修士論文の和文要旨

研究科・専攻	大学院　情報システム学研究科　社会知能情報学専攻　博士前期課程		
氏　　名	徐　劍	学籍番号	1351008
論文題目	Biterm Topic Model を用いた e ラーニングコースのレポート分類		

要　旨

近年、Computer Supported Collaborative Learning（CSCL）システムが開発されている。CSCLはコンピュータ技術を利用して、学習コミュニティの中での知識の共有と建設を特徴としている。しかし、CSCLは同時に同一トピックを学習するメンバによって構成される学習コミュニティを支援するので,メンバの熟達レベルの多様性が小さく，他者から学び方や学習成果を共有できる範囲は限定される。

この制限を克服するために、eポートフォリオシステムは提案されている。E-ポートフォリオシステムは長年にわたって学習者の成果や情報を収集することができる。これらのデータから有用な情報を見つけて、他の学習者を助けるために、トピックモデルが適用されているeポートフォリオシステムが提案されている。

トピックモデルは、ドキュメントのコレクションで発生する抽象的な「トピック」を発見するための統計モデルの一種である。Latent Dirichlet Allocation（LDA）は、e ポートフォリオに適用することが提案されている。しかし、LDA はデータがスパースな場合、推定精度が落ちるなどの問題がある。まず、短い文書では、ほとんどの単語が一度だけしか出現しない。つまり、単語の出現頻度から、重要な単語を識別なことが困難である。第二に、多くの単語の意味は、その単語が出現する文脈によって決定される。短い文章では、関連する単語の数によって制限されてきたので、それが曖昧な単語のトピックを識別することは困難である。こんなデータのスパースは、伝統的なトピックモデルの推定精度に影響を与える。この問題に対処するために、Biterm Topic Model（BTM）が提案されている。本研究では、文書分類のための代わりに LDA の BTM を使用するように触発されている。

BTM のパフォーマンスを測定するために、本研究は、e ラーニングシステム"samurai"に蓄積されている学習者レポートを用いた。実験の結果は、1）BTM は LDA より推定したトピックを構成する単語の一貫性が高い。2）BTM は LDA よりトピックの推定精度が高い。

平成 26 年度修士論文

Biterm Topic Model Based Classification of E-learning Course Reports

Biterm Topic Model を用いた e ラーニングコースのレポート分類

電気通信大学大学院情報システム学研究科

社会知能情報学専攻知識創産システム学講座

学籍番号　1351008

氏名　　徐　剣

主任指導教員

植野　真臣　教授

指導教員

田中　健次　教授

田原　康之　准教授

2015 年 1 月 26 日提出

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent years, various Computer Supported Collaborative Learning (CSCL) systems are developed. CSCL uses instructional methods designed to encourage or require students to work together on learning tasks. Collaborative learning is distinguishable from the traditional approach to instruction in which the instructor is the principal source of knowledge and skills. For example, the neologism "e-learning1.0" refers to the direct transfer method in computer-based learning and training systems (CBL). In contrast to the linear delivery of content, often directly from the instructor's material, CSCL uses blogs, wikis, and cloud-based document portals (i.e., Google and Dropbox). With technological Web 2.0 advances, sharing information between multiple people in a network has become much easier and been widely use. [1] One of the main reasons for its usage states that it is "a breeding ground for creative and engaging educational endeavors."[1]

However, CSCL is generally providing support to the learning community of learners that learning the same subjects at the same time. As the diversity of proficiency level of these learners is relatively small, their learning methods and learning outcomes are less helpful for the others from deferent learning community. [2]

In order to overcome the above limitation, the e-portfolio system comes into our sights. The e-portfolio system makes it possible to share the achievements of various learners, such as grades, learning diary and learning history. [3]The e-portfolio has been popular as a tool to facilitate the reflection of learning individuals. Since various learning user data can been collected over many years, the e-portfolio is potentially combines the features of the learning community, which makes it become a useful tool for the learners from other learning communities. [2]

The learner information in the e-portfolio can be effectively used in learning for the others with a study of Ueno-Uto [2]. This system is capable of finding the other learner information useful to the learner system. For example, excellent past learners learned the same topic and it is possible for others to consult them from their e-portfolio such as learning objectives and learning methods. However, there is still a problem that it is difficult for e-portfolio to discover useful data with a large amount of information.

To solve this problem, Ueno [4], have designed and developed a system for recommending useful information to others target learners. The system using a decision tree to recommend the learners with high evaluation of learning process that statistically similar to the other learners, and induces learned from others. Specifically, it is performed results, scores in each quiz, whether writing to the bulletin board, a recommendation by analyzing the statistical data, such as duration of the content. However, this recommendation system uses superficial statistical data (i.e., learned fields, learning frequency, and the level of understanding of each field) for recommendation, the contents of the learning outcomes is not taken into consideration.

Kato [5] proposed to analysis the learning artifacts stored in the e-portfolio (for example, report) and aim to make effective use of it to learning. Specifically, Kato [5] developed a recommendation system to provide learners useful report recommendation. And this system uses the Latent Dirichlet Allocation (LDA) [6] topic model for document classification.

Although LDA is one of successful topic models, it suffers from the data sparse problem on short reports. First, in short reports, most words only occur once. So it is impossible to tell which words are more important from their counts. Second, many words in human language are ambiguous, which their senses are decided by their contexts. In short reports, context is limited with few relevant words. So it is difficult to identify the topics of the ambiguous words. The severe data sparse problem makes conventional topic models less effective on short reports. To address this problem, the Biterm Topic Model (BTM) [7] has been proposed. In this study, we are inspired to use BTM to instead of LDA for document classification.

To measure the performance of BTM, we conducted extensive experiments on learner reports of the learning management system named "samurai", i.e., Experimental results show that 1) BTM can discover more prominent and coherent topics than the LDA. 2) Compared to the LDA, the BTM is much more accurate.

The rest of the paper is organized as follows. In Chapter 2, we give some introductions of e-learning, learner report and our goals. Chapter 3, we give a brief survey of related works. Chapter 4 introduces biterm topic model based classification for e-learning course reports.

Experimental results are presented in Chapter 5. Finally, conclusions are made in the Chapter 6.

Chapter 2
E-leaning and Its Problem

E-learning is an inclusive term that describes educational technology that electronically or technologically supports learning and teaching. Bernard Luskin, a pioneer of e-learning, advocates that the "e" should be interpreted to mean "exciting, energetic, enthusiastic, emotional, extended, excellent, and educational" in addition to "electronic." This broad interpretation focuses on new applications and developments, and also brings learning and media psychology into consideration.[8] Parks suggested that the "e" should refer to "everything, everyone, engaging, easy".[9]

E-Learning does not just mean distance education, online education can also play an important role in the traditional teaching on campus, also in the remote network education, some conventional teaching methods and teaching methods is also very important. E-Learning completely replace the traditional classroom is not realistic, traditional classroom teaching in imparting knowledge, social, interactive aspects have a huge advantage.

