Latent Dirichlet Markov Allocation for Sentiment Analysis

Ayoub Bagheri

Isfahan University of Technology, Isfahan, Iran
Intelligent Database, Data Mining and Bioinformatics Lab, Electrical and Computer Engineering Department
a.bagheri@ec.iut.ac.ir

Mohamad Saraee

University of Salford, Manchester, UK
School of Computing, Science and Engineering
m.saraee@salford.ac.uk

Franciska de Jong

University of Twente, Enschede, Netherlands
Human Media Interaction, P.O. Box 217, 7500 AE
f.m.g.dejong@utwente.nl

Abstract. In recent years probabilistic topic models have gained tremendous attention in data mining and natural language processing research areas. In the field of information retrieval for text mining, a variety of probabilistic topic models have been used to analyse content of documents. A topic model is a generative model for documents, it specifies a probabilistic procedure by which documents can be generated. All topic models share the idea that documents are mixture of topics, where a topic is a probability distribution over words. In this paper we describe Latent Dirichlet Markov Allocation Model (LDMA), a new generative probabilistic topic model, based on Latent Dirichlet Allocation (LDA) and Hidden Markov Model (HMM), which emphasizes on extracting multi-word topics from text data. LDMA is a four-level hierarchical Bayesian model where topics are associated with documents, words are associated with topics and topics in the model can be presented with single- or multi-word terms. To evaluate performance of LDMA, we report results in the field of aspect detection in sentiment analysis, comparing to the basic LDA model.

Keywords: topic model, latent Dirichlet allocation (LDA), hidden markov model (HMM), Latent Dirichlet Markov Allocation (LDMA), sentiment analysis.

1 INTRODUCTION

With the explosion of web 2.0, user generated content in online reviews present a wealth of information that can be very helpful for manufactories, companies and other customers. Mining these online reviews to extract and summarize users' opinions is a challenging task in the field of data mining and natural language processing. One main task in sentiment review analysis is to find aspects that users evaluate in their reviews. Aspects are topics on which opinion are expressed about. In the field of sentiment analysis, other names for aspect are: features, product features or opinion targets [1-10]. Aspects are important because without knowing them, the opinions expressed in a sentence or a review are of limited use. For example, in the review sentence "after using iPhone, I found the size to be perfect for carrying in a pocket", "size" is the aspect for which an opinion is expressed. Likewise aspect detection is critical to sentiment analysis, because its effectiveness dramatically affects the performance of opinion word detection and sentiment orientation identification. Therefore, in this study we concentrate on aspect detection for sentiment analysis.

Different approaches have been proposed for aspect detection from reviews. Previous works like double propagation [4] and supervised learning methods have the limitation that they do not group semantically related aspect expressions together [2]. Supervised methods, additionally, are not often practical due to the fact that building sufficient labeled data is often expensive and needs much human labor. In contrast, the effectiveness of unsupervised topic modeling approaches has been shown in identifying aspect words. Probabilistic topic models are a suite of algorithms whose aim is to extract latent structure from large collection of documents. These models all share the idea that documents are mixtures of topics and each topic is a distribution over words [11-15]. We follow this promising line of research to extend existing topic models for aspect detection in sentiment analysis. Current topic modeling approaches are computationally efficient and also seem to capture correlations between words and topics but they have two main limitations: the first limitation is that they assume that words are generated independently to each other, i.e. the bag of words assumption. In other words, topic models only extract unigrams for topics in a corpus. We believe that a topic model considering unigrams and phrases is more realistic and would be more useful in applications. The second limit for current topic modeling approaches is that of the assumption that the order of words can be ignored is an unrealistic oversimplification. Topic models assume the subsequent words in a document or a sentence have different topics, which is not a true assumption. In our model, in addition to extracting unigrams and phrases for topics we assume that the topics of words in a sentence form a Markov chain and that subsequent words are more likely to have the same topic. Therefore, in this paper we propose a new topic modeling approach that can automatically extract topics or aspects in sentiment reviews. We call the proposed model: Latent Dirichlet Markov Allocation Model (LDMA), which is a generative probabilistic topic model based on Latent Dirichlet Allocation (LDA) and Hidden Markov Model (HMM), which emphasizes on extracting multi-word topics from text data. In addition LDMA relaxes the "bag of words" assumption from topic modeling approaches to yield to a better model in terms of extracting latent topics from text.

We proceed by reviewing the formalism of LDA. We then propose the LDMA model and its inference procedure to estimate parameters. We demonstrate the effectiveness of our model in our experiments by comparing to basic LDA model for aspect detection. Finally, we conclude our work and outline future directions.

2 LDA

LDA, introduced by David Blei et al. [12], is a probabilistic generative topic model based on the assumption that each document is a mixture of various topics and each topic is a probability distribution over different words.

