Semantic Topic Discovery for Lecture Video

Jiang Bian[1] and Maolin Huang[2]

University of Technology Sydney, Sydney, Australia,
bianjiang22@gmail.com,
University of Technology Sydney, Sydney, Australia,
Mao.Huang@uts.edu.au

Abstract. With more and more lecture videos are available on the Internet, on-line learning and e-learning are getting increasing concerns because of many advantages such as high degree of interactivity. The semantic content discovery for lecture video is a key problem. In this paper, we propose a Multi-modal LDA model, which discovers the semantic topics of lecture videos by considering audio and visual information. Specifically, the speaking content and the information of presentation slides are extracted from the lecture videos. With the proposed inference and learning algorithm, the semantic topics of the video can be discovered. The experimental results show that the proposed method can effectively discover the meaningful semantic characters of the lecture videos.

Keywords: lecture video, topic model, Multi-modal LDA model

1 Introduction

Lecture videos are knowledge sources, intellectual properties of university and material for multimedia course ware and teaching evaluation. Compared with text in the book, lecture videos have many unique advantages: it is more salient and attractive, so it grabs users attention instantly; it carries more visual information that can be comprehended more quickly. Data, speech and digital TV broadcast are regarded as the most considerable contents of online learning or e-learning, which are rapidly emerging in the world, and separate education to students distributed around the world. According to the booming of Internet and digital technology, Internet based distance learning has many advantages such as high degree of interactivity, a variety of courses are available at any time, uses less bandwidth.

To give background, Web Based Training is a computer-based educational service that uses Internet to support distance learning. This technology is becoming popular for providing university courses and business training so that it allows students to learn wherever they are. Therefore, how to mine the related knowledge and corresponding multimedia from the Internet is a key for online-learning. In this paper, we propose a system classifying videos into categories, so users can efficiently find their interesting multimedia sources from the Internet. Furthermore, users can publish their notes or comments for some

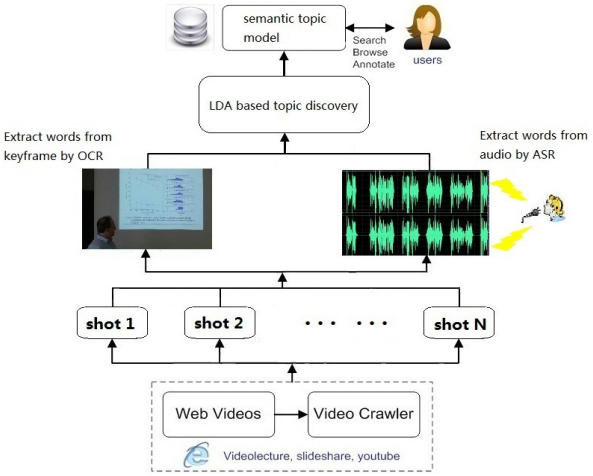

Fig. 1. Multi-modal LDA based topic discovery for lecture videos.

lectures and share with others through the network. This makes a community for online-learning, which is interactive and interesting. Therefore, the content-based multimedia information retrieval is a critical issue for our system.

Different from general multimedia content on the Internet, lecture videos contain many information sources. Specially, a lecture video contains the speaking content and the presentation slides for the lecture, as shown in Fig. 1. Furthermore, there are some scripts, user's notes or PowerPoint for the lecture on the Internet. The fusion of all these information is important for content-based multimedia retrieval. Therefore, we propose Multi-modal LDA model to achieve the goal.

As shown in Fig. 1, our system first segments each video into shots, and obtains two kinds of information from each shot: the speaking content and the presentation slides. Regards to the speaking content, we extract what the speaker has said by Automatic Speech Recognition (ASR). For the content of slides, we first extract key frames of the shot, which is defined as the frame that contains the slides of lecture. Since the lecture video usually containing two kinds of frames: one is an image of speaker, the other is an image of slides, so it is obviously that the frames which contain slides are defined as key frame. After key frame extraction, we extract what the slide shown by Optical Character Recognition (OCR). Therefore, we obtain two sets of texts from the shot and they are treated as evidences for semantic content analysis.