E-Learning to enter the campus, not a substitute for traditional teaching style to enter, but continue to collide with traditional teaching, the gradual integration of the collision, the fusion constantly replenished and improved, forming an effective and feasible in practice under the IT environment teaching methodology.

E-Learning cannot completely replace face to face learning, but in danger of being marginalized. The reason is that e-Learning can only solve part of the process of learning issues. And in the learning effect, since it creates a lack of classroom teaching effectiveness and positive interaction, the learning effect will be greatly reduced.

2.1 E-learning Course

A massive open online course (MOOC) [10] is a kind of e-learning courses aimed at unlimited participation and open access via the web. In addition to traditional course materials such as filmed lectures, readings, and problem sets, many MOOCs (i.e., Coursera, edX) provide interactive user forums to support community interactions between students, professors, and teaching assistants (TAs). MOOCs are a recent development in distance education which was first introduced in 2008 and emerged as a popular mode of learning in 2012.[11][12]

Early MOOCs often emphasized open-access features, such as open licensing of content, structure and learning goals, to promote the reuse and remixing of resources. Some later MOOCs use closed licenses for their course materials while maintaining free access for students. [13][14][15]

Many MOOCs use video lectures, employing the old form of teaching using a new technology.[16] Thrun testified before the President's Council of Advisors on Science and Technology (PCAST) that MOOC "courses are 'designed to be challenges,' not lectures, and the amount of data generated from these assessments can be evaluated 'massively using machine learning' at work behind the scenes. This approach, dispels 'the medieval set of myths' guiding teacher efficacy and student outcomes, and replaces it with evidence-based, 'modern, data-driven' educational methodologies that may be the instruments responsible for a 'fundamental transformation of education' itself". [17]

Because of massive enrollments, MOOCs require instructional design that facilitates large-scale feedback and interaction. The two basic approaches are:

- Peer-review and group collaboration.

- Automated feedback through objective, online assessments, e.g. quizzes and exams.

Assessment can be the most difficult activity to conduct online, and online assessments can be quite different from the bricks-and-mortar version.[18] Special attention has been devoted to proctoring and cheating.[19]The two most common methods of MOOC assessment are machine-graded multiple-choice quizzes or tests and peer-reviewed written assignments.[18]Machine grading of written assignments is also underway.[20]Peer review is often based upon sample answers or rubrics, which guide the grader on how many points to award different answers. These rubrics cannot be as complex for peer grading as for teaching assistants. Students are expected to learn via grading others [21] and become more engaged with the course. [22] Exams may be proctored at regional testing centers. Other methods, including "eavesdropping technologies worthy of the C.I.A." allow testing at home or office, by using webcams, or monitoring mouse clicks and typing styles.[19]Special techniques such as adaptive testing may be used, where the test tailors itself given the student's previous answers, giving harder or easier questions accordingly.

Course delivery involves asynchronous access to videos and other learning material, exams and other assessment, as well as online forums. Although MOOCs provide interactive learners forums to support community interactions between students, professors, and teaching assistants (TAs), learners’ learning outcomes are not taken used for recommendations.

2.2 Learner Report

Learner report is one kind of learners' learning outcomes. Sometimes the assessment of course is based on the reports proposed by learners. The reports can be collected from learners who take the same course or in the same learning community by LMS (Learning Management System). We will give the example of learner report and introduce the process of learner reports in Chapter 6.

2.3 Goals

In this study, the learner reports are collected from the lecture of graduate school by LMS (Learning Management System) called "Samurai" developed by Ueno [2], [23] ~ [27]. In addition, we are inspired to make use of these actual data for text analysis based on biterm topic model (BTM) in order to prove the possibility to improve the recommendation accuracy.

In summary, our goals are mainly to apply BTM to classify the reports with fewer contents and to improve the accuracy of the classification.

Chapter 3
Related Works

In this section, we briefly summarize the related works from the following two perspectives: topic model, report recommendation.

3.1 Topic Model

Topic models are widely used to uncover the latent semantic structure from text corpus. The effort of mining the semantic structure in a text collection can be dated from latent semantic analysis (LSA) [17], which employs the singular value decomposition to project documents into a lower dimensional space, called latent semantic space. Probabilistic latent semantic analysis (PLSA) [6] improves LSA with a sound probabilistic model based on a mixture decomposition derived from a latent class model. In PLSA, a document is represented as a mixture of topics, while a topic is a probability distribution over words. Extending PLSA, Latent Dirichlet Allocation (LDA) [7] adds Dirichlet priors for the document-specific topic mixtures, making it possible to generate unseen documents. Due to its nice generalization ability and extensibility, LDA has achieved huge success in text mining.

The sparse content in short texts brings new challenges to topic modeling. To address this question, Yan [7] propose a generative biterm topic model (BTM), which learns topics over short texts by directly modeling the generation of biterms in the whole corpus. Compared to conventional topic models, the major differences and advantages of BTM lay in that 1) BTM models the word co-occurrence patterns (i.e., biterms) explicitly, rather than implicitly (via document modeling), to enhance topic learning; and 2) BTM uses the aggregated word co-occurrence patterns in the corpus for topic discovering, which avoids the problem of sparse patterns at document level.

3.2 Report Recommendation

Kato [5], have developed a report recommendation function within e-portfolio system based on LDA. Specifically, it estimates the potential topic of the report by using LDA, and recommends reports in e-portfolio based on the result generalized by LDA. This recommendation system is expected to provide support for learners to write a high quality report.

In addition, the topic similarity compared to superficial word similarity is more important. By recently proposed LDA, it has become possible to automatically classify a potential theme or topic sentence. So it is also possible to recommend a report with similar topics estimated by LDA. In this case, it is desirable that the report recommendations are variable.

Chapter 4
Biterm Topic Model Based Classification of E-learning Course Reports

Before we detail the model, we first introduce the notation of “biterm”, which denotes an unordered word pair co-occurring in a short context (i.e., an instance of word co-occurrence pattern). Here a short context refers to a small, fixed-size window over a term sequence. In short texts with limited document length, such as tweets and text messages, we can simply take each document as an individual context unit. In such case, any two distinct words in a document construct a biterm. For example, a document with three distinct words will generate three biterms:

$$\{w1; w2; w3\} \quad \Rightarrow \quad \{(w1; w2); (w2; w3); (w1; w3)\}$$

where (word; word) is unordered. After extracting biterms in each document, the whole corpus now turns into a biterm set. The biterm extraction process can be completed via a single scan over the documents.

