A TEXT MINING RESEARCH BASED ON LDA TOPIC MODELLING

Zhou Tong[1] and Haiyi Zhang[2]

Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada
[1]zhoutong@acadiau.ca
[2]haiyi.zhang@acadiau.ca

ABSTRACT

A Large number of digital text information is generated every day. Effectively searching, managing and exploring the text data has become a main task. In this paper, we first represent an introduction to text mining and a probabilistic topic model Latent Dirichlet allocation. Then two experiments are proposed - Wikipedia articles and users' tweets topic modelling. The former one builds up a document topic model, aiming to a topic perspective solution on searching, exploring and recommending articles. The latter one sets up a user topic model, providing a full research and analysis over Twitter users' interest. The experiment process including data collecting, data pre-processing and model training is fully documented and commented. Further more, the conclusion and application of this paper could be a useful computation tool for social and business research.

KEYWORDS

topic model, LDA, text mining, probabilistic model

1. INTRODUCTION

As computers and Internet are widely used in almost every area, more and more information is digitized and stored online in the form of news, blogs, and social networks. Since the amount of the information is exploded to astronomical figures, searching and exploring the data has become the main problem. Our research is intended to design a new computational tool based on topic models using text mining techniques to organize, search and analyse the vast amounts of data, providing a better way understanding and mining the information.

2. BACKGROUND

2.1. Text Mining

Text mining is the process of deriving high-quality information from text [1]. Text mining usually involves the process of structuring the input text, finding patterns within the structured data, and finally evaluation and interpretation of the output. Typical text mining tasks include text categorization, text clustering, document summarization, keyword extraction and etc. In this

Jan Zizka et al. (Eds) : CCSEIT, AIAP, DMDB, MoWiN, CoSIT, CRIS, SIGL, ICBB, CNSA-2016
pp. 201–210, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60616

research, statistical and machine learning techniques will be used to mine meaningful information and explore data analysis.

2.2. Topic Modelling

In machine learning and natural language processing, topic models are generative models, which provide a probabilistic framework [2]. Topic modelling methods are generally used for automatically organizing, understanding, searching, and summarizing large electronic archives.

The "topics" signifies the hidden, to be estimated, variable relations that link words in a vocabulary and their occurrence in documents. A document is seen as a mixture of topics. Topic models discover the hidden themes through out the collection and annotate the documents according to those themes. Each word is seen as drawn from one of those topics. Finally, A document coverage distribution of topics is generated and it provides a new way to explore the data on the perspective of topics.

2.3. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar [3]. LDA has made a big impact in the fields of natural language processing and statistical machine learning and has quickly become one of the most popular probabilistic text modelling techniques in machine learning.

Intuitively in LDA, documents exhibit multiple topics [4]. In text pre-processing, we exclude punctuation and stop words (such as, "if", "the", or "on", which contain little topical content). Therefore, each document is regarded as a mixture of corpus-wide topics. A topic is a distribution over a fixed vocabulary. These topics are generated from the collection of documents [5]. For example, the sports topic has word "football", "hockey" with high probability and the computer topic has word "data", "network" with high probability. Then, a collection of documents has probability distribution over topics, where each word is regarded as drawn from one of those topics. With this document probability distribution over each topic, we will know how much each topic is involved in a document, meaning which topics a document is mainly talking about.

A graphical model for LDA is shown in Figure 1:

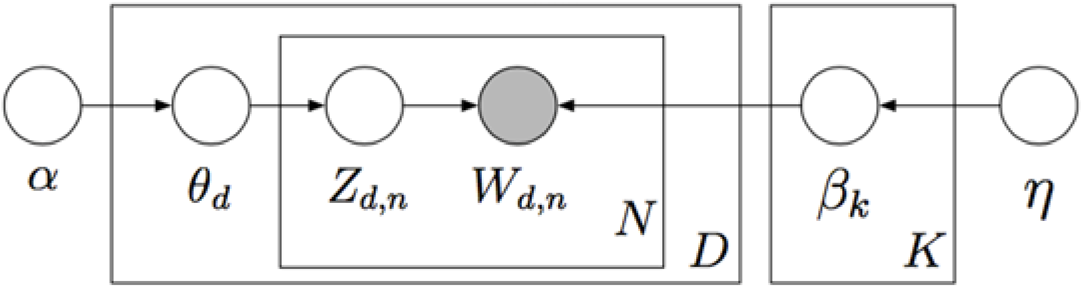


Figure 1. Graphic model for Latent Dirichlet allocation

As the figure illustrated, we can describe LDA more formally with the following notation. First, α and η are proportion parameter and topic parameter, respectively. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary. The topic proportion for the d th document are θ_d , where $\theta_{d,k}$ is the topic proportion for topic k in document d. The topic assignments for the d th document are Z_d, where $Z_{d,n}$ is the topic assignment for the nth word in document d. Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary.

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables:

$$p(\beta_{1:K},\theta_{1:D},z_{1:D},w_{1:D})=\prod_{i=1}^{K}p(\beta_i)\prod_{d=1}^{D}p(\theta_d)(\prod_{n=1}^{N}p(z_{d,n}\mid\theta_d)p(w_{d,n}\mid\beta_{1:K},z_{d,n}))$$

Notice that this distribution specifies a number of dependencies. The topic assignment $Z_{d,n}$ depends on the per-document topic distribution θ_d; and the word $w_{d,n}$ depends on all of the topics $\beta_{1:K}$ and the topic assignment $Z_{d,n}$.

2.4. Jensen-Shannon Divergence

In probability theory and statistics, the Jensen-Shannon divergence is a popular method of measuring the similarity between two probability distributions. It is also known as information radius or total divergence to the average. It is based on the Kullback-Leibler divergence. The square root of the Jensen-Shannon divergence is a metric often referred to as Jensen-Shannon distance [6].

For discrete probability distributions P and Q, Kullback-Leibler divergence of Q from $P$is defined to be:

$$D_{KL}(P\parallel Q)=\sum_i P(i)\log\frac{P(i)}{Q(i)}$$

So, the Jensen-Shannon divergence of Q from P is defined by:

$$JSD(P\parallel Q)=\frac{1}{2}D_{KL}(P\parallel M)+\frac{1}{2}D_{KL}(Q\parallel M)$$

where $M=\frac{1}{2}(P+Q)$

Jensen-Shannon divergence measures the similarity between two distributions. By applying Jensen-Shannon divergence to the topic assignment for the dth document Z_d, it will allow us to measure the distance and similarity between each document.

3. DESIGNS AND EXPERIMENTS

In order to apply topic modelling and propose a new text mining solution on topics, we have designed two experiments fulfilling our goal. The first one is to use topic modelling manage and explore Wikipedia, and the second is a Twitter application on topic modelling. The former is a new solution on a typical problem and the latter is building up a new model on Twitter data analysis.

3.1. Wikipedia on Topic Modelling

3.1.1. Experiment Overview

Wikipedia is a free-access, free-content Internet encyclopaedia, supported by non-profit Wikimedia Foundation. It has millions of articles for people to search, explore or even edit. In this experiment, the text data is from simplified Wikipedia (English version) with over 200,000 articles. By applying Latent Dirichlet allocation (LDA) and topic modelling, a solution of topic searching, exploring and recommending system will be achieved.

3.1.2. Data Pre-processing

The simplified Wikipedia English version is free for download from Wikipedia Foundation database backup dumps. The backup is in a format of XML. The first step of data pre-processing is to parse the XML file and extract the text data. R package *XML* provides a series of function parsing XML file. By using those functions, we will get a relatively clear data of all the articles in a data frame.

The next step is text-cleaning process. The purpose of text cleaning is to simplify the text data, eliminating as much as possible language dependent factors. Articles are written in natural language for human to understand. But in text mining, those data are not always easy for computers to process. In this experiment, there are three steps in text cleaning:

- Tokenization: a document is treated as a string, removing all the punctuations and then partitioned into a list of tokens.

