LDA performs variational inference of  $\theta$  and *z* for a document, and estimation of the topics  $\beta$ .

The relationship among courses and topics, which is represented by  $p(w|z, \beta)$  in LDA, depends on  $\alpha$ .  $\alpha$  is regarded as a single parameter in Blei's model, though  $\alpha$  is a vector in general. In this paper, the single parameter model is employed. The parameter  $\alpha$  can be set fixed or be estimated by the EM algorithm. If we use smaller  $\alpha$ , LDA relates a course to a few topics strongly, while a course is related to various topics if we use larger  $\alpha$ . Because the relationship among courses and topics directly determines the structure of a map, it is important to use an appropriate  $\alpha$ .

#### **3.3** Syllabus Map Creation by Isomap

For generating a map, we use p(w|z) as coordinates of a syllabus. Because the number of topics in our experiments is higher than 2 or 3, we have to reduce its dimension for visualization. In order to utilize the neighborhood structure of syllabi in the high dimensional structure for visualization, we employ Isomap for the dimension reduction. Isomap is a method used to connect nearby points to form a manifold in a higher dimensional space and unfold it into a low-

dimensional space. Therefore, we can reduce the overlay of clusters to satisfy Criterion 2.

The shape of a map generated by Isomap depends on a distribution of syllabi in an original higher dimensional space. To enable the comparison on the same map, we employ the following procedure.

- 1. Analyzing the reference curriculum  $C^{\text{ref}}$  by means of LDA, we get  $\beta$  estimated by the EM algorithm, which is a probability distribution of words over topics, and also we get  $p^{\text{ref}}(w|z)$ , which is a probability distribution of course syllabi of  $C^{\text{ref}}$  over topic *z*.
- 2.  $v_{LDA}(w^{\text{ref}}) = (p^{\text{ref}}(w^{\text{ref}}|z_1), \cdots, p^{\text{ref}}(w^{\text{ref}}|z_T))$  represents a course syllabus  $w^{\text{ref}}$  in the topic space. We reduce its dimension using Isomap (see Equation 1) and generate a map of  $C^{\text{ref}}$  in the 2D space as follows:

$$\boldsymbol{v}_{iso}(\boldsymbol{w}^{\text{ref}}) = \Pi_{\text{Isomap}}(\boldsymbol{v}_{LDA}(\boldsymbol{w}^{\text{ref}})) \tag{1}$$

where  $\Pi_{\text{Isomap}} : \mathbf{R}^T \to \mathbf{R}^2$  is the Isomap projection.

- 3. Analyzing the test curriculum  $C^{\text{test}}$  by LDA with  $\beta$  acquired at Step 1, we get  $p^{\text{test}}(w|z)$ , which is a probability distribution of course syllabi of  $C^{\text{test}}$  over topic *z*. From  $p^{\text{test}}(w|z)$ , we get  $v_{LDA}(w^{\text{test}})$  which represents a course syllabus  $w^{\text{test}}$  in the topic space.
- 4. For each  $v_{LDA}(w^{\text{test}})$ , pick up  $k_{\text{test}}$ -nearest neighboring syllabi of  $v_{LDA}(w_i^{\text{ref}})(1 \le i \le k_{\text{test}})$  where the distance is given as the Euclidean distance between  $v_{LDA}(w^{\text{test}})$  and  $v_{LDA}(w^{\text{ref}})$ .
- 5. Calculate a projection of  $v_{LDA}(w^{\text{test}})$  using the following equation:

$$v_{iso}(\boldsymbol{w}^{\text{test}}) = \frac{1}{k_{\text{test}}} \sum_{i=1}^{k_{\text{test}}} \Pi_{\text{Isomap}}(\boldsymbol{v}_{LDA}(\boldsymbol{w}_i^{\text{ref}}))$$
 (2)

Using this algorithm, we can plot two curricula in the same map to satisfy Criterion 3.

#### 4. EXPERIMENT

7

<sup>&</sup>lt;sup>2</sup>See [4] for details.



Figure 1: Accuracy of the k-nearest neighbor method for the different number of neighbors k for the number of topics T = 10, 20, 30, 40, and 50.

#### 4.1 **Determination of Hyperparameters**

In order to determine appropriate values for the following hyperparameters: T (number of topics for LDA),  $\alpha$  (parameter of Dirichlet distribution), and  $\hat{k}_{iso}$  (number of neighbors for Isomap) mentioned in Section 3, we estimate the classification accuracy for given hyperparameters by the wellknown k-nearest neighbor method with leave-one-out cross validation (for the details, see some textbooks on machine learning, e.g. [2]). This test shows how near documents which belong to the same category are mapped.

We applied our syllabus map creation method to map the articles of the Twenty Newsgroups Data Set which is available at UCI Machine Learning Repository<sup>3</sup>[1]. For each parameter values, the procedure was repeated ten times and their results were averaged (to alleviate the randomized effects). The average accuracies on the experiments with  $\alpha =$  $1, T = \{10, 20, 30, 40, 50\}, k_{iso} = 5$  are shown in Figure 1. From this result, we determined that the appropriate value of *T* is the same as the number of categories. As to  $\alpha$ , the experiments with  $\alpha = 1$  gave a good result. As to  $k_{iso}$ , the small value of  $k_{iso}$  seemed slightly better though there are no much differences.