4.1 Biterm Topic Model

Unlike most topic models that learn the latent topic components in a corpus by modeling the generation of documents, BTM performs this task by modeling the generation of biterms. The key idea is that if two words co-occur more frequently, they are more likely to belong to a same topic. Based on this idea, it assumes that the two words in a biterm are drawn independently from a topic, where a topic is sampled from a topic mixture over the whole corpus.

Given a corpus with $\mathrm{N_D}$ documents, suppose it contains $\mathrm{N_B}$ biterms $\mathbf{B}=\{\mathrm{b_i}\}_{\mathrm{i=1}}^{\mathrm{N_B}}$with $\mathrm{b_i}= (\mathrm{w_{i,1}}, \mathrm{w_{i,2}})$, and **K** topics expressed over **W** unique words in the vocabulary. Let $\mathbf{z}\in$ [1; **K**] be a topic indicator variable, we can represent the prevalence of topics in the corpus (i.e.,P(z)) by a K-dimensional multinomial distribution $\boldsymbol{\theta}=\{\theta_\mathrm{k}\}_{\mathrm{k=1}}^{\mathbf{K}}$with θ_k =P(z=k) and $\sum_{\mathrm{k=1}}^{\mathbf{K}} \theta_\mathrm{k} = 1$. The word distribution for topics (i.e.,P(w|z)) can be represented by a **K** × **W** matrix **Φ** where the kth row ϕ_k is a W-dimensional distribution $\phi_{\mathrm{k,w}} = \mathrm{P(\,w \mid z = k\,)}$ with entry $\sum_{\mathrm{w=1}}^{\mathbf{W}} \phi_{\mathrm{k,w}} = 1$.

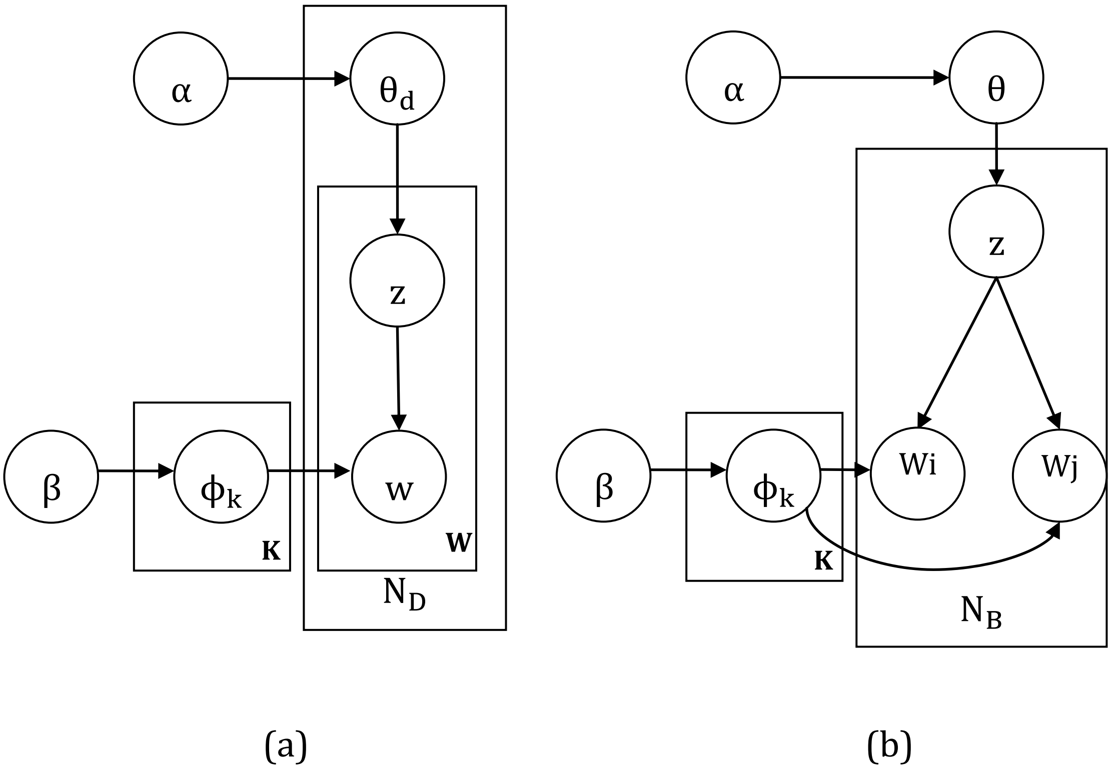


Figure 4.1 Graphical representation of (a) LDA and (b) BTM.

Each node in the graph denotes a random variable, where shading represents an observed variable. A plate denotes replication of the model within it. The number of replicates is given in the bottom right corner of the plate.

Following the convention of LDA [30], we use symmetric Dirichlet priors for $\boldsymbol{\theta}$ and ϕ_k with single-valued hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Formally, the generative process of BTM is described as follows.

1) Draw $\boldsymbol{\theta}$ Dirichlet($\boldsymbol{\alpha}$);

2) For each topic $k \in [1; \mathbf{K}]$:

a) To draw ϕ_k;

3) For each biterm $b_i \in \mathbf{B}$:

a) To draw $z_i \sim$ Multinomial($\boldsymbol{\theta}$);

b) To draw $w_{i,1}$, $w_{i,2} \sim$ Multinomial(ϕ_{z_i});

Its graphical representation is shown in Figure 1(b).Note that it assumes that the biterms are generated independently for simplicity.

Following the above procedure, we can write the probability of biterm b_i conditioned on the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$:

$$\mathrm{P}\ (b_i|\boldsymbol{\theta}, \boldsymbol{\Phi}) = \sum_{k=1}^{K} \mathrm{P}(\ w_{i,1}, w_{i,2}, z_i = k \mid \boldsymbol{\theta}, \boldsymbol{\Phi})$$

$$= \sum_{k=1}^{K} \mathrm{P}(\ z_i = k \mid \theta_k)\mathrm{P}\left(\ w_{i,1} \,\middle|\, z_i = k, \phi_{k,\, w_{i,1}}\right) \cdot$$

$$\mathrm{P}(\ w_{i,2} \mid z_i = k, \phi_{k,\, w_{i,2}}).$$

$$= \sum_{k=1}^{K} \theta_k\ \phi_{k,\, w_{i,1}} \phi_{k,\, w_{i,2}}. \qquad (1)$$

Given the hyper parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we can obtain the probability of b_i by integrating over $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$:

$$\mathrm{P}\ (b_i|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \iint \sum_{k=1}^{K} \theta_k\ \phi_{k,\, w_{i,1}} \phi_{k,\, w_{i,2}}\ d\boldsymbol{\theta}\ d\boldsymbol{\Phi}\ . \qquad (2)$$

Taking the product of the probability of single biterms, we obtain the likelihood of the whole corpus:

$$\mathrm{P}\ (\mathbf{B}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{B} \iint \sum_{k=1}^{K} \theta_k\ \phi_{k,\, w_{i,1}} \phi_{k,\, w_{i,2}}\ d\boldsymbol{\theta}\ d\boldsymbol{\Phi}. \qquad (3)$$

For better understanding the uniqueness of BTM, we compare it with one typical model for topic learning, i.e., LDA [7]. In literature, LDA has been employed for topic discovering over short texts [1], [2],

[31], and [26]. Figure 1 shows the graphical representation of the two models.

