Journal of Engineering and Applied Sciences 12 (16): 4123-4126, 2017
ISSN: 1816-949X

Image Categorization using Topic Modeling with the Latent Dirichlet Allocation

[1]Ghaidaa A. Al-Sultany and [2]Suha Kamal
[1]Department of Information Network,
[2]Department of Software, Babylon University, Babel, Iraq

Abstract: In this study, the issue of text modeling was considered with an annotated textual description for diverse categories of images. A topic modeling with the Latent Dirichlet Allocation technique were proposed on a set of images text descriptions to realize the precise topic for each set of images with respect to their attached descriptions from which classifying new images based on the observed topics would be more facilitated. The research was evaluated on the benchmark CLEF dataset; the results were encouraging with regards to the enhancing image retrieval using topics extraction.

Key words: Topic modeling, image text descriptions, LDA, descriptions, encouraging, benchmark

INTRODUCTION

In recent decades, huge amount of textual data are available and more information can be obtained. However, the most difficult issue is how to discover that information efficiently and accurately.

A number of probabilistic analyses on text data have arisen for covering various text-mining problems such as topic modeling. Topic modeling provides a significant manner for summarizing huge amount of data through organizing, understanding and involving new relations to discover the hidden topics of those data. Topic modeling is considered one of the best techniques that research powerfully on a variety of text as it has the ability to discover the hidden topics that found in text.

Surfing websites for particular images have become a trend for the users of Internet. Consequently, categorized them in terms of their objects, attached text and their meaning suitably is crucial. Topic Modeling has been one of the techniques that were applied for image clustering and extracting the integrated topics of the related images. Latent Dirichlet Allocation-LDA is one of a generative probabilistic model for collections of discrete data such as text corpora in which it is dealing in particular for topic modeling by modeling every item of a collection as a finite mixture over a set of topics.

In this study, the issue of text modeling was considered with an annotated textual description for diverse categories of images. Itsgoal is to realize the precise topic for each set of images with respect to their attached short descriptions. This research is a part of additional work in which an efficient matching process among large collection of text documents and images with respect to their topics would be accomplished.

Literature review: Topic modeling were utilized by the researchers (Nguyen *et al*., 2010) as a method for image annotation by using two models: a model of feature-word distributions based on multiple instance learning and mixture hierarchies which it is similar to the way of supervised multiclass labeling. And a model of word-topic distributions based on probabilistic Latent Semantic Analysis-pLSA they used UWDB and Core l5 k dataset and represented images as a set of real feature vectors in our method used another data set and lda approach (Nguyen *et al*., 2010). Also, Chong *et al*. (2009) used both the annotations of images and class labeling to develop a new probabilistic model for jointly modeling the images. They derive an approximate inference and estimation algorithms based on variation methods for classifying and annotating new images. Wang *et al*. (2014), the research used topic modeling for tagging images using both the statistics of tags and the visual similarities of images in the source data. This research was done by applying the regularized Latent Dirichlet Allocation-rLDA. Another work for classifying a collection of images were done by Zang *et al*. (2014) they performed the LDA for extracting features from set of images. They attempted to generate a codebook before applying it to the LDA to infer topics and then applied the labeled images to classifier as a training data (Zang *et al*., 2014). Likewise, Anh *et al*. (2016) proposed a new visual search system to retrieval the relevant images their method was focused on encoding technique based on soft-assignment of local features to convert an entire image into a single vector. They used thena probabilistic topic model to extract the object and background regions from collection of images (Anh *et al*., 2016).

In our research a topic modeling with LDA technique were proposed on a set of images text descriptions for

Corresponding Author: Ghaidaa A. Al-Sultany, Department of Information Network, Babylon University, Babel, Iraq

extracting the proper topics of those images from which classifying new images based on the observed topics would be more facilitated.

MATERIALS AND METHODS

Latent Dirichlet allocation: There are many collections of text that have similar meaning in their content but apparently it looks different in their structure and syntax. The latent Dirichlet allocation is the technique that is emerged to address the needs for topic modeling to assign documents with a certain topics; hidden or latent variables are inferred using posterior inference. Posterior inference is where the hidden variables are estimated based on relevant background evidence (Abbey, 2015). In the LDA the topics are hidden variables and are inferred from the words in the text documents as it considers aligning the words with specific probabilities (Blei, 2012). It groups words together based on how likely they are to appear in a text document together. Correlated topic models explore the correlation of words to other words within a text as shown in Fig. 1.

Where α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the parameter of the Dirichlet prior on the per-topic word distribution from a collection of text documents, it infers:

- Per-word topic assignment ($z_{d,n}$)
- Per-document topic proportions (θ_d)
- Per-corpus topic distributions (β_k)

The system modeling: The proposed system considers topics based image categorization as a significant area where the goal is to assign an image with specified set of topics. The LDA algorithm was implemented for learning the images textual descriptions to mine the multinomial distribution of the images topics.