After that, we propose a model to discover semantic content from the obtained evidences. Our model is derived from topic model, which is originally proposed for text processing and generates the topics to describe a distribution of words, for example, probability Latent Semantic Analysis (pLSA) [1][2] and Latent Dirichlet Allocation (LDA) [3]. Specifically, LDA is a generative prob-

abilistic model introduced, and it is a three-level hierarchical Bayesian probabilistic model that a mixture of a latent set of distributions of discrete semantic data to set topics. Our multi-modal LDA model is derived from LDA model.

In this model, we have two evidence variables, w and l, corresponding to

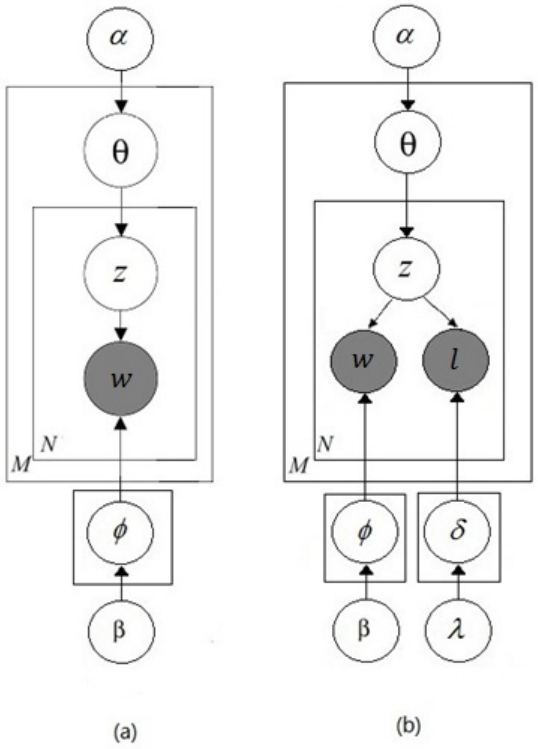

Fig. 2. (a). standard LDA model. (b). multi-modal LDA model

speaking content and the presentation slides in the video. They are generated from topic z, which indicates the semantic topic of the shot. The document corresponds to a video.

2 Multi-modal LDA model for Semantic Topic Discovery

2.1 Evidence extraction

Before we classifying videos by semantic information, we need to extract it by the methods. In lecture video, speakers teach knowledge to students or Internet

users through speaking and showing slides, so semantic information can be extracted from lectures oral presentation and slides.

ASR extracts semantic information from the speaking through capturing spoken words and then connects words to form a sentence. It is a computer-driven method which transcribes spoken language into text, so the semantic information can be read in real time. To ensure ASR working normal, it must follows three steps. The system will capture the words that are recording by any storage device, and then converts the digital signals of the speaking into syllables or even phoneme directly. This is referred as feature analysis. Next, the system will match the spoken syllables to a phoneme sequence that is kept in an acoustic model database. This is called pattern classification. Finally, the system tries to make sense what is being said by comparing the word phonemes from a language model database. Based on ASR, speech information can be extracted.

OCR extracts semantic information on slides through scanning and translating image of handwritten, typewritten or printed text into machine-encoded text. An OCR system recognizes the fixed static shape of the characters and uses a dictionary to match recognized words, so that it can increase recognition rates. Through this technology, the semantic information on slides can be extracted.