LDA, illustrated in Figure 1(a), models the generation of a document d as follows: For each word in $\mathbf{d}$, we first draw a topic $\mathbf{z}$ from the document-specific topic distribution θ_d, and then draw a word $\mathbf{w}$ from topic $\mathbf{z}$. From this figure, we can see that the topic $\mathbf{z}$ of word $\mathbf{w}$ depends on the other words in the same document through sharing the topic distribution θ_d. Hence, LDA excessively relies on the document-level context for the inference of $\mathbf{z}$ and θ_d. It makes LDA susceptible to the data sparsity problem when documents are short, resulting in poor estimation of $\mathbf{z}$ and θ_d, in turn, hurting the learning of the topic-word distributions $\mathbf{\Phi}$.

In a word, the major trouble of LDA lies in modeling the short documents improperly. For such extremely sparse data, it is difficult to directly model and infer the latent topics in single short documents. However, we argue that it is not necessary to model documents for topic discovering in a corpus. BTM, illustrated in Figure 1(b), just chooses another way to discover topics by modeling the generation of biterms, rather than documents. Compared to LDA, BTM avoids the data sparse problem by learning a global topic distribution $\boldsymbol{\theta}$.

4.2 Parameter Estimation

Similar to LDA, it is intractable to exactly solve the coupled parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ by maximizing the likelihood in Eq. (3). Following [30], we conduct approximate inference for $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ using Gibbs sampling [14], which estimates the parameters using samples drawn from the posterior distributions of latent variables sequentially conditioned on the current values of all other variables and the data.

In the setting of BTM, there are three types of variables (i.e., the topic assignments of $\mathbf{z}$, the multinomial distribution parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$) to be estimated. But using the technique of collapsed Gibbs sampling [32], $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ can be integrated out due to the use of conjugate priors. Thus, for biterm b_i, we only need to sample its topic z_i according the following conditional distribution (the derivation is provided in the supplemental material):

$$P(z_i = k \mid \mathbf{z}_{-i}, \mathbf{B}) \propto (n_{-i,k} + \boldsymbol{\alpha})\frac{(n_{-i,w_{i,1}|k}+\boldsymbol{\beta})(n_{-i,w_{i,2}|k}+\boldsymbol{\beta})}{(n_{-i,.|k}+\mathbf{W}\boldsymbol{\beta}+1)(n_{-i,.|k}+\mathbf{W}\boldsymbol{\beta})} \quad . \qquad (4)$$

where z_{-i} denotes the topic assignments for all biterms except b_i, $n_{-i,k}$ is the number of biterms assigned to topic k excluding b_i, $n_{-i,w|k}$

is the number of times word w assigned to topic k excluding *bi*, and $n_{-i,.|k}=\sum_{w=1}^{\mathbf{W}} n_{-i,w|k}$. The right hand of Eq. (4) is quite intuitive: the first factor is proportional to the probability of topic k in the corpus, and the second part expresses the product of the probabilities of $w_{i,1}$ and $w_{i,2}$ under topic k.

We summarize the overall procedure of Gibbs sampling: Firstly, we randomly assign a topic to each biterm as the initial state. In each iteration, we update the topic assignment for each biterm by examining Eq. (4) sequentially. After a sufficient number of iterations, we count the number of biterms in each topic k, denoting by n_k, and the number of times that each word *w* assigned to topic k, denoting by $n_{w|k}$. These counts are used to estimate $\mathbf{\Phi}$ and $\boldsymbol{\theta}$ as follows (the derivation is presented in the supplemental material):

$$\phi_{k,w}=\frac{n_{w|k}+\boldsymbol{\beta}}{n_{.|k}+\mathbf{W}\boldsymbol{\beta}}. \tag{5}$$

$$\theta_k=\frac{n_k+\boldsymbol{\alpha}}{N_{\mathbf{B}}+\mathbf{K}\boldsymbol{\alpha}}. \tag{6}$$

4.3 Topic Inference

Besides learning the topic components (i.e.,$\left\{\phi_k\right\}_{k=1}^{\mathbf{K}}$), another common task in topic models is to infer the topics in a document, i.e., evaluating the topic posterior P (z|d) for document d. However, as BTM does not model documents, we cannot directly obtain P (z|d) from the estimated model. Fortunately, we can derive the topic proportion of a document via the topics of biterms.

Suppose d contains N_d biterms,$\left\{b_i^{(d)}\right\}_{i=1}^{N_d}$, using the chain rule we have

$$\mathrm{P}\,(\mathrm{z|d}) = \sum_{i=1}^{N_d} \mathrm{P}(\mathrm{z}, b_i^{(d)} \mid \mathrm{d}) \;=\; \sum_{i=1}^{N_d} \mathrm{P}(\mathrm{z} \mid b_i^{(d)}, \mathrm{d})\mathrm{P}(b_i^{(d)} \mid \mathrm{d})\,. \quad (7)$$

Given biterm $b_i^{(d)} = (w_{i,1}^{(d)} \;; w_{i,2}^{(d)})$, we assume its topic z is conditionally independent of d, i.e, $\mathrm{P}(\mathrm{z}, b_i^{(d)} \mid \mathrm{d}) = \mathrm{P}(\mathrm{z} \mid b_i^{(d)})$. Then, we can simplify the above equation:

$$\mathrm{P}(\mathrm{z} \mid \mathrm{d}) \;=\; \sum_{i=1}^{N_d} \mathrm{P}(\mathrm{z} \mid b_i^{(d)})\mathrm{P}(b_i^{(d)} \mid \mathrm{d})\,. \qquad (8)$$

In Eq. (8), $\mathrm{P}(\mathrm{z} \mid b_i^{(d)})$ can be calculated via Bayes' formula based on the parameters learned in BTM:

$$P(z = k \mid b_i^{(d)}) = \frac{\theta_k \phi_{k,w_{i,1}^{(d)}} \phi_{k,w_{i,2}^{(d)}}}{\sum_{k'} \theta_{k'} \phi_{k',w_{i,1}^{(d)}} \phi_{k',w_{i,2}^{(d)}}} \qquad (9)$$

Meanwhile, $P\,(b_i^{(d)} \mid d)$ can be estimated empirically:

$$P\,(b_i^{(d)} \mid d) = \frac{n(b_i^{(d)})}{\sum_{i=1}^{N_d} n(b_i^{(d)})}$$

where $n(b_i^{(d)})$ is the frequency of biterm $b_i^{(d)}$ in d.