- Removing stop words: stop words such as "the", "if", "and" ... are frequently occurring but no significant meanings which need to be removed.

- Stemming word: stemming word that converts different word form into similar canonical form. For example, computing to compute, happiness to happy. This process reduces the data redundancy and simplifies the later computation [7].

3.1.3. Model Training

The training process requires R package *topicmodels* with its package dependencies (*tm* and others) to be loaded. An LDA model of simplified English Wikipedia on a sample of 1000 articles with more than 1000 characters, returned after 2000 iterations of Gibbs sampling, with $K = 50$ topics, and Dirichlet hyper-parameters $\beta = 0.1$ and $\alpha = 50 / K$. Meanwhile, topic distribution coverage for each document is generated. This distribution represents how much each document is related to each topic. A new way of search and explore documents over topics can be implemented.

Table 1. A few selected topics generated from Wikipedia topic distribution

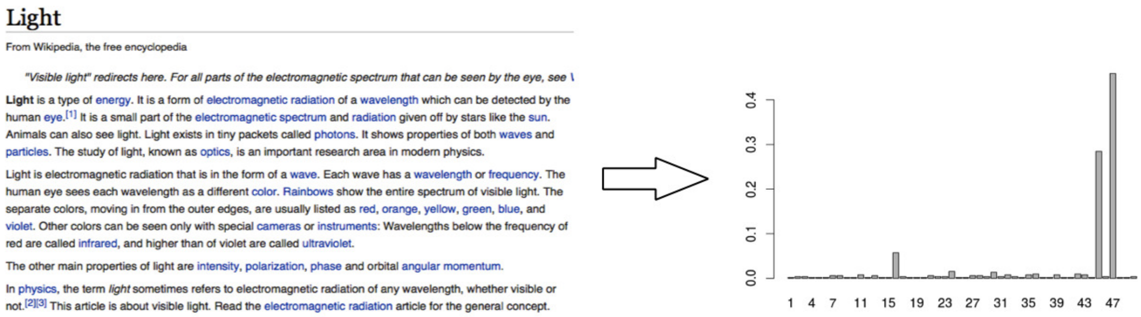
Topic 2	Topic 13	Topic 22	Topic 31	Topic 46
athlete	album	universal	movie	hurricane
olympia	song	college	categories	major
field	music	school	fiction	season
summer	record	categories	solstice	minor
track	band	new	alien	key
men	release	economic	star	storm
metre	single	institut	sun	tropic
image	rock	educate	direct	chord
women	singer	science	drama	verse
medal	pop	work	southern	end

Table 1 shows five selected topic terms after the model is trained, where top ten terms are listed for each topic. With LDA training, the terms in the same topic tend to be similar. Formally speaking, they are highly associated. For example, topic 13 is about music, topic 22 is about education and topic 46 is about weather. This topic distribution provides a way to search topic and explore between topics in order to find the document the user is looking for.

After the model is built, Jensen-Shannon divergence is applied to calculate the similarity of each distribution. Sorting the similarity of one document between every other distribution, a topic recommender system can be implemented.

3.1.4. Results

Here is an example of article *Light* from the experiment. The original article is shown in the left part of Figure 2, which can be also accessed in simple Wikipedia online (the data for this experiment is retrieved as a backup version on 11/1/2016). After the model is trained, we got a series of article distribution over each topic. The right part of Figure 2 is the bar plot of article *light* topics distribution. In total of 50 topics, we can easily find there are 3 topics with obviously high probabilities – topic 47, topic 45 and topic 16. Table 2 shows the top 5 probabilities topics. These probabilities are how tightly this article is associated with each topic. Table 3 shows the terms in these 5 topics with 10 terms each.



Light

From Wikipedia, the free encyclopedia

*"Visible light" redirects here. For all parts of the electromagnetic spectrum that can be seen by the eye, see *

Light is a type of energy. It is a form of electromagnetic radiation of a wavelength which can be detected by the human eye.[1] It is a small part of the electromagnetic spectrum and radiation given off by stars like the sun. Animals can also see light. Light exists in tiny packets called photons. It shows properties of both waves and particles. The study of light, known as optics, is an important research area in modern physics.