2.2 Multi-modal LDA model

The Markov chain Monte Carlo (MCMC) algorithm is constructed to converge to the target distribution [4]. As a sampling method deriving from MCMC algorithm, Thomas reaches the target distribution by using Gibbs Sampler, which is a heat bath algorithm in statistical physics and the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data. Therefore, Thomas proposed the equation based on this thinking as follows,

$$P\left(z_i = j \,|\mathbf{z}_{-i}, \mathbf{w}\right) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(w_i)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,}^{(d_i)} + T\alpha} \tag{1}$$

where $n_i^{(\cdot)}$ is the counting that does not include the current assignment of z_i. This result is quite intuitive; the first ratio expresses the probability of w_j under topic j, and the second ratio expresses the probability of topic j assigned in document d_i. After obtained the full conditional distribution, the Monte Carlo algorithm is then straightforward. $\{z_i\}$ are initialized to values in $\{1, 2, \cdots, T\}$, that determines the initial states of the Markov chain. We do this by an online version of Gibbs sampler. Furthermore, all variables are list in Tab. 1,

According to Thomass thinking, we derive Gibbs sampler based LDA model so that it can input two variables and cluster them based on semantic information, which present as words from videos by ASR and OCR. Furthermore, given documents contain topics that express over words. Based on [4], we derive equations that could include two variables.

Table 1. Multimodal-LDA variable list

α	Dirichlet prior parameter on the document-topic distributions
β	Dirichlet prior parameter on the topic-word distribution based on ASR
λ	Dirichlet prior parameter on the topic-word distribution based on OCR
θ_i	the topic distribution for document i, and $\theta \sim Dir(\alpha)$
ϕ_i	the word distribution for topic i based on ASR, and $\Phi \sim Dir(\beta)$
δ_i	the word distribution for topic i based on OCR, and $\Delta \sim Dir(\lambda)$

Algorithm 1 Multi-modal LDA generative model

1: Choose $\theta_i \sim Dir\,(\alpha)$, where $i \in \{1, \cdots, M\}$ and $Dir\,(\alpha)$ is Dirichlet Distribution;
2: For each ASR words vector w_i of W_{ASR}, choose $\phi_i \sim Dir\,(\beta)$, where $w_i \in [1, \cdots, w_{W_1}]$;
3: For each OCR words vector l_j of W_{OCR}, choose $\delta_j \sim Dir\,(\lambda)$, where $l_j \in [1, \cdots, l_{L_1}]$;
4: For each word w_i in the ASR word set:
5: • Choose $z_i \sim Multinomial\,(\theta_i)$.
6: • Choose w_i from $p\,(w_i|z_i, \beta)$, a multinomial probability conditional on the topic z_i
7: For each word l_j in the OCR word set:
8: • Choose $z_j \sim Multinomial\,(\theta_j)$.
9: • Choose l_j from $p\,(l_j|z_j, \beta)$, a multinomial probability conditional on the topic z_i

Φ is $T \times V$ Markov matrix, where V is the dimension of the vocabulary and T is the dimension of the topics, and each row of matrix denotes the word distribution of a topic. Our strategy for discovering topics differs from previous approaches in not explicitly representing θ and δ as parameters to be estimated, but instead of considering the posterior distribution over the assignments of words to topics $p\,(\mathbf{L}|\mathbf{Z})$. We then obtain estimates of θ and δ by examining this posterior distribution.

We can get formulas including single variable [4], and integrating out δ_i gives the first term

$$
\begin{aligned}
P\left(\mathbf{Z}\right) &= \int_{\theta} \prod_{j=1}^{D} p(\theta_j;\alpha) \prod_{t=1}^{W} p(Z_{j,t}|\theta_j) d\theta \\
&= \prod_{j=1}^{D} \frac{\Gamma(T\alpha)}{\prod_{i=1}^{T} \Gamma(\alpha)} \frac{\prod_{i=1}^{T} \Gamma(n^i_{j,(\cdot)} + \alpha)}{\Gamma(\sum_{i=1}^{T} n^i_{j,(\cdot)} + \alpha)}
\end{aligned} \tag{2}
$$