Chapter 5
Experiments

In this study, we use the reports from the actual course of e-leaning system "samurai ".The name of the course is **Knowledge Computing and Building 2.**The contents of the course contains the knowledge management and the knowledge of statistics foundation.

5.1 Experimental Settings

5.1.1 Datasets

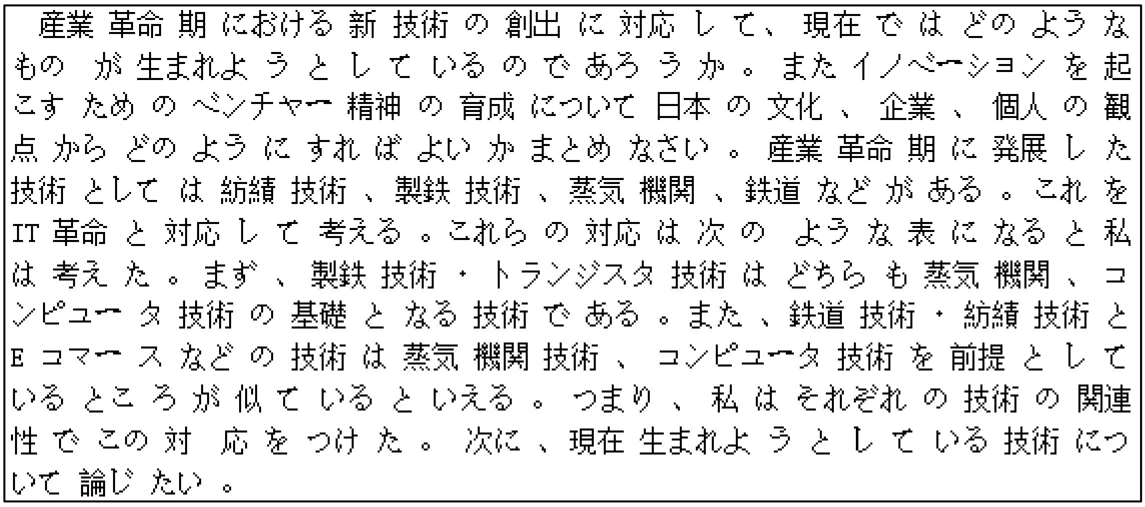
In order to show the performance of our approach over different on reports with different length, we use the learner reports collection for evaluation. The features of learner reports and main parameters are shown in the Table 5.1.

Table 5.1 The features of learner reports and main parameters

Number of reports	90
Number of topics	10
α	0.5
β	0.01
Number of words	5436
Average length of reports	311.8

5.1.2 Processing of Learner Reports

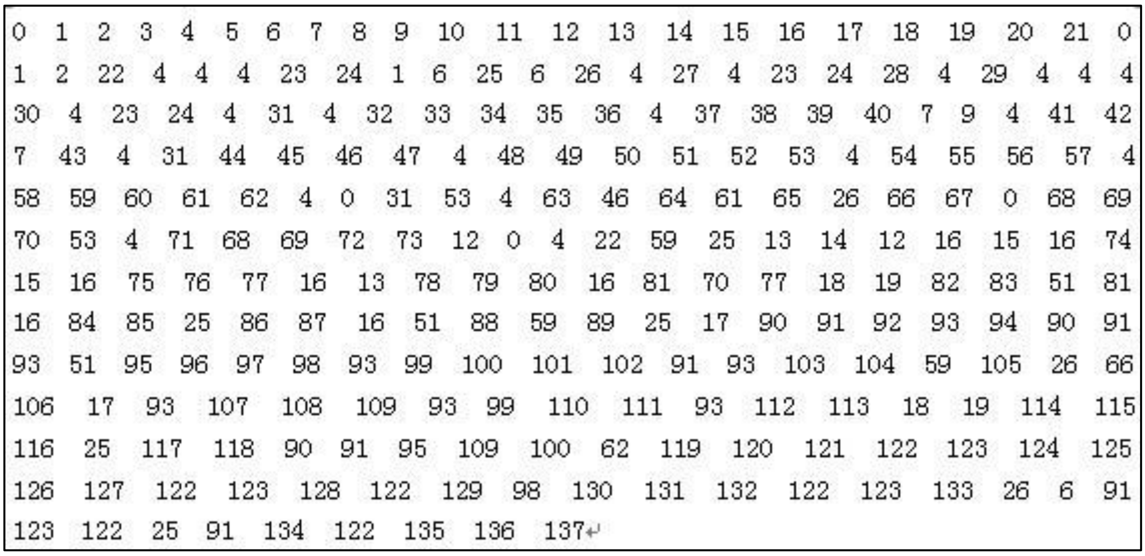
Figure 5.1 is the example of learner report collected from the actual course of e-leaning system "samurai ".



　産業 革命 期 における 新 技術 の 創出 に 対応 し て、 現在 で は どの よう な もの が 生まれよ う と し て いる の で あろ う か 。 また イノベーション を 起こす ため の ベンチャー 精神 の 育成 について 日本 の 文化 、 企業 、 個人 の 観点 から どの よう に すれ ば よい か まとめ なさい 。 産業 革命 期 に 発展 し た 技術 として は 紡績 技術 、製鉄 技術 、蒸気 機関 、鉄道 など が ある 。これ を IT 革命 と 対応 し て 考える 。これら の 対応 は 次 の よう な 表 に なる と 私 は 考え た 。 まず 、 製鉄 技術 ・ トランジスタ 技術 は どちら も 蒸気 機関 、 コンピュー タ 技術 の 基礎 と なる 技術 で ある 。また 、鉄道 技術 ・ 紡績 技術 と E コマー ス など の 技術 は 蒸気 機関 技術 、 コンピュータ 技術 を 前提 と し て いる とこ ろ が 似 て いる と いえる 。 つまり 、 私 は それぞれ の 技術 の 関連性 で この 対 応 を つけ た 。 次に 、現在 生まれよ う と し て いる 技術 について 論じ たい 。