Light is electromagnetic radiation that is in the form of a wave. Each wave has a wavelength or frequency. The human eye sees each wavelength as a different color. Rainbows show the entire spectrum of visible light. The separate colors, moving in from the outer edges, are usually listed as red, orange, yellow, green, blue, and violet. Other colors can be seen only with special cameras or instruments: Wavelengths below the frequency of red are called infrared, and higher than of violet are called ultraviolet.

The other main properties of light are intensity, polarization, phase and orbital angular momentum.

In physics, the term *light* sometimes refers to electromagnetic radiation of any wavelength, whether visible or not.[2][3] This article is about visible light. Read the electromagnetic radiation article for the general concept.

Figure 2. Topics Distribution of Article *Light*

Table 2. Top 5 Topic Probabilities of Article *Light*

Topics	Topic 47	Topic 45	Topic 16	Topic 24	Topic 30
Probabilities	0.05692600	0.45920304	0.28462998	0.01518027	0.01328273

Table 3. Top 5 Topic Terms of Article *Light*

Topic 47	**Topic 45**	**Topic 16**	**Topic 24**	**Topic 30**
light	use	color	computable	cleaner
beamline	one	style	equation	people
beam	also	background	fluid	use
radiated	can	hex	categories	clean
don	people	rgb	image	make
synchrotron	mania	magenta	program	chemical
wavelength	call	ffffff	protocol	thing
physical	time	fuchsia	mathematical	paint
station	two	red	function	made
carlo	like	pink	design	put

So just like this example, every article of the whole collection is represented as a vector of probabilities over 50 topics. This is the core data of our model, where we can do all sorts of applications. Here is an example of finding the most related article of article *Light*. We will use Jensen-Shannon divergence to calculate the distance between article *Light* and every other article. The shortest distance will be the most related article. After calculating the distance, Table 4 shows the top 10 of the shortest articles. To be noted, the distance should have been the square root of the distance below, but to simplify the calculation and higher the accuracy, we will stay with the squared number, as there would be no difference on comparing the closet distance.

Table 4. Top 10 Shortest Distance of Article *Light*

Articles	Article 855	Article 820	Article 837	Article 299	Article 911
Distance^2	0.0000000	0.2271741	0.3404084	0.3881467	0.4583745
Articles	Article 341	Article 287	Article 328	Article 544	Article 606
Distance^2	0.4671845	0.4728383	0.4802241	0.4803357	0.4874499

As shown on Table 4, these 10 articles are the closest distance with Article *Light*. The distance of Article 855 is 0, because it is article *Light* itself. So Article 820 is the closest distance with *Light*, meaning their contents most related. Meanwhile, what we get is a sorted list of closest distance and will also work if we require more than one most related article. This method is based on calculating the probabilities each meaningful word in the model. So the accuracy is much better than calculating the keywords, or titles, which is widely used on many documents management system. Lastly, if we look up the title of Article 820, it is *Beamline*. Based on the 2000 article dataset, it is a convincing answer to the most related article of *Light*.

3.2. Twitter Data Analysis

3.2.1. Experiment Overview

Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". Currently, Twitter has more than 332 million active users posting 340 million tweets a day. Twitter has a big impact in everyday life. Twitter mining is not only a big research task in computer science or statistics, but also a key factor on social and business research. In this experiment, by applying LDA and topic modelling, a deep research and analysis on Twitter users is proposed. A detail model of Twitter user's personality and preference will be inferred.

3.2.2. Data Pre-processing

Generally, a tweet from any twitter user is public and free to read for everyone. Having registered as a Twitter application developer will allow you to access all the tweets and a set of APIs to manipulate the data [8]. The first step is to collect tweets from the users. By using the APIs and R package *twitteR*, a sample of 10,000 valid users is gathered into a data frame. To improve the data quality, a standard is set up for the twitter user to be a "valid" user:

- The user profile is an unprotected, which means user's information and tweets is public to everyone. If the user sets the profile to be protected, his or her information cannot be gathered by developers.

