Research on a Text Topic Mining Algorithm Based on Big Data

Zequn Yang
Information Management and Information System, North China University of Science and Technology, Tangshan, Hebei, 063200

Abstract. Topic modeling has been widely used to mine topics from documents. However, a key disadvantage of topic modeling is that it requires large amounts of data (e.g. thousands of documents) to provide reliable statistics to generate coherent topics. However, in practice, many document collections do not have as many documents. Given a small number of documents, the topic generated by the classic topic model LDA is very poor. Even in the case of a large amount of data, unsupervised learning of the topic model will still produce unsatisfactory results. In recent years, knowledge-based topic models have been proposed which require human users to provide some prior domain knowledge to guide the model to generate better topics. Our research takes a completely different approach. We recommend learn like humans, i.e. retain the results learned in the past and use them to help future learning. When faced with a new task, we first mine some reliable (transcendental) knowledge from past learning/modeling results, and then use it to guide model inference to produce more coherent topics. This approach is possible because there is readily available big data on the Web. The algorithm mines two forms of knowledge: the chain must be (Meaning two words should be in the same topic) and cannot be linked (meaning two words should not be in the same topic). Two issues of automatic knowledge mining are discussed, namely the problem of erroneous knowledge and knowledge transferability. Use from the experimental results of the review documents of 100 product areas show that the method proposed in this paper has a significant improvement over the state-of-the-art baseline.

1. Introduction
Topic models (e.g. LDA, pLSA, etc.) have been widely used to extract topics from text documents [1]. However, these models often require large amounts of data, such as thousands of documents, to provide reliable statistics to generate coherent topics [2]. This is a major flaw, because in practice very few document collections have so many documents [3]. For example, in the task of finding product features or aspects from online reviews for opinion mining, most products are in there are not even more than 100 comments (documents) in the review site [4]. As we will see in the experimental section, given 100 reviews, the classic theme model LDA produces very poor results [5].

There are three main ways to solve this problem:

(1) Inventing a Better Topic Model: This method may be effective if a large number of documents are available [6].
(2) Require users to provide prior domain knowledge: An obvious form of external information is the user’s prior knowledge of the domain [7].

(3) Life-long learning like humans: We still use knowledge-based methods, but automatically mine prior knowledge from the results of past learning [8]. This method works similar to human learning.

Existing research focuses on the first two methods [9]. We believe that it is time to create algorithms and build systems that learn like humans [10]. In our context, lifelong learning is possible, thanks to two key observations:

1. Although each area is different, there is considerable cross-cutting of topics across domains.
2. In themes previously generated from many fields, you may also find that pictures and prices should not be in the same theme.

This article proposes a new topic model (AMC), whose reasoning can make use of automatically mined knowledge to deal with erroneous knowledge and transitive problems, resulting in excellent topics. Our experiments using comment sets from 100 fields show that the performance of the proposed AMC model is significantly better than the latest baseline model.

2. Overview of the proposed algorithm
This section introduces the proposed algorithm flow, which follows the lifelong learning ideas described in the introduction. The algorithm consists of two phases:

Phase 1 - Initialization
Given a set of previous document sets \( D = (D_1, ..., D_n) \), this step first runs standard LDA on each domain set \( D_i \in D \) to generate a set of topics \( S_i \). Result topics from all \( n \) domains are combined to produce a set \( S \) of all topics, e.g. \( S = \bigcup_i S_i \). We call \( S \) a collection of previous topics (or p topics). Then mine a set of necessary links from \( S \), which will be covered in detail in section 4.1. Note that this initialization phase is only applied at the beginning. It is not used for every modeling of a new document collection.

| Algorithm 1 | AMC(D, S, M) |
|-------------|--------------|
| 1. \( A^l \) ← GibbsSampling\( (D^l, N, M, \emptyset) \); // \( \emptyset \): no cannot-links. |
| 2. for \( r = 1 \) to \( R \) do |
| 3. \( C \) ← \( C \cup \text{MineCannotLinks}(S, A^l) \); |
| 4. \( A^l \) ← GibbsSampling\( (D^l, N, M, C) \); |
| 5. end for |
| 6. \( S \) ← Incorporate\( (A^l, S) \); |
| 7. \( M \) ← MiningMustLinks\( (S) \); |