#### 4.2 Curriculum Analysis based on CS2008

The Review Task Force commissioned by the ACM Education Board and the IEEE Computer Society's Education finalized CS2008 as a reference curriculum in computer science. The Task Force identified a set of 14 knowledge areas, each of which contains about 10 knowledge units which correspond to syllabi.

We used LDA-C[3] which was a C-implementation of Blei's LDA model. With LDA-C we can compute probability distributions of words over topic  $p(w|z_n)$  of one document set for the analysis of other document sets. Therefore, we can analyze computer science curricula based on CS2008. In order to achieve its compatibility with other curriculum analysis, we should use the LDA topics which are as similar to the knowledge areas proposed by CS2008 as possible. For generating such intended topics, we applied LDA to the description of units augmented with the description of their knowledge areas.

Table 1: CS2008 Knowledge Areas and Topics Topic Knowledge Areas and Numbers of Units GV: 13/13, HC: 2/10, IM: 1/15 1 2 IM: 9/15, HC: 1/10 PF: 8/8, HC: 1/10, IM: 1/15, PL: 1/11 3 IS: 11/11, IM: 2/15, HC: 1/10 4 5 DS: 6/6, HC: 2/10 OS: 12/14 6 7 OS: 2/14, HC: 1/10 8 PL: 10/11 9 SE: 14/14 10 AL: 11/11, IM: 2/15 11 HC: 2/10 12 NC: 9/9, CN: 3/3

The two-letter codes stand for the CS2008 knowledge areas such as: GV: Graphics and Visual Computing, HC: Human-Computer Interaction, IM: Information Management, PF: Programming Fundamentals, IS: Intelligent Systems, DS: Discrete Structures, OS: Operating Systems, PL: Programming Languages, SE: Software Engineering, AL: Algorithms and Complexity, NC: Net Centric Computing, CN: Computational Science, AR: Computer Architecture, and SP: Social and Professional Issues.



Figure 2: The curriculum map of MIT

Table 1 shows the LDA topics and their related knowledge areas. Two numbers after each knowledge area stand for numbers of knowledge units of the area. The former is the number of units with highest  $p(w|z_n)$  for the LDA topic. The latter is the total number of units of each area. If most of the units in a knowledge area are strongly-related to the topic, it is shown in boldface. For example, GV at Topic 1 is in boldface because all the 13 units of GV are strongly-related to Topic 1. The results show that the generated LDA topics were strongly-related to the knowledge areas in CS2008

#### 4.3 Analysis of Syllabi of MIT

We used the computer science-related course syllabi of Massachusetts Institute of Technology (MIT) as a target of analysis. This is because MIT provides many computer science-

- 13 AR: 10/10
- SP: 11/11 14

<sup>&</sup>lt;sup>3</sup>http://archive.ics.uci.edu/ml/



Figure 3: The course chains which start from "18.03 Differential Equations"

related courses and its course syllabi are available at OCW<sup>4</sup>. As computer science-related courses, we picked up 299 courses of the departments of "Electrical Engineering and Computer Science" and "Mathematics" from 1999 to 2008. We extracted a text from "Course Home" and "Syllabus" web pages of each course, and eliminated obviously unnecessary words such as HTML tags, header and footer, stop words, and some frequent words specific to OCW.

All courses of MIT are related to several topics of CS2008. We used the probability p(w|z) as the degree of relation between the syllabus w and the topic z. Figure 2 shows all the course syllabi, each of which is represented as a small circle and a course id in a different color according to the most strongly related topic. Our method satisfies Criteria 1 and 3 because we can extract topics from CS2008 and map course syllabi of the MIT curriculum based on the CS2008 topics.

From Figure 2, we can see the following characteristics of the MIT curriculum. Relatively many courses are provided for Topics 4, 5, and 12 (DS, IS, NC, and CN). This shows the emphasis of the MIT curriculum. On the other hand, a few MIT courses are related to Topic 6 (OS).

Courses related to the same topic are aggregated, and some multi-disciplinary courses lie between their related disciplinary courses. For example, "18.404J/6.840J Theory of Computation" lies between Topic 5 and Topic 10 (DS and AL)" in Figure 2. This feature of the map satisfies Criterion 2.

In order to take some courses of the MIT curriculum, students have to take other "prerequisite" courses beforehand. Some of such prerequisite courses also require the students to take other prerequisite courses. So the prerequisite relationships form a chain of courses in the specific educational area. Figure 3 shows an example of chains, which start from "18.03 Differential Equations." In the figure, each arrow connects





Figure 4: The distributions of distances from target courses to their three closest courses

from prerequisite courses to the target courses. We notice that the arrows extend spirally and deepen to "descendant" courses.