- The user has at least 100 tweets. According to a statistical research from Twitter on January 2012, the average length of a tweet is 67.9 characters. Therefore, before pre-processing, there are 6790 characters for one user in a sample. Less than 100 tweets per user will lower the calculation quality.

- The user must use English as major language in the tweets. Some odd non-English word will not affect the model, but a number of total non-English users' data will longer the pre-processing time, add confusion to the topics, or even mess up the result.

The text cleaning process is basically the same as the Wikipedia experiment. Tokenization, stop words removing and word stemming is required in this process. However most tweets are oral and informal language, a few details need to be noticed:

- Some tweets may contain URLs, using hash tags on a topic, using @ to mention other users. In text cleaning process, these situations need particular functions to remove or parse.

- Some Internet terms, such as "LOL", "OMG", "BTW", are abbreviation of a phrase. Those terms can treat as stop words to delete.

- Some words are written as shorthand, such as "ppl" (people), "thx" (thanks), "fab" (fabulous). Those words need to stem to the original form.

3.2.3. Model Training

The training process is also similar to the Wikipedia experiment. However, Twitter data are formed with natural daily language, which has a narrow topic range compared to Wikipedia. The topics number is less and not with equally clear boundaries. Nevertheless, the topic model still has a good performance and the coverage distribution can easily illustrate the user's personal interest. Here is an experiment a sample of 100 twitter user with more than 100 tweets, returned after 2000 iterations of Gibbs sampling, with 30 topics.

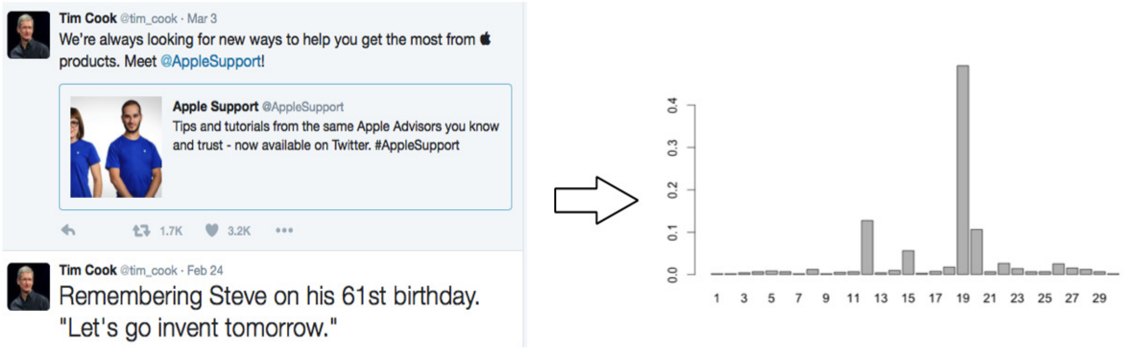
Table 5. A few selected topics generated from Twitter topic distribution

Topic 1	**Topic 5**	**Topic 9**	**Topic 17**	**Topic 28**
halifax	check	stream	follow	weight
nova	reward	live	win	loss
scotia	one	league	enter	diet
man	kangaroo	communities	canada	news
refuge	new	check	retweet	lose
media	facebook	ea	card	natural
say	get	design	sunday	tip
woman	post	weekend	away	plan
central	photo	chat	chance	health
student	coffee	gold	donate	techno

Table 2 shows five selected topics terms after the Twitter topic model is trained. Similar to the previous experiment, the LDA model has a good performance on dividing topics. But as we expected, twitter data is limited by the daily language that leads to less clear boundaries as Wikipedia topics. For example, in topic 5, it is hard to label this topic into a particular category. However, the Twitter application is successful on building up topic models over users, and it will certainly benefit the statistic analysis and even a big impact on social and business research.