Figure 1. Algorithm 1

Stage 2 - Lifelong Learning through AMC:
Given a new/test document collection DT, this stage generates topics from the DT using the proposed AMC model. To distinguish these topics from p-topics, we refer to them as the current topic (or c-topic for short). AMC is given in Algorithm 1. The first line runs the proposed Gibbs sampler (introduced in Section 5.3), using only the M that must be linked from the p-topic set (S) generated so far to generate a set of topics, where \( N \) is the number of Gibbs sampling iterations. The third line cannot-link based on the current topic and p topic S. Then, the fourth line uses both must link and not link to improve the result topic. Please note that this process can run iteratively. We refer to these iterations as learning iterations, which is different from Gibbs iterations. In each learning iteration, we hope to get better topic results. We will experiment with the number of learning iterations later. Currently, merge (At, S) The function (line 6 in Algorithm 1) is very simple. If the At field exists in S, replace those topics of the field in S with at; otherwise, add At to S. Using the updated S, dig one the new group must be linked (line 7) and it will be called in the next new build by calling AMCUsed in mold tasks.
3. Amc model
We now introduce the proposed AMC model. As mentioned earlier, due to errors in the topic model results, some of the necessary and unlinkable automatic mining may be wrong. AMC has the ability to handle this incorrect knowledge. The idea is, the correct semantic links that must be linked and cannot be linked should also be reasonably summarized by the statistical information behind the domain set. If a piece of knowledge (which must be linked or cannot be linked) is inconsistent with the domain set, the knowledge is likely to be in this particular test domain is usually incorrect or incorrect. In either case, the model should not trust or leverage this knowledge. AMC still uses lda's graphical model and its generation process. Therefore, we do not give a graphical model. However the reasoning mechanism of AMC is completely different from that of LDA. The reasoning mechanism cannot be reflected in the graphical model using the plate symbol.

Errors that cannot be linked are often more difficult to detect and verify than errors that must be linked. Due to the power-law distribution of natural language words, most words are rare and do not co-occur with most other words. Low co-occurrence of two words does not necessarily mean negative correlation (cannot be linked). Therefore, we detect and balance unlinkable in the sampling process. More specifically, we extended the Pólya black box model, added unlinkable knowledge, and resolved the above problem.

4. Gibbs sampler proposed in this paper
This section introduces the Gibbs sampler of the proposed AMC model, which is different from LDA because AMC requires additional mechanisms to utilize prior knowledge and handle the problems of prior knowledge during the sampling process. We propose a multi-generalization of the task Polya Black Box Model (M-GPU). Below, we will introduce it:

In M-GPU, when a ball is drawn randomly, a certain number of extra balls of each color are returned to the urn, instead of just returning two balls of the same color as in the SPU. This is inherited from the GPU. Therefore, the proportion of these colored balls has increased, making them more likely to be drawn in this urn in the future. We call this a promotion of these colored balls. Applying this idea to our case, when the item w. When assigned to topic k, each term w0 that must be linked with w is also assigned to topic k a certain amount, which is determined by the matrix $\lambda_{w0, w}$ (see Equation 1). Therefore, w0 is promoted by w. As a result, in topic the probability of w0 under k also increases.

$$
\lambda_{w0, w} = \begin{cases} 
1 & \text{if } w = w' \\
\mu \times PMI(w, w') & \text{is a must-link} \\
0 & \text{otherwise}
\end{cases}
$$

According to the definition of unavailable link, there is a high probability that two words in the unavailable link cannot be connected under the same topic. Since M-GPU allows multiple URNs to interact, when sampling from the term URN UW k represents the term w. We want to transfer the ball representing the non-terms of w, such as WC (shared with w cannot be linked) to other URNs, that is, reduce the probability of those non-terms under this topic, and under some other topic increase its corresponding probability. In order to correctly transfer the ball representing the term WC, it should be transferred to an ash box with a higher proportion of WC. That is, we randomly sample an ash box with a higher proportion of Wcto transfer Wcto to Wcto. However, there is a situation where no other urn has a higher percentage of toilets. It is recommended to create a new urn to move Wcto assuming that the knowledge that cannot be linked is correct. The knowledge that cannot be linked may be incorrect. For example, assuming the model puts battery and life in the same topic k, where both battery and life have the highest probability, you cannot link to see them in the same topic and want to separate them. In this case, we should not trust can't link because it might break related terms into different topics.
5. Experimental results and analysis

5.1. Data Set Used in This Article
We have created two large datasets for our experiments. The first dataset contains reviews from 50 types of electronics or field (see the first row of Table 1). The second dataset contains from 50 types Mixed types of reviews for non-electronic products or fields (see the second row of Table 1). There are 1000 reviews for each domain name. Using the first dataset, we want to show the performance of AMC when there is a considerable topic overlap. Use For the second data set, we want to show the performance of AMC in the case of less overlapping topics. We follow the preprocessing principle to preprocess the data set. These data sets can be publicly available on the author's website. Figure 1 shows the average topic correlation value for each model on the 50 test sets.