Figure 5.1 Example of learner report

For preprocessing, we removed meaningless words such as stop words, low frequency words, and characters not in Latin or Japanese. Figure 5.2 shows the example of learner report after preprocessing.



```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 0
1 2 22 4 4 4 23 24 1 6 25 6 26 4 27 4 23 24 28 4 29 4 4 4
30 4 23 24 4 31 4 32 33 34 35 36 4 37 38 39 40 7 9 4 41 42
7 43 4 31 44 45 46 47 4 48 49 50 51 52 53 4 54 55 56 57 4
58 59 60 61 62 4 0 31 53 4 63 46 64 61 65 26 66 67 0 68 69
70 53 4 71 68 69 72 73 12 0 4 22 59 25 13 14 12 16 15 16 74
15 16 75 76 77 16 13 78 79 80 16 81 70 77 18 19 82 83 51 81
16 84 85 25 86 87 16 51 88 59 89 25 17 90 91 92 93 94 90 91
93 51 95 96 97 98 93 99 100 101 102 91 93 103 104 59 105 26 66
106 17 93 107 108 109 93 99 110 111 93 112 113 18 19 114 115
116 25 117 118 90 91 95 109 100 62 119 120 121 122 123 124 125
126 127 122 123 128 122 129 98 130 131 132 122 123 133 26 6 91
123 122 25 91 134 122 135 136 137↵
```

Figure 5.2 Example of learner report after preprocessing

At the same time, we also create the dictionary after the words numbered. Table 5.2 shows the dictionary of words.

Table 5.2 The dictionary of words

The dictionary of words	
Index	**Words**
0	産業
1	革命
2	期
3	新
4	技術
5	創出

After preprocessing, we can extract biterms. Table 5.3 shows the data format of the biterms. The format contains three parts: First part is the Id of each report since the biterms are extracted in each report and the whole corpus turns into a biterm set. The biterm extraction process can be completed via a single scan over the reports; Second part is a pair of words named biterm contains two words not unordered; Last part is counts of the same biterms in each report named frequency.

Table 5.3 The data format of biterms

The format of biterms			
Report Id	**Biterms(wi,wj)**		**Frequency**
1	0	10	1
1	0	11	1
2	0	10	1
2	0	11	2
3	0	10	4
3	0	11	3

In order to evaluate the BTM, we need to calculate topic rates of reports estimated by BTM while the topic number is 10.

Table 5.4 and Table 5.5 show the top 10 words under 10 topics.

Table 5.6 shows topic rates of reports.

Table 5.7 shows the possibility of each word generalized in topic 6.

Table 5.4 The top 10 words of topic 0~topic 4

Topic 0	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**
年	C	learning	情報	社会
システム	リエンジニアリング	information	ベンチャー	情報
日本	プロセス	japanese	社会	ベンチャー
ため	フォード	century	れる	技術
商品	ため	innovations	産業	育成
向上	書類	eventually	ため	れる
マクドナルド	人	industry	技術	イノベーション
化	システム	world	概念	企業
導入	書	companies	企業	ため
http	クレジット	market	化	創出

Table 5.5 The top 10 words of topic 5~topic 9

Topic 5	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**
労働	産業	リエンジニアリング	企業	経営
科学	革命	会社	技術	評価
作業	企業	ため	もの	化
テーラー	技術	商品	失敗	企業
化	日本	プロセス	化	システム
実践	社会	化	ため	bsc
という	ため	情報	コスト	制度
システム	ベンチャー	市場	ザッポス	という
標準	イノベーション	年	できる	ため
仕事	られる	業務	れる	コスト

Table 5.6 Topic rates of reports with the number of topics is 10

Topic rates of reports	
Topic0	0.0139144
Topic1	0.0180729
Topic2	0.00826359
Topic3	0.188758
Topic4	0.0513287
Topic5	0.0308924
Topic6	0.664545
Topic7	0.0130962
Topic8	0.0111291
Topic9	1.32015e-07

Table 5.7 The possibility of each word generalized in topic 6

Topic6	
産業	0.023277
革命	0.020134
企業	0.017637
技術	0.016695
日本	0.011878
社会	0.011130
ため	0.010477
ベンチャー	0.010028
イノベーション	0.008412
られる	0.008088

5.1.3 Measures and Methodology

We aim to evaluate the effectiveness and efficiency of BTM on learner reports. Note that the evaluation of effectiveness of a topic model is not a trivial problem. A typical metric is the perplexity or marginal likelihood evaluated on a held-out test set [6], [28], [29], but it is not suitable for us for two reasons. First, the marginal likelihoods of LDA and BTM are not comparable, since LDA optimizes the likelihood of word occurrences in documents, while BTM optimizes the likelihood of biterm occurrences in the corpus. Second, these metrics disconnect with our expectations of topic models [30], e.g., the interpretability of topics and usefulness in real applications. It is argued that topic models with better held-out likelihood may infer less semantically meaningful topics [31]. Considering that we are often interested in two parts of the results of topic models, i.e., the topic components and documents' topic proportions, we would like to evaluate the quality of them separately.

In recent years, some automatic evaluation methods are proposed to measure the quality of the topics discovered. One is the *coherence score* [32], which says that a topic is more coherent if the most probable words in it co-occurring more frequently in the corpus.

This idea is consistent with the basic assumption of BTM, i.e., words co-occurring more frequently should be more possible to belong to a same topic. Thus it is not surprising that BTM always obtains better coherence scores than the baselines [7]. Another popular metric for automatic evaluation is the PMI-Score [33], which measures the coherence of a topic based on point wise mutual information using large scale text datasets from external sources, e.g., Wikipedia and Baike8. Since these external datasets are model-independent, PMI-Score is fair for all the topic models. Therefore, we exploit PMI-Score to verify the topic quality. Given the **T** most probable words of a topic k, (w_1,..., w_T), PMI-Score measures the pair wise association between them:

$$\text{PMI-Score (k)} = \frac{1}{\mathbf{T}(\mathbf{T}-1)} \sum_{1\le i\le j\le \mathbf{T}} \mathrm{PMI}(w_i, w_j)$$

where $\mathrm{PMI}(w_i, w_j) = \log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}$, $P(w_i, w_j)$ and $P(w_i)$ are the probabilities of co-occurring word pair (w_i, w_j) and word w_i estimated empirically from the external datasets, respectively. For evaluation, we compute the PMI-Score using learner reports.

To measure the quality of the documents' topic proportions, we use document classification to see how accurate and discriminative of the learned topical representations from different models are. For each document d, its topical representation is a vector $[\mathrm{P}(\ z = 1 \mid \mathrm{d}\,), \dots, \mathrm{P}(\ z = \mathbf{K} \mid \mathrm{d}\,)]$. We randomly split the dataset into training and test subsets with the ratio 4: 1, and employed the linear SVM classifier for classification with 5-fold cross validation.

5.2 Evaluation of the BTM

In this section, we empirically evaluate the quality of topics, report classification and clustering of BTM. We take one typical topic model as our baseline method, namely LDA.