A graphical model of LDA is shown in Figure 1, wherein nodes are random variables and edges indicate the dependence between nodes [12, 13]. As a directed graph, shaded and unshaded variables indicate observed and latent (i.e., unobserved) variables respectively, and arrows indicate conditional dependencies between variables while plates (the boxes in the figure) refer to repetitions of sampling steps with the variable in the lower right corner referring to the number of samples.

Given a corpus with a collection of *D* documents, each document in the corpus is a sequence of *W* words, each word in the document is an item from a vocabulary index with *V* distinct terms, and *T* is the total number of topics. The procedure for generating a word in a document by LDA is as follows:

```
1. for each topic z = 1…K,
     Draw $\varphi_z$ ~ Dirichlet($\beta$);
2. for each document d = 1…D,
     (a) Draw topic distribution $\theta^d$ ~ Dirichlet($\alpha$);
     (b) for each word $w_i$ in document d,
           i. Draw a topic $z_{d,i}$ ~ $\theta^d$;
           ii. Draw a word $w_{d,i}$ ~ $\varphi_{z_{d,i}}$;
```

The goal of LDA is therefore to find a set of model parameters, topic proportions and topic-word distributions. Standard statistical techniques can be used to invert the generative process of LDA, inferring the set of topics that were responsible for generating a collection of documents. The exact inference in LDA is generally intractable, and we have to appeal to approximate inference algorithms for posterior estimation. The most common approaches that are used for approximate inference are EM, Gibbs Sampling and Variational method [12, 13, 15].

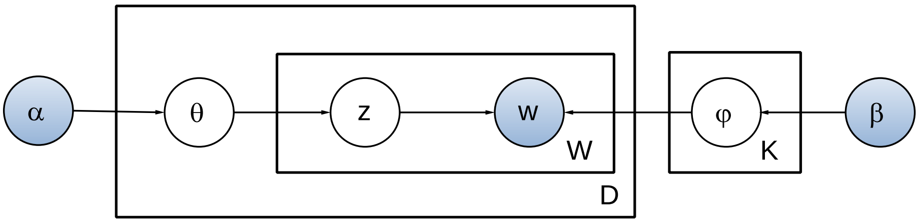

Figure 1 *LDA Topic Model*

3 LDMA

LDA model makes an assumption that words are generated independently to each other in a document. Also LDA ignores the order or positions of words or simplicity. Here we propose LDMA, a generative probabilistic topic model based on Latent Dirichlet Allocation (LDA) and Hidden Markov Model (HMM), to relax the “bag of words” assumption from LDA to yield to a better model in terms of extracting latent topics from text. Figure 2 shows the graphical model corresponding to the LDMA generative model.

LDMA, in addition to the two sets of random variables *z* and *w*, introduces a new set of variables *x* to detect an n-gram phrase from the text. LDMA assumes that the topics in a sentence form a Markov chain with a transition probability that depends on θ, a distribution ψ_{zw}, a random variable x_i and the topic of previous word z_{i-1}. Random variable x denotes whether a bigram can be formed with previous term or not. Therefore LDMA has the power to decide whether to generate a unigram, a bigram, a trigram or etc. Here we only consider generating unigrams and bigrams from LDMA. If the model sets x_i equal to one, it means that w_{i-1} and w_i form a bigram and if it is equal to zero they do not.

The generative process of the LDMA model can be described as follows:

```
3. for z = 1…K,
     Draw multinomial $\varphi_z$ ~ Dirichlet($\beta$);
4. for z = 1…K,
     for w = 1…W,
       (a) Draw binomial $\psi_{zw}$ ~ Dirichlet($\gamma$);
       (b) Draw multinomial $\sigma_{zw}$ ~ Beta($\delta$);
5. for d = 1…D,
     (c) Draw multinomial $\theta^d$ ~ Dirichlet($\alpha$);
     (d) for s = 1…S in document d,
           for each word $w_i$ in sentence s,
             i. Draw $x_i^s$ from binomial $\psi_{z_{i-1}^s w_{i-1}^s}$;
             ii. if($x_i^s$ = 0) Draw $z_i^s$ from multinomial $\theta^d$;
                 else $z_i^s = z_{i-1}^s$;
             iii. if($x_i^s$ = 1) Draw $w_i^s$ from multinomial $\sigma_{z_i^s w_{i-1}^s}$;
                  else draw $w_i^s$ from multinomial $\varphi_{z_i^s}$;
```