and

$$
\begin{aligned}
P\left(\mathbf{L}|\,\mathbf{Z}\right) &= \int_{\delta} \prod_{i=1}^{T} p(\delta_i;\lambda) \prod_{j=1}^{D} \prod_{t=1}^{W} p(L_{j,t}|\delta_{Z_{j,t}}) d\delta \\
&= \prod_{i=1}^{T} \frac{\Gamma(V\lambda)}{\prod_{r=1}^{V} \Gamma(\lambda)} \frac{\prod_{r=1}^{V} \Gamma(n^i_{(\cdot),r} + \lambda)}{\Gamma(\sum_{r=1}^{V} n^i_{(\cdot),r} + \lambda)}
\end{aligned} \tag{3}
$$

where $n^i_{j,(\cdot)}$ is the counting of word i has been assigned to topic j in the vector of assignments $\mathbf{z}$, $n^i_{(\cdot),r}$ indicates the counting of word i has been assigned to vocabulary $\mathbf{r}$, and $\Gamma\left(\cdot\right)$ is the standard gamma function. Then, we get the Gibbs Sampling Equation as follows

$$
P(Z_k|\mathbf{Z}_{-k}, \mathbf{L}; \alpha, \lambda) = \frac{n^k_{(\cdot),v} + \lambda}{\sum_{r=1}^{V} n^k_{(\cdot),r} + \lambda} \frac{n^k_{u,(\cdot)} + \alpha}{\sum_{i=1}^{T} n^i_{u,(\cdot)} + \alpha} \tag{4}
$$

Now we can advise the formula that allow us adding another variable in this model and Gibbs Sampler based LDA model with two variables. We get the equation $P(\mathbf{Z}|\mathbf{W}, \mathbf{L})$

$$
P(\mathbf{Z}|\mathbf{W}, \mathbf{L}) = P\left(\mathbf{W}|\mathbf{Z}\right) \times P\left(\mathbf{Z}\right) \times P\left(\mathbf{L}|\mathbf{W}, \mathbf{Z}\right). \tag{5}
$$

Through clustering words by $P\left(\mathbf{W}|\mathbf{Z}\right)$ and $P\left(\mathbf{L}|\mathbf{Z}\right)$, so we get that $P\left(\mathbf{L}|\mathbf{W}, \mathbf{Z}\right) = P\left(\mathbf{L}|\mathbf{Z}\right)$. Therefore, (4) can be express as follows

$$
P\left(\mathbf{W}|\mathbf{Z}\right) \times P\left(\mathbf{Z}\right) \times P\left(\mathbf{L}|\mathbf{W}, \mathbf{Z}\right) = P\left(\mathbf{W}|\mathbf{Z}\right) \times P\left(\mathbf{Z}\right) \times P\left(\mathbf{L}|\mathbf{Z}\right) \tag{6}
$$

With analyzing this formula, the equation to process two variables is

$$
\begin{aligned}
&p(Z_k|\mathbf{Z}_{-k}, \mathbf{W}, \mathbf{L}) \\
&\quad = \frac{n^k_{(\cdot),v} + \beta}{\sum_{r=1}^{V_{ASR}} n^k_{(\cdot),r} + \beta} \cdot \frac{n^k_{(\cdot),v} + \lambda}{\sum_{r=1}^{V_{\mathrm{OCR}}} n^k_{(\cdot),r} + \lambda} \cdot \frac{n^k_{u,(\cdot)} + \alpha}{\sum_{i=1}^{T} n^i_{u,(\cdot)} + \alpha}
\end{aligned} \tag{7}
$$

3 Experiment

In our experiment, we choose 100 lecture videos from *http://videolectures.net* to have an experiment, and these 100 videos have different titles which focus on Beyesian Process or Dirichlet Process in Machine Learning, which means the words relating to probability knowledge, characterizing method and describing models are involved. Then, we segment them into many shots based on time, for example, we segment a 20-mins video into 20 shots that each shot has 1 minute. Therefore, we totally get 1640 shots and 11000 key frames from these 100 videos. After that, we use ASR and OCR to extract semantic information from these key frames. The words extracted by ASR and OCR are treated as two kinds of evidences.