5.2.1 Topic Coherence

To evaluate the quality of topics discovered, we calculated the average PMI-Score, i.e., $\frac{1}{\mathbf{K}} \sum_{\mathrm{k}} \mathrm{PMI} - \mathrm{Score(k)}$ for BTM and LDA. Table 5.8 lists the results on learner reports with the number of most probable words **T** ranging from 5 to 10. We find that the PMI-Scores of BTM outperform LDA consistently. The results show that BTM can discover more coherent topics than the other three methods.

Table 5.8 Average PMI-Scores of BTM and LDA

Number of topics	10	
Topic Models	Top 5 words	Top 10 words
BTM	2.35±0.05	1.87±0.03
LDA	2.16±0.05	1.72±0.03

5.2.2 Document Classification

We further compare the classification performance of BTM and LDA. Considering topic model as a way for dimensionality reduction, which reduces a document to a fixed set of topical features P (z|d), we would like to see how accurate and discriminative of the topical representation of documents for classification. We reported the accuracy in Figure 5.3.

From the results, we can see that BTM always dominates the LDA. Moreover, the advantage of BTM becomes more notable as the topic number K grows. That is because when the number of topics is small, topics discovered are usually very general. In such case, a report is more likely to belong to a single topic. In contrast, with the increase of the topic number K, BTM will learn more specific topics. At the same time, a large topic number will aggravate the data sparse problem of LDA by introducing more parameters, thus the gap between BTM and LDA also increases.

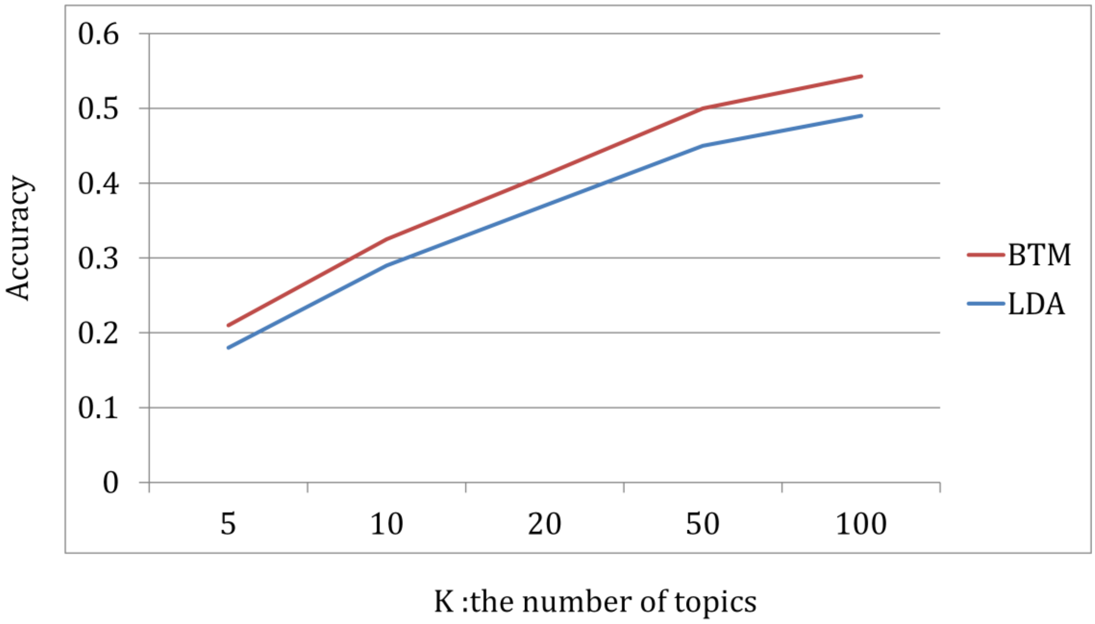


Figure 5.3 Classifying performance of BTM and LDA

5.2.3 Document Clustering

For quantitative evaluation, we compare the clustering performance of BTM and LDA. Document clustering evaluation is a direct way to measure the effectiveness of a topic model without depending on any extrinsic methods. For document clustering, we take each topic as a cluster, and assign each document d to the topic z with highest value of conditional probability P (z|d).

We adopt purity in clustering evaluation as follows. Let **Ω** = (w_1,..., $w_{\mathbf{K}}$) be the set of output clusters, and **C** = (c_1,..., $c_{\mathbf{P}}$)be **P** labeled classes of the documents.

Purity: Suppose documents in each cluster should take the dominant class in the cluster. Purity is the accuracy of this assignment measured by counting the number of correctly assigned documents and divides by the total number of test documents. Formally:

$$\text{Purity}\ (\boldsymbol{\Omega}, \mathbf{C}) = \frac{1}{n} \sum_{i=1}^{\mathbf{K}} \max_{j} \mid w_i \cap c_j \mid$$

Note that when all the documents in each cluster are with the same class, purity is highest with value of 1.Conversely; it is close to 0 for bad clustering.

In this experiment, we separate the reports into three groups according to their length. The groups are listed in Table 5.9.

Table 5.9 The groups setting for clustering

	Group A	Group B	Group C
Number of reports	55	35	90
Length of reports	＜311	＞311	Average length =311

The results are shown in Figure 5.4. On the whole, it is clear that BTM outperforms LDA significantly. As the length of reports increases, more word co-occurrence patterns are included, which improves the performance of BTM substantially.

