Topics Discovery in Text Mining

Anacleto Correia[1(✉)] and António Gonçalves[2]

[1] CINAV, Center of Naval Research, Portuguese Naval Academy, Almada, Portugal
cortez.correia@marinha.pt
[2] INESC-ID, Instituto Superior Técnico, Lisbon University, Lisbon, Portugal
antonio.leonardo.goncalves@tecnico.ulisboa.pt

Abstract. Text data has been increasingly growing in the last years, due to the advances of web based technologies that enable the publishing of an overwhelming amount of data. One can say that, many knowledge about the world in text data, besides being stored in articles and books, is also available on blogs, tweets, web pages. This paper overviews some general techniques for text data mining, based on text retrieval models, that can be applicable to any text in natural language. The techniques are targeted to problems requiring minimum or no human effort. These techniques, which can be used in many applications, allow the discovery of main topics of a document in text data with different levels of granularity.

Keywords: Applied statistics · Bayesian theory and methods · Data and text mining

1 Introduction

The Web increased the textual revolution by made available a great amount of on-line information. Information and knowledge about almost any subject is encoded in text data available on-line, in articles or books. The process of text mining refers to extraction of high quality information from text data. Quality of collected information is related with elicited patterns and trends. Text mining usually involves the process of structuring the input text (through parsing, and adding or removing linguistic features), deriving patterns within the structured data, and eventually analyzing and interpreting the output.

When looking at text data in any support, people may have expectations regarding its content, which can be: (1) to discover aspects about a specific natural language, its usage, as well as the patterns on it; (2) mining knowledge from content of text data about the observed world, getting the essence of it or extracting information about relevant aspects of the world; (3) mining knowledge about an observer, which means using text data to infer properties of a person; and (4) making predictive analytics using text mining to infer real world variables. When real world variables are inferred they can also use intermediate results of other predictions. So, multiple types of knowledge can be mined from text in general [2–4].

Á. Rocha et al. (eds.), *Recent Advances in Information Systems and Technologies*,
Advances in Intelligent Systems and Computing 569, DOI 10.1007/978-3-319-56535-4_25

Text mining techniques are surveyed in several works [7–13]. In this paper we focus on statistical approaches regarding topic mining used for extracting specific knowledge from documents [14]. The next sections describe mining techniques for topic mining and analysis, a way to analyze content of text [17].

2 Mining of Topics

In the current context, a topic [18] is a main idea discussed, a theme or subject of discussion or conversation in text data. The topic can be mined at different levels of granularity (e.g. a sentence, an article, a paragraph or several articles). In general, topic mining is the process of collecting knowledge about the world. This process involves: (2) find out the covered topics in analyzed documents and; (2) measure the extent of the coverage.

The setup of topic models [3] begins by taking test data as input. After the mining process, a set of topics is the output, with each topic characterized by a word distribution, as well as the proportions by which the topics are covered in each document.

In a formal way, the topic mining problem can be defined as having as input a collection of n text documents $C = d_1, \ldots, d_n$ and the number k of topics, with C denoting a text collection, and d_i each text article.

The output of the process is the k discovered topics, $\theta_1, \ldots, \theta_k$, and the coverage degree of topics by each document, denoted by $\pi_{i1}, \ldots, \pi_{ij}$, with each π_{ij} being the probability of the document d_i covering topic θ_j. Topics' coverage are constrained by $\forall\, i \in [1,N], \sum_{j=1}^{k} \pi_{ij} = 1$, i.e., the topics in each document must be within the discovered topics.

The following sections present different ways of discovering the topic θ_j, beginning with a simplistic approach and introducing, step by step, more realistic assumptions.

2.1 A Topic as a Simple Term

The simplistic approach is a topic being a simple term (e.g. a word). The first step to perform in mining these simple topics is the elicitation of candidate terms by parsing the text data of documents collection. Next, a scoring function is used to evaluate the adequacy of the terms to actually being one of the k topics. Eventually, the degree of topics' coverage of each document in the collection is determined.

Since choosing frequent terms as candidate topics is preferable, some approaches from information retrieval are used, namely Term Frequency (TF) and Inverse Document Frequency (IDF) weighting. Care should be taken regarding semantically equivalent terms when building the scoring function. Redundancy of terms must be tackled since not desirable. This can be done using for instance, the maximal marginal relevance ranking algorithm. The process of selecting terms can be done moving through the list of scored terms and find the threshold of the k top ranked terms, thus avoiding to choose semantically close terms.

The degree of topic coverage π_{ij}, in each document, can be computed by counting the normalized occurrences of the terms in the document, in order that each topic coverage sums to one, as formalized by $\pi_{ij}=\frac{\text{count}(\theta_j,d_i)}{\sum_{L=1}^{k}\text{count}(\theta_L,d_i)}$. The expression can also be seen as a distribution of the topics for the document, and a characterization and the coverage of different topics by the document.