Since most existed OCR tools focus on recognize characters instead of words in OCR, a word will be recognized incorrectly even if there is a character in the word is recognized incorrectly. Thus, we need do some spelling check and correction operation after OCR extraction.

In ASR, some speakers teach their lectures with their assent, relating to different countries or different regions, for example, someone speak "dream" as "doraemon"' and "very" as "vely". Besides that, speakers say some prepositions such as "is", "and", "or", so we also need to delete these words and the similar ones after extraction step.

To address these problems, we build up a stop-word dictionary as follows: firstly, we find a stop-word list on the network as initial dictionary; and then through analyzing the extracting information, the words such as "enough" and "local", can be added to dictionary. Repeating this program one by one, we can get dictionary and use it as filter to shield words that have no meaning for discovery knowledge to classify shots.

Now we take experiment with extracting semantic information to discovery knowledge and classify these shots into different topics. The dataset of clustering is too large, so we show a part of them.

In Tab. 2 and Tab. 3, we can find some words like "cftp" is not a word that

Table 2. Extracting words from the first lecture video.

	ASR	OCR
Extracting words	acknowledged	approach
	act	approaches
	action	bayesian
	active	beliefs
	actual	borrow
	added	borrow
	additional	cftp

Table 3. Extracting words from the second lecture video.

	ASR	OCR
Extracting words	act	applicable
	additional	bayesian
	africa	bound
	against	bounds
	ago	cholesky
	ahead	concave
	analyst	concavity

has some mean. Therefore, we state that "cftp" means "coupling from the past" and it occurs in the sentence *We borrow methods for inference UGMs: CFTP in Ising and Potts models*. Thus, we need to note that this type of word means Abbreviation.

Because they have different titles mainly focus on 3 fields, so we set $T = 3$, which means we should classify 10 videos include 1100 key frames into 3 topics. Then, we adjust the dimension of vocabulary and topic matrix, so we can index vocabulary and word number clearly. α, λ are the assuming symmetric Dirichlet priors, we set $\alpha = 5/T$, $\beta = 0.01$ and $\lambda = 2$.

In Tab. 4, we can find that the words in Topic 1 mostly mean describing

Table 4. Result of Multi-modal LDA with Gibbs Sampler

Topic 1	Topic 2	Topic 3
reg	features	gp-lvm
prior	parameter	portfolios
changes	algorithm	performance
extend	rsthq	modular
vary	bars	sensitive

probability distribution, the words in Topic 2 focus on the feature of model or characterizing the objects, and Topic 3 mainly involves in analyzing experiment results and performance of models in science field. Then, we can prove multi-modal LDA model, processing with two variables with Gibbs Sampler, is useful.

Based on these topics and through analyzing the meaning of every topic, we can define topics and may be one topic can be given two or three words for indexing. Users can browse videos through searching these words, and scanning lecture videos directly.

4 Conclusion and Future Works

We have shown in this paper that multi-modal LDA model is able to discovery semantic information from lecture videos. For the purpose that helping users searching and browsing videos directly, we presented an advised model based on LDA and added another input, so that system classifies lecture videos into few same titles efficient. The main contributed idea is adding one more variable into LDA model based on Gibbs Sampler. Through the model, we can use the semantic information from speaking and slides, but also based on the effect of different extracting technologies, lecture videos can be classified properly. Even with the professional words, this model can achieve the goal, so it will be more effective used in other video like sports, news or movies.

Our future work: (1) in order to improve the accuracy of classification, ASRs or OCRs ability of correction will be improved; (2) using model in video scene classification, extracting semantic information from the dialogue and background; (3) with more other features, we derive novel LDA model to process two different type features parallelly.