Preprocessing: In our research, the data were extracted from the dataset namely CLEF which it is a freely available benchmark for image retrieval (Grubinger, 2007). This dataset consists of set of color images with annotations available in three languages. Each image is attached with an XML data (Algorithm 1) that comprise of a number of sentences describing it. The minimum number of sentences is one as no image should be without an alpha numeric representation and should not exceed six sentences. A parsing process on the annotated XML format file was performed for reading and extracting the file tags. According to the proposed work, the tags "description" and "title" were chosen to retrieve their content as they both contain textual features that can be passed to the LDA algorithm for analyzing and extracting

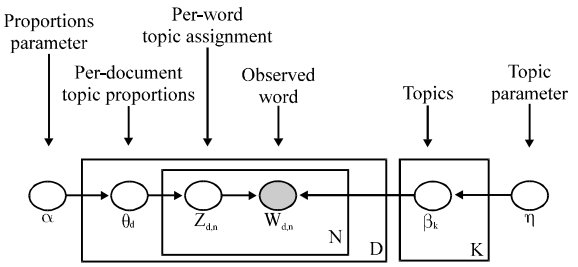

Fig. 1: A graphical model representation of the Latent Dirichlet Allocation-LDA

relevance topic of. After extracting all the description text of all the images in the dataset, the text were preprocessed by removing all the stop words which it would not help in the topic extracting process and the remained text are gathered as a bag of words.

Algorithm 1; sample of annotated xml file correspond to an image:

<DOC><DOCNO>annotations/02/2007.eng</DOCNO><TITLE>Sleeping alligator in the jungle</TITLE><DESCRIPTION>An alligator in the water between leafs;</DESCRIPTION><NOTES></NOTES><LOCATION>Puerto Maldonado, Peru</LOCATION><DATE>November2002</DATE><IMAGE>images/02/2007.jpg</IMAGE><THUMBNAI L>thumbnails/02/2007.jpg</THUMBNAIL></DOC>

Topic extraction: For topic extracting process, LDA was applied on the whole preprocessed text with regards to the following stages.

Stage 1: Initially, the number of topics should be specified at first. It can either use an informed estimate (e.g., results from a previous analysis) or in terms of how much data are you going to treat with (as a training data). In our research, the number of topics was inferred by eyeballing the documents.

Stage 2: In addition, every word in the text should be assigned to one of the initial temporary topics. Topics assignment process is appliedin a semi-random manner according to the Dirichlet distribution (regularly indicated Dir (α) is a group of ceaseless multi variate likelihood appropriations parameterized by a vector α of positive real). This also means that if a word appears twice each word may be assigned to different topics.

Stage 3: The system learning where the LDA were utilized to learn the topic representation for the all images text descriptions in the dataset. In addition, all the words in all the description were checked and updated in terms of their relations and similarity to a certain topic. For each word, its topic assignment is updated based on two criteria:

- The word's prevalent across topics
- Thetopic's prevalent across the image text description

The process of checking topic assignment is repeated for each word in every text, cycling through the entire collection of text descriptions multiple times. This iterative updating is the key feature of LDA that generates a final solution with coherent topics as depicted in the Algorithm 2.

Algorithm 2; the learning process of the LDA:

Input: Set of files contains images (text descriptions)
Output: Subjects consist of some words for each
Step 1: For each file f
1- for every subject s, process two things
- p(subject s/file f)
- p(word w/subject s)

2- Experience every word w in file f.
Step 2: Generate subject for w by computing p(subject s|file f)*p(word w| subject s)
Step 3: Repeat the Step 2 an extensive number of times until assignments are quite great
Step 4: Estimate the subject mixtures of each file by
For every subject, checking the extent of words doled out to it inside the file
Counting the extent of words appointed to every subject general

RESULTS AND DISCUSSION

After the preprocessing stage on the collection of images text description files, 100 images text descriptions files were chosen for testing the proposed system with five main topics. Each file was accumulated with some distributed topics where each topic defines a distribution over words. We assumed that particular topics are associated with a collection and that each text description defines a distribution over (hidden) topics. The number of words in topics was depended on the size of text descriptions that would be usedin the training process.

The 80% of the collection of images text descriptions was grouped for training process and 20% were left for testing process. The results were encouraging through obtaining a good outcome of those topics that describe the images collection. The held-out probability estimation was used for evaluating the tested images as shown in Fig. 2a and j for a set of images with their description taken from IAPR TC-12 benchmark dataset. And the resulted topics in terms of 8 and 5 words in each extracted topics as listed in the.