3.2.4. Results

This experiment has a similar structure with the Wikipedia articles. Instead of each article, we will treat every user's tweets as an article. With the topic model, we can also calculate the distribution over each topic. Here is an example of Tim Cook's Twitter. As shown in Figure 3, the left part is the Twitter user, and the right part is the bar plot of topic distribution. This distribution represents what kind of topic the user talks more and more interested. Table 6 shows the top 5 topics of Tim Cook talks most about. By applying Jenson-Shannon divergence to calculate the distance, we can find the people that talks the most similar topics or even with similar personality.



Tim Cook @tim_cook · Mar 3
We're always looking for new ways to help you get the most from products. Meet @AppleSupport!

Apple Support @AppleSupport
Tips and tutorials from the same Apple Advisors you know and trust - now available on Twitter. #AppleSupport

1.7K 3.2K

Tim Cook @tim_cook · Feb 24
Remembering Steve on his 61st birthday. "Let's go invent tomorrow."

Figure 3. Topics Distribution of Twitter of Tim Cook

Table 6. Top 5 Topic Terms of Tim Cook's Tweets

Topic 19	Topic 12	Topic 20	Topic 15	Topic 22
day	new	apple	custom	summer
today	great	iphone	service	student
thank	canada	trump	expect	job
will	join	deal	job	american
great	congrat	say	product	wed
time	congratulations	app	donet	still
get	communities	product	photo	campus
ea	forward	vs	price	fair
can	proud	court	easier	act
new	event	camera	facebook	hall

4. CONCLUSIONS AND FUTURE WORK

In this paper, an introduction of text mining and topic model LDA is represented. We proposed two experiments, which built up topic models on Wikipedia articles and Twitter users' tweets. A brief introduction of each experiment including overview, pre-processing and model training is given and analysed.

With these data and model foundation, a number of future works can be done for further research and experiment.

- As the limitation of the computation power, this research is based on a relatively small sample by the time we start writing. However, the result is quite convincing even with the small size. Applying to a larger dataset will more likely achieve better results.
- An application on topic modelling to manage, search and explore offline Wikipedia articles could be implemented.
- A full research on Twitter users' interest could be applied. Further more, this application could be a useful tool for social and business research
- In Twitter application on topic modelling, we ignore the pictures users posted. What if we can combine image processing and topic model to provide a better performance?

REFERENCES

[1] Martin Ponweiser (2012) Latent Dirichlet Allocation in R, Vienna University of Business and Economics.

[2] Bettina Grun, kurt Hornik (2011) "topicmodels: An R Package for Fitting Topic Model", Journal of Statistical Software Vol. 40, No. 13.

[3] Qi Jing (2015) Searching for Economic Effects of User Specified Event Based on Topic Modelling and Event Reference, Jordery School of Computer Science, Acadia University.

[4] David M.Blei (2012) "Probabilistic Topic Models", Communications of the ACM Vol. 55, No. 4, pp77-84.

[5] David M.Blei, John D. Lafferty (2006) “A Correlated Topic Model of Science”, Annals of Applied Statistics Vol. 1, No. 1, pp17-35.

[6] Jianhua Lin (1991) “Divergence Measures Based on the Shannon Entropy”, IEEE Transactions on Information Theory Vol. 37, No. 1, pp145-151.

[7] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan (2010) “A Review of Machine Learning Algorithms for Text-Document Classification”, Journal of Advances in Information Technology Vol. 1, No. 1, pp4-20.

[8] Yanchang Zhao (2015) R and Data Mining, http://www.rdatamining.com.

AUTHORS

Zhou Tong is currently a master student of computer science at Acadia University, Canada. His research is focusing on text mining.

Haiyi Zhang received his MS degree in 1990 from the Computer Science department of New Jersey Institute of Technology of USA, and his Ph.D in 1996 from Harbin Institute of Technology in China. He was a post-doctor in information department of ABO, Finland in 2000. His research interests are machine learning, data mining. He has more than 50 academic papers published. Currently he is an associate professor at Acadia University, Canada.