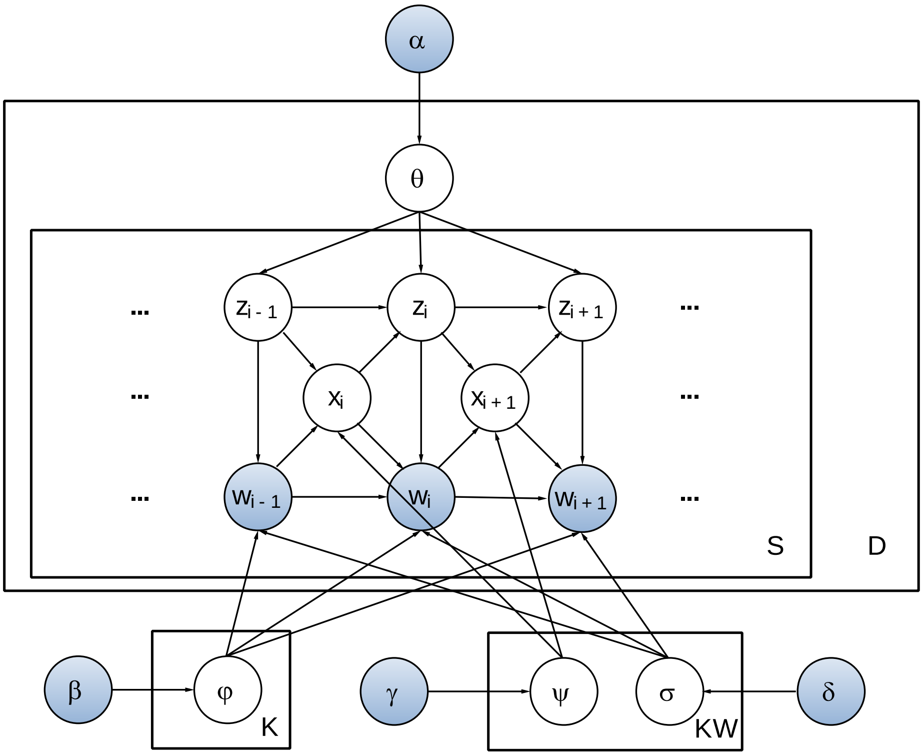

Figure 2 *A graphical view of LDMA Topic Model*

LDMA model is sensitive to the order of the words, which it is not assumes that a document is a bag of words, i.e. successive words in LDMA tend to have the same topics. Therefore unlike LDA approach, LDMA will not give the same topic to all appearances of the same word within a sentence, a document or corpus.

4 INFERENCE

The exact inference of learning parameters in LDMA is intractable due to the large number of parameters in the model [13, 15]. Therefore approximate techniques are needed. We use Gibbs sampling to approximate the latent variables in the model. At each transition of Markov chain, the aspect (topic) of *i*th sentence, z_i, and the n-gram variable x_i are drawn from the following conditional probability:

If $x_i = 0$ then:

$$P(z_i, x_i | z_{-i}, x_{-i}, w, \alpha, \beta, \gamma, \delta) \propto \frac{\gamma_{x_i} + \mathrm{p}_{z_{i-1} w_{i-1} x_i}}{\sum_{k=0}^{1}\left(\gamma_k + \mathrm{p}_{z_{i-1} w_{i-1} k}\right)} \left(\alpha_{z_i} + q_{sdz_i}\right) \left(\frac{\beta_{w_i} + n_{z_i w_i}}{\sum_{v=1}^{V}(\beta_v + n_{z_i v})}\right)$$

If $x_i = 1$ then:

$$P(z_i, x_i | z_{-i}, x_{-i}, w, \alpha, \beta, \gamma, \delta) \propto \frac{\gamma_{x_i} + \mathrm{p}_{z_{i-1} w_{i-1} x_i}}{\sum_{k=0}^{1}\left(\gamma_k + \mathrm{p}_{z_{i-1} w_{i-1} k}\right)} \left(\alpha_{z_i} + q_{sdz_i}\right) \left(\frac{\delta_{w_i} + m_{z_i w_{i-1} w_i}}{\sum_{v=1}^{V}(\delta_v + m_{z_i w_{i-1} v})}\right)$$

Where x_{-i} denotes the bigram status for all words except w_i, z_{-i} represents the topic (aspect) assignments for all words except w_i, n_{zw} represents how many times word *w* is assigned to aspect *z* as a unigram, m_{zwv} represents how many times word *v* is assigned to aspect *z* as a second term of a bigram word given the previous word *w*, p_{zwk} denotes how many times the status variable *x* is set to *k* given the previous word w and the previous word's aspect *z*, and q_{sz} represents how many times a word is assigned to aspect *z* in sentence *s* of document *d*.