References

1. Hofmann, Thomas. "Probabilistic latent semantic indexing." In ACM SIGIR Forum, vol. 51, no. 2, pp. 211-218. ACM, 2017.
2. Hofmann, Thomas. "Unsupervised learning by probabilistic latent semantic analysis." Machine learning 42, no. 1-2 (2001): 177-196.
3. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.
4. Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101, no. suppl 1 (2004): 5228-5235.
5. Azaiez, Ikbel, and Mohamed Ben Ahmed. "An approach of a semantic annotation and thematisation of AV documents." In Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on, pp. 1406-1411. IEEE, 2010.
6. Lee, Lin-Shan, Shun-Chuan Chen, Yuan Ho, Jia-Fu Chen, Ming-Han Li, and Te-Hsuan Li. "An initial prototype system for Chinese spoken document understanding and organization for indexing/browsing and retrieval applications." In Chinese Spoken Language Processing, 2004 International Symposium on, pp. 329-332. IEEE, 2004.
7. Li, Weitao, Xiaojie Zhou, and Tianyou Chai. "Bag of visual words and latent semantic analysis-based burning state recognition for rotary kiln sintering process." In Control and Decision Conference (CCDC), 2011 Chinese, pp. 377-382. IEEE, 2011.
8. Ide, Ichiro, Hiroshi Mo, Norio Katayama, and Shin'ichi Satoh. "Exploiting topic thread structures in a news video archive for the semi-automatic generation of video summaries." In Multimedia and Expo, 2006 IEEE International Conference on, pp. 1473-1476. IEEE, 2006.
9. Mase, Motohiro, Seiji Yamada, and Katsumi Nitta. "Extracting topic maps from Web pages by Web link structure and content." In Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on, pp. 1232-1239. IEEE, 2008.

10. Huang, Rui, Fen Lin, and Zhongzhi Shi. "Focused crawling with heterogeneous semantic information." In Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on, vol. 1, pp. 525-531. IEEE, 2008.
11. Hennig, Leonhard, Thomas Strecker, Sascha Narr, Ernesto William De Luca, and Sahin Albayrak. "Identifying sentence-level semantic content units with topic models." In Database and Expert Systems Applications (DEXA), 2010 Workshop on, pp. 59-63. IEEE, 2010.
12. Zhang, Weifeng, Zengchang Qin, and Tao Wan. "Image scene categorization using multi-bag-of-features." In Machine Learning and Cybernetics (ICMLC), 2011 International Conference on, vol. 4, pp. 1804-1808. IEEE, 2011.
13. Kong, Sheng-Yi, and Lin-Shan Lee. "Semantic analysis and organization of spoken documents based on parameters derived from latent topics." IEEE Transactions on Audio, Speech, and Language Processing 19, no. 7 (2011): 1875-1889.
14. Liu, Huasheng, Gang Liu, and Yuqin Lv. "Semantic Hyperlink Analysis Model." In Semantics, Knowledge and Grid, 2008. SKG'08. Fourth International Conference on, pp. 416-419. IEEE, 2008.
15. Vretos, Nicholas, Nikos Nikolaidis, and Ioannis Pitas. "A perceptual hashing algorithm using latent dirichlet allocation." In Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on, pp. 362-365. IEEE, 2009.
16. Ngo, Chong-Wah, Ting-Chuen Pong, and Thomas S. Huang. "Detection of slide transition for topic indexing." In Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on, vol. 2, pp. 533-536. IEEE, 2002.
17. Mirylenka, Katsiaryna, Christoph Miksovic, and Paolo Scotton. "Applicability of latent dirichlet allocation for company modeling." In Industrial Conference on Data Mining (ICDM2016). 2016.
18. Hu, Chuan, Huiping Cao, and Qixu Gong. "Sub-Gibbs Sampling: A New Strategy for Inferring LDA." In Data Mining (ICDM), 2017 IEEE International Conference on, pp. 907-912. IEEE, 2017.