#### 4.4 Analysis of Syllabi of the Open University

Next, we compared the curriculum of MIT with the computer science-related course syllabi of the Open University (OU)<sup>5</sup>. OU is the United Kingdom's only university dedicated to distance learning, and it offers over 600 courses which are categorized into 14 fields. We picked up 55 courses in "Computing and ICT courses". We extracted a text from "Summary" and "Course content" web pages of each course.

Using the algorithm mentioned in Section 3.3, we plotted the course syllabi of OU into the map of MIT generated in Section 4.3.  $k_{\text{test}} = 3$  was employed in this paper, which is determined empirically.

We first calculated the distance to the closest three other courses in the higher dimensional LDA topic space as shown in Figure 4. As to the MIT course syllabi (shown in red "+") the distance between courses peaks at 18. As to OU (shown in green "x"), the peak is 28. As to the closest three MIT courses to each OU course, the distance peaks at 24 and 34 (shown in "\*"). Because these distributions are similar, we can expect the OU courses are properly plotted in the map of the MIT courses.

Figure 5 shows how the OU course, "T837 Systems engineering" was plotted in the map of MIT. In a higher dimensional LDA topic space, three MIT courses, "6.033 Computer System Engineering," "6.163 Strobe Projects Laboratory," and "6.938 Engineering Risk-Benefit Analysis" are closest to T837. These three MIT courses are mapped to nearby points by Isomap, so the OU course is plotted at a point close to these nearby points.

Figure 6 shows all the courses of MIT and OU (white-bordered circles). From this figure, we can see the characteristics of the OU curriculum. Most course syllabi in OU are plotted in the left area in the map of MIT. This means that the many courses in OU are offered for Topic 12 and 14 (NC, CN, and SP), and few courses are for Topic 5 (DS). This practical nature of courses in OU is consistent with the fact that about 70 percent of undergraduates are in full-time employment in OU. The additional data are available at our web site<sup>6</sup>.

<sup>&</sup>lt;sup>5</sup>http://www.open.ac.uk/

<sup>&</sup>lt;sup>6</sup>http://www.sekiya.ecc.u-tokyo.ac.jp/coursedb/



Figure 5: The plot of the OU course T837 in the MIT curriculum map

### 5. SUMMARY

In this paper, we proposed a method with LDA to analyze syllabi and to construct a two-dimensional map from Isomap, so that syllabi could be holistically understood. We applied our methods to CS2008, the curricula of MIT, and these of OU. Characteristics of curriculum of these two universities could be detected using CS2008 as reference. We have been developing a web-based tool for visualizing the map of a curriculum[11].

Now, we are planning to extend our experiments to syllabi from other universities. We are expecting to find similarities and differences among curricula of different universities, and sparse fields which are not sufficiently educated.

#### 6. REFERENCES

- [1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.
- [2] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, New York, NY, USA, 2006.
- [3] D. M. Blei. Lda-c, 2006. http://www.cs.princeton.edu/~blei/lda-c/.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [5] G. S. Carvalho and P. Clément. Construction and validation of the instruments to compare teachers ' conceptions and school textbooks of 19 countries : the european biohead-citizen project. 2007.
- [6] Harvard Graduate School of Education. The Collaborative Curriculum Design Tool (CCDT), 2010. http://learnweb.harvard.edu/ccdt/ (accessed 2010-02-22).
- [7] M. Ida, T. Nozawa, F. Yoshikane, K. Miyazaki, and H. Kita. Syllabus database and web service on higher education. In 7th International Conference on Advanced Communication Technology (ICACT2005), volume 1, pages 415 – 418, 2005.
- [8] H. Mima. Mima search: a structuring knowledge system towards innovation for engineering education. In Proceedings of the COLING/ACL on Interactive



Figure 6: The curriculum map of MIT and OU

presentation sessions, pages 21–24, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

- [9] M. Pedroni, M. Oriol, and B. Meyer. A framework for describing and comparing courses and curricula. In Proceedings of the 12th annual SIGCSE conference on Innovation and technology in computer science education, ITiCSE '07, pages 131–135, New York, NY, USA, 2007. ACM.
- [10] M. Ronchetti and J. Sant. Towards automatic syllabi matching. In *ITiCSE '09: Proceedings of the 14th annual* ACM SIGCSE conference on Innovation and technology in computer science education, page 379, New York, NY, USA, 2009. ACM.
- [11] T. Sekiya, Y. Matsuda, and K. Yamaguchi. Development of a curriculum analysis tool. to appear in ITHET 2010, 9th International Conference on Information Technology Based Higer Education and Training, 2010.
- [12] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [13] The CS2008 Review Taskforce. Computing science curriculum 2008: An interim revision of cs2001, 2008. http://www.acm.org/education/curricula/ ComputerScience2008.pdf.
- [14] M. Tungare, X. Yu, W. Cameron, G. Teng, M. A. Perez-Quinones, L. Cassel, W. Fan, and E. A. Fox. Towards a syllabus repository for computer science courses. In *Proceedings of the 38th SIGCSE technical symposium on Computer science education*, pages 55 – 59, 2007.
- [15] M. S. Wiske, M. Sick, and S. Wirsig. New technologies to support teaching for understanding. *International Journal of Educational Research*, 35(5):483 – 501, 2001.