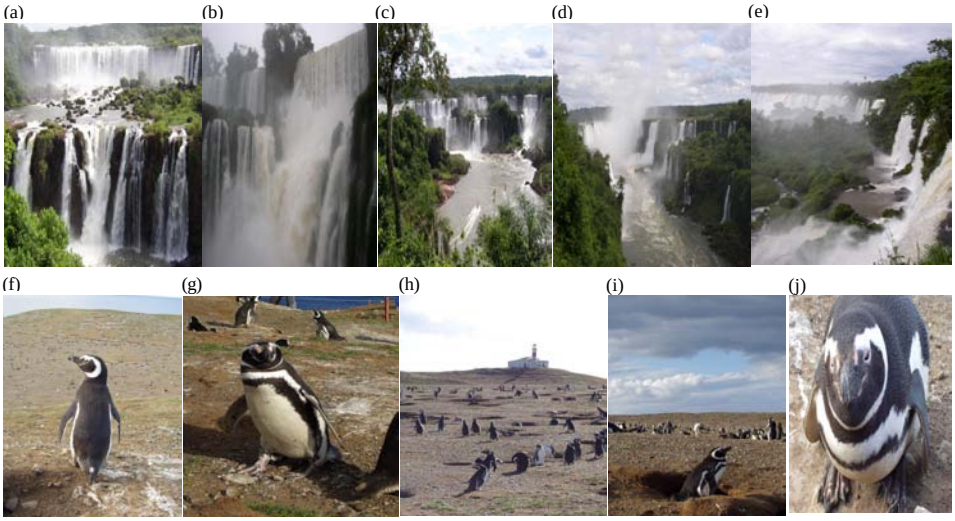

Fig. 2: Set of images with their text description: a) <DESCRIPTION> view of giant waterfalls in the middle of the jungle; <DESCRIPTION>; b) <DESCRIPTION> voluminous bodies of water are falling in the middle of the jungle; <DESCRIPTION>; c) <DESCRIPTION> view of giant waterfalls in the middle of the jungle; two boats in the water and a tree in the fore ground on the left; <DESCRIPTION>; d) <DESCRIPTION> view of giant waterfalls in the middle of the jungle; <DESCRIPTION>; e) <DESCRIPTION> numerous giant, thunderous waterfalls in the middle of the jungle; <DESCRIPTION>; f) <DESCRIPTION> a penguin on a brown, bald island with many other penguins in the background; <DESCRIPTION>; g) <DESCRIPTION> penguins on a brown, bald island; <DESCRIPTION>; h) <DESCRIPTION> penguins on a brown, bald island with a lighthouse in the background; <DESCRIPTION>; i) <DESCRIPTION> a penguin in his nest on a brown, bald, rocky island with many other penguins in the back ground; <DESCRIPTION> and j) <DESCRIPTION> close-up photo of a black and white penguin; <DESCRIPTION>

The extracted topics with 8 words:

- Middle; penguins; flower; snow; brown
- Jungle; bald; leaves; covered; shades
- Waterfalls; brown; green; sky; canyon
- Giant; terrain; blooms; wooded; sky
- Thunderous; island; trunks; foreground; blue
- Water; penguin; front; bank; clouds
- Falling; flat; red; grey; giant
- Bodies; white; big; blue; white

The extracted topics with 5 words:

- Jungle; flower; brown; penguins; snow
- Middle; leaves; shades; bald; covered
- Waterfalls; green; canyon; brown; sky
- Giant; blooms; sky; terrain; foreground
- Thunderous; trunks; blue; island; grey

CONCLUSION

Basically, topic modeling works powerfully on a variety of text as it has the ability to discover the hidden topics that found in text. The latent dirichlet allocation has been approved as one of the best techniques for topic modeling from which it was implemented in this research, for learning the images textual descriptions to mine the multinomial distribution of the images topics. The research was tested using the Benchmark CLEF Dataset and the topics results were encouraging with respect to the topics based image retrieval. This research is a part of additional work in which an efficient matching process among large collection of text documents and images with respect to their topics would be accomplished. For a group images topic modeling found topic (some of word) to describe the content of them that used to retrieval similar image or to classify the image with the similar another images.

REFERENCES

Abbey, R.K., 2015. The statistics of topic modeling. Master Thesis, University of Canterbury, Christchurch, New Zealand.

Anh, N., D. Tu, M. Dinh, K. Rasel and Y. Lee, 2016. Topic modeling and improvement of image representation for large-scale image retrieval. Inf. Sci., 366: 99-120.

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM., 55: 77-84.

Chong, W., D. Blei and F.F. Li, 2009. Simultaneous image classification and annotation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR 2009, June 20-25, 2009, IEEE, Miami, Florida, ISBN:978-1-4244-3992-8, pp: 1903-1910.

Grubinger, M., 2007. Analysis and evaluation of visual information systems performance. Ph.D Thesis, Victoria University, Wellington, New Zealand.

Nguyen, C.T., N. Kaothanthong, X.H. Phan and T. Tokuyama, 2010. A feature-word-topic model for image annotation. Proceedings of the 19th ACM International Conference on Information and Knowledge Management, October 26-30, 2010, ACM, Toronto, Ontario, Canada, ISBN:978-1-4503-0099-5, pp: 1481-1484.

Wang, J., J. Zhou, H. Xu, T. Mei and X.S. Hua *et al*., 2014. Image tag refinement by regularized latent Dirichlet allocation. Comput. Vision Image Understanding, 124: 61-70.

Zang, M., D. Wen, K. Wang, T. Liu and W. Song, 2014. A novel topic feature for image scene classification. Neurocomput., 142: 282-290.