Table 1. List of 100 domain names: electronics (first line) and non-electronics (second line).

| Electronic Domain | Non-Electronic Domain |
|-------------------|-----------------------|
| Alarm Clock       | Apple                  |
| Amplifier         | Bat                    |
| Battery           | Blue-Ray Player        |
| Cable Modem       | Camera                 |
| Camcorder         | Camera, Car Stereo     |
| Camera            | CD Player, Cell Phone  |
| Computer          | DVD Player, Fan        |
| DVR               | Graphics Card          |
| Headphone         | Hard Drive             |
| Home Theater System | Iron, Keyboard          |
| Iron              | Kindle, Lamp           |
| Laptop            | Laptop, Media Player   |
| Media Player      | Memory Card, Microphone |
| Microwave         | Monitor, Mouse        |
| Monitor           | MP3 Player, Network Adapter |
| Mouse             | Printer, Projector    |
| MP3 Player        | Radar Detector, Remote Control |
| Network Adapter   | Rice Cooker, Skimmer  |
| Printer           | Speaker, Subwoofer    |
| Projector         | Tablet, Telephone, TV  |
| Radar Detector    | Vacuum, Video Player   |
| Remote Control    | Video Recorder, Voice Recorder |
| Table             | Watch, Webcam, Wireless Router, Xbox |
| Electronics       | Accessories, Automotive, Baby, Bag, Beauty, Bike, Books, Cable, Car, Clothing, Conditioner, Diaper, Dining, Drum, Flashlight, Food, Gloves, Golf, Home Improvement, Industrial Scientific, Jewelry, Kindle Store, Kitchen, Knife, Luggage, Magazine, Newspaper, Matt, Mattress, Movie, TV, Music, Musical Instruments, Office Product, Patio Lawn Garden, Pet Supplies, Pillow, Sandal, Scooter, Shoes, Software, Sports, Table Chair, Tint, Tire, Toys, Video Games, Vitamin Supplement, Wall Clock, Water Filter |

Figure 2. Average topic relevance per model

5.2. Test Results
Figure 2 shows the average P-repeat rate @5 (top) and P-repeat rate @10 (middle) of the topic words for each model in each domain with only coherent topics (not considering non-coherent topics). Obviously, AMC the highest p@5 and p@10 values were achieved in all 10 domains. LTM is also generally better than LDA, but significantly worse than AMC. This is consistent with the subject coherence results in Section 6.2. In the case of data, the results of LDA are very bad. On average, for p@5 and p@10, AMC improves LTM and LDA by 8% and 14% and 25% and 25% respectively. Using paired t test for significant the sex test showed that the maximum likelihood ratio (p<0.0002) and the maximum likelihood ratio (p<0.0001) showed significant improvement on p@5 and p@10.
Figure 3. Top and middle: the topic words pcumulant @5 and cumulant @10 related topics for each model; bottom: the number of related (#related) topics found by each model. The left-to-right bars in each group are respectively used for AMC, LTM, and LDA.

6. Conclusion

This article proposes an advanced topic model AMC capable of lifelong learning. For such learning, it mines prior knowledge from the results of past modeling and uses this knowledge to help future modeling. Our system is based on a large number of previous the subject generated by the document set (big data) is automatically mined in two forms of prior knowledge, that is, must be linked and cannot be linked. The system also identifies some issues related to the knowledge of automatic mining. The proposed AMC model can not only mine the learned Knowledge, and can process the knowledge issues mined to generate more accurate topics. Experimental results using 100 domain review sets show that the performance of the proposed AMC model is significantly better than the existing latest models. In our future work we plan to study other aspects of lifelong learning in the context of topic modeling, such as how to maintain previous topics, and how to incrementally update the knowledge that must be linked when new topics are added to previous topic sets.

References

[1] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In ICML, pages 25–32, 2009.

[2] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In IJCAI, pages 1171–1177, 2011.

[3] D. M. Blei and J. D. McAuliffe. Supervised Topic Models. In NIPS, pages 121–128, 2007.

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

[5] S. R. K. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. Learning Document-Level Semantic Properties from Free-Text Annotations. In ACL, pages 263–271, 2008.

[6] Zhao Yunsong. Application of Topic Model in Medical Service Text Data Mining [D]. PLA Air Force Military Medical University, 2018.

[7] Dai Hengyu. Research on User Complaint Data of Mobile Operators Based on Data Mining Technology [D]. Nanjing University, 2016.

[8] Lu Jian. Text-oriented Topic Mining Technology and Implementation [D]. Jinan University, 2010.
[9] Xu Qinan. Research on Key Technologies of Chinese Text Automatic Summarization for web Text Mining [D]. xidian university, 2009.

[10] Yang Danbo. Design and Implementation of Topic Meta Search Engine Based on Web Data Mining [D]. Tsinghua University, 2009.