This approach, however, has some issues: (1) it does not consider synonyms words when the words belonging to a topic are counted; (2) the inherent ambiguity of certain words requires that their context must be considered to derive the corresponding topic; and (3) lack of expressiveness power in case of complex topics.

To solve these problems, more words must be used to describe a topic. The incompleteness of vocabulary coverage, by a topic represented as one term, can be addressed through the use of weights on words assigned to the topic. This allows to distinguish subtle differences in topics and fuzziest semantically related words. To solve the problem of word ambiguity, ambiguous words should be split, so the topic can be disambiguated. The next approach addresses the issues by using a probabilistic topic model.

2.2 A Topic as a Word Distribution

The basic idea in this approach is to improve the representation of the topic with word distributions. So, each topic become a word distribution with known probabilities that sum to one, for all the words (w) in the vocabulary (V), which can be denoted as $\forall\, j\in[1,k], \sum_{w=V} p(w|\theta_j)=1$. Another constraint imposed to the topic coverage, is that all the π_{ij}'s should sum to one for the same document.

As input, the topics' discovery problem has: (1) a collection of N text documents $C = d_1, \ldots, d_n$; (2) a vocabulary set $V = w_1, \ldots, w_M$; and (3) a specified number of k topics. The problem solution are: (1) the k topics; (2) each word distribution $\theta_1, \ldots, \theta_k$; (3) for each document d_i, the coverage of topics $\pi_{i1}, \ldots, \pi_{ij}$; and (4) the probability, π_{ij}, of each d_i cover each topic θ_j.

The general way of solving this text mining problem, using statistical modeling, is the generative model. The model data generation with a probabilistic model is given by P(Data|Model, Λ). The main idea is first design a model for data. The probabilistic model is designed next, to model how the data are generated. This gives the probability distribution of the data, with the particular model and parameters being denoted by $\Lambda = (\{\theta_1, \ldots, \theta_k\},$
$\{\pi_{11}, \ldots, \pi_{1k}\}, \ldots, \{\pi_{N1}, \ldots, \pi_{Nk}\})$.

After the set up of the model it can be fitted with actual data. Meaning that the most likely parameter values, Λ^*, can be estimated based on the data set. So, $\Lambda^* = \text{argmax}_\Lambda \text{p}(\text{Data}|\text{Model}, \Lambda)$, would maximize the probability of the observed data. The parameters are the outcome of the data mining algorithm that can then be used to discover knowledge from text.

Statistical Language Models (SLM) cover probabilistic topic models as a special cases, and can be regarded as probabilistic mechanism for generating text. The simplest language model is called the Unigram Language Model (ULM).

It simply assume that the text is generated word by word independently, i.e., $\mathrm{p}(w_1, w_2, \ldots, w_n) = \mathrm{p}(w_1)\,\mathrm{p}(w_2)\ldots\mathrm{p}(w_n)$.

With this assumption, the probability of the sequence of words is the product of the probability of each word, and the model has as many parameters as the number of words in the vocabulary. Assuming that there are n words, it means there are n probabilities, and the result of their sum is $\mathrm{p}(w_1)+\ldots\mathrm{p}(w_n) = 1$. So, the text is a sample drawn according to the word distribution. Since this model ignores the words' sequence it may not be suited for some kind of problems. However, it is quite sufficient for many tasks that involve topic analysis.

For the best estimation of the parameters of the Unigram Language Model, there are two different ways of estimating the parameters: the Maximum Likelihood (ML) estimator and the Maximum a Posteriori (MAP) estimator. The ML estimator gives the observed data the maximum probability and is formalized as $\hat{\theta} = \arg\max_{\theta} \mathrm{P}(X|\theta)$, where arg max means a function that returns the argument θ that gives the function maximum value of X.

When the sample is small, unseen words could get zero probability, which sometimes may not be reasonable if the distribution is to characterize topics. This problem is addressed through the Bayesian estimation, since it uses both the data, and prior knowledge about the parameters, such that $\hat{\theta} = \arg\max_{\theta} \mathrm{P}(X|\theta)\mathrm{P}(\theta)$. Where the prior is defined by $\mathrm{P}(\theta)$, which means that preferences are imposed on certain θ's. So, by using Bayes Rule, the MAP estimator is derived to combine the likelihood function with the prior and give the posterior probability of the parameters.

The MAP estimator is more general than the ML estimator because if the prior is defined as a non-informative it is reduced to the maximum likelihood estimated.

2.3 Mining a Single Topic from a Document

The Unigram Language Model is used for mining a single topic from one document. The generating model discovers the words' probabilities for the single topic using as data the document d, with x_i's as sequence of words from the document. The document is formalized as $d = x_1\, x_2 \ldots x_{|d|}$ with $x_i \in V = \{w_1, \ldots, w_M\}$. The model, is a word distribution that denotes the topic θ, with M parameters (as many as words in the vocabulary), and θ_i as the probability of choosing word w_i. The ULM is formally expressed as $\theta : \{\theta_i = p(w_i|