5 EXPERIMENTS

In this section, we apply the proposed LDMA model to customer review datasets of digital cameras and compare the results with the original LDA model. These datasets of customer reviews [1] contain two different products: Canon G3 and Nikon Coolpix 4300. Table 1 shows the number of review sentences and the number of manually tagged product aspects for each product in the dataset.

Table 1

Summary of customer review datasets.

Dataset	Number of review sentences	Number of manual aspects
Canon	597	100
Nikon	346	74

We first start by preprocessing review document from the datasets. We extract the sentences according to the delimiters '.', ',', '!', '?', ';'. And then by removing Stopwords and words with frequency less than three we extract a feature vector to represent review documents. By applying LDMA and the original LDA models examples of most probable aspects (topics) extracted are shown in Tables 2 and Table 3.

Table 2

Exampled aspects discovered by LDMA and LDA form Canon reviews

Canon	
LDMA	**LDA**
canon	canon
canon powershot	purchased
durability	durability
photos	easy
battery	life
digital zoom	battery
picture	quality
size	powershot
picture quality	zoom
optical zoom	size
photos quality	picture
battery life	digital

Table 3

Exampled aspects discovered by LDMA and LDA form Nikon reviews

Nikon	
LDMA	**LDA**
easy use	Battery
auto mode	perfect
camera	transfer
pictures	camera
nikon	nikon
transfer cable	Manual
battery	macro
manual mode	small
print quality	mode

From the tables we can find that the LDMA model discovered more informative words for aspects or topics. In addition to the unigrams, LDMA can extract phrases, hence the unigram and bigram list of aspects are more pure in LDMA. Also LDMA associates words together to detect the multi-word aspects which are only highly probable in this model. Based on the results, LDMA can successfully find aspects that consist of words that are consecutive in a review document.

6 CONCLUSIONS

Managing the explosion of digital content on the internet requires new tools for automatically mining, searching, indexing and browsing large collections of text data. Recent research in data mining and statistics has developed a new brand of techniques call probabilistic topic modeling for text. In this paper, we proposed LDMA model, a model which extends the original topic modeling approach, LDA, by considering the underlying structure of a document and order of words in document. LDMA ignores the bag of words assumption of LDA to extract multi-word and n-gram aspects and topics from text data. We find that ignoring this basic assumption allows the model to learn more coherent and informatics topics.

REFERENCES

[1] Hu, M., Liu, B. (2004). Mining opinion features in customer reviews. In: Proceedings of 19th Na-tional Conference on Artificial Intelligence AAAI.

[2] B. Liu, L. Zhang, (2012). A survey of opinion mining and sentiment analysis, Mining Text Data, 415-463.

[3] C. Lin, Y. He, R. Everson, S. Ruger, (2012). Weakly supervised joint sentiment-topic detection from text, IEEE Trans. Knowl. Data Eng. 24, no. 6, 1134-1145.

[4] G. Qiu, B. Liu, J. Bu, C. Chen. (2011). Opinion word expansion and target extraction through double propagation, Computational linguistics, 37(1), 9-27.

[5] I. Titov, R. McDonald, (2008). A joint model of text and aspect ratings for sentiment summarization, in: Proceedings of the Annual Meeting on Association for Computational Linguistics and the Human Language Technology Conference (ACL-HLT). pp. 308–316.

[6] S. Brody, N. Elhadad, (2010). An unsupervised aspect-sentiment model for online reviews, in: Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics. Publishing, Association for Computational Linguistics, pp. 804-812.

[7] S. Moghaddam, M. Ester, (2011). ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. Publishing, ACM, pp. 665-674.

[8] X. Fu, G. Liu, Y. Guo, Z. Wang, (2013). Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, Knowledge-Based Systems, 37, 186-195.

[9] Z. Zhai, B. Liu, H. Xu, P. Jia, (2011). Constrained LDA for grouping product features in opinion mining, in: Proceedings of 15th Pacific-Asia Conference, Advances in Knowledge Discovery and Data Mining, pp 448-459.

[10] Jo, Y., & Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 815-824). ACM.

[11] C. Zhai, J. Lafferty, (2001). Model-based feedback in the language modeling approach to information retrieval, in: Proceedings of 10th International Conference on Information and knowledge management. Publishing, pp. 403-410.

[12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022.

[13] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. (2005). Integrating topics and syntax. In Lawrence K. Saul, Yair Weiss, and L´eon Bottou, editors, Advances in Neural Information Processing Systems 17, pages 537–544. MIT Press, Cambridge, MA.

[14] Hanna M. Wallach. (2006). Topic modeling: Beyond bag-of-words. In ICML ’06: 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, USA.

[15] Xuerui Wang and Andrew McCallum. (2005). A note on topical n-grams. Technical Report UM-CS-071, Department of Computer Science University of Massachusetts Amherst.