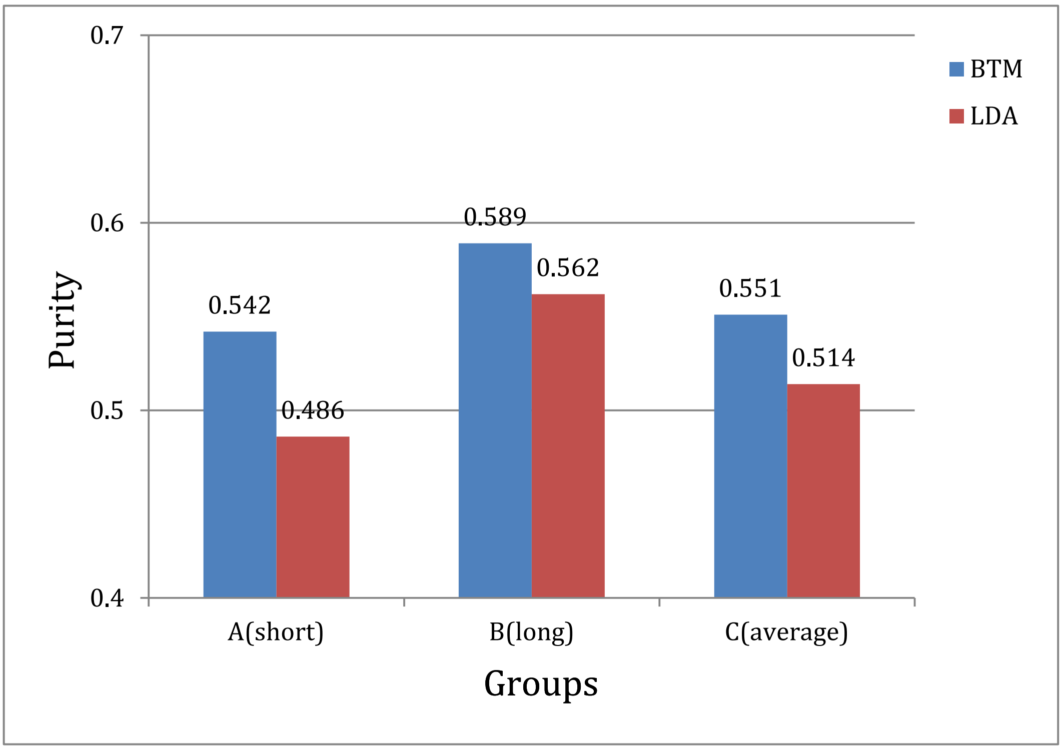


Figure 5.4 Clustering performance of BTM and LDA

Chapter 6
Conclusions

In this study, we apply the BTM to classification on learner reports. The results show that BTM can discover more prominent and coherent topics than the LDA. And compared to the LDA, the BTM are much more accurate. In addition, BTM can performance much better to the reports with fewer contents than LDA.BTM can be an effective topic model in reports recommendation.

In future, the reports recommendation system based on BTM could be developed.

References

[1] Crane,Beverly E. "Using Web 2.0 Tools in the K-12 Classroom". Neal-Schuman Publishers, Inc., 2009

[2] 真臣植野，雅輝宇都，“他者からの学びを誘発する e ポートフォリオ (¡特集¿新時代の学習評価)，日本教育工学会論文誌，vol.35，no.3，pp.169–182，Dec 2011.

[3] 康彦森本，“e ポートフォリオの理論と実際，教育システム情報学会誌，vol.25，no.2，pp.245–263，2008．http://ci.nii.ac.jp/naid/40016344499/

[4] 植野真臣，“他者からの学びを誘発する e ポートフォリオ推薦システムの実践， 日本教育工学会，pp.··–··，2013．

[5] Yoshihiro KATO, Takatoshi ISHII and Maomi UENO, e-Portfolio System with Reports Recommender Function Based on LDA.

[6] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, pp.··–··, 2003.

[7]Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A Biterm Topic Model for Short Texts. In Proceedings of the 22nd international conference on World Wide Web, WWW'13, pages 1445-1456, Rio de Janeiro, Brazil, 2013, ACM.

[8] "Think "Exciting": E-Learning and the Big "E"". Retrieved 8 September 2014.

[9] Eric Parks. "What's the "e" in e-Learning?". Askinternational.com. Retrieved 2013-10-22.

[10] http://en.wikipedia.org/wiki/Massive_open_online_course

[11] Pappano, Laura. "The Year of the MOOC". The New York Times. Retrieved 18 April 2014.

[12] Lewin, Tamar (20 February 2013). "Universities Abroad Join Partnerships on the Web". New York Times. Retrieved 6 March 2013.

[13] Wiley, David. "The MOOC Misnomer". July 2012.

[14] Cheverie, Joan. "MOOCs an Intellectual Property: Ownership and Use Rights". Retrieved 18 April 2013.

[15] David F Carr (20 August 2013). "Udacity hedges on open licensing for MOOCs". Information Week. Retrieved 21 August 2013

[16] Shirky, Clay (8 July 2013). "MOOCs and Economic Reality". The Chronicle of Higher Education. Retrieved 8 July 2013.

[17] "Librarians and the Era of the MOOC". Nature.com. 9 May 2013. Retrieved 11 May 2013.

[18] Degree of Freedom – an adventure in online learning, MOOC Components – Assessment, 22 March 2013.

[19] Eisenberg, Anne (2 March 2013). "Keeping an Eye on Online Test-Takers". New York Times. Retrieved 19 April 2013.

[20] Rivard, Ry (19 April 2013). "EdX Rejected". Inside Higher Education. Retrieved 22 April 2013.

[21] Wong, Michael (28 March 2013). "Online Peer Assessment in MOOCs: Students Learning

from Students". Centre for Teaching, Learning and Technology Newsletter. University of British Columbia. Retrieved 20 April 2013.

[22] P. Adamopoulos, "What Makes a Great MOOC? An Interdisciplinary Analysis of Student Retention in Online Courses," ICIS 2013 Proceedings (2013) pp. 1–21 in AIS Electronic Library (AISeL)

[23] M. Ueno, "Data mining and text mining technologies for collaborative learning in an ilms"samurai"," Proceedings of the IEEE International Conference on Advanced Learning Technologies, p.1052{1053, ICALT '04, IEEE Computer Society, Washington, DC, USA, 2004.http://dl.acm.org/citation.cfm?id=1018423.1020205

[24] M. Ueno, "On-line contents analysis system for e-learning," Advanced Learning Technologies, 2004.Proceedings. IEEE International Conference on,pp.762{764, Aug. 2004.

[25] M. Ueno, "Animated pedagogical agent based on decision tree for e-learning," Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies, pp.188{192, ICALT'05,IEEEComputerSociety,Washington,DC,USA,2005.http://dx.doi.org/10.1109/ICALT.2005.63

[26] M. Ueno and T. Okamoto, \Online mdl-markov analysis of a discussion process in cscl," Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies, pp.764{768, ICALT'06, IEEE Computer Society, Washington, DC, USA,2006. http://dl.acm.org/citation.cfm?id=1156068.1156177

[27] 真臣植野，"多機能型 e ポートフォリオシステム "samurai-folio" の開発，日本教育工学会研究報告集，vol.2010，no.3，pp.33–40，jul 2010．

[28] A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Hidden topic Markov models," AISTATS, 2007.

[29] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum, "Integrating topics and syntax," NIPS, vol. 17, pp. 537–544, 2005.

[30] D. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, pp. 77–84, 2012.

[31] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei,"Reading tea leaves: How humans interpret topic models," in NIPS, 2009.

[32] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in EMNLP. ACL, 2011, pp. 262–272.

[40] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in NAACL, 2010.

Acknowledgement

First of all, I would like to express my gratitude to my supervisor, Prof. Maomi Ueno for giving me the continuous guidance and invaluable encouragement throughout the whole work.

And I am deeply grateful to Prof. Kenji Tanaka and Assoc. Prof. Yasuyuki Tahara for giving invaluable comments and guidance.

Finally, I wish to thank all the members of Ueno Laboratory for their cooperation, especially Senior Yoshihiro Kato for the guidance throughout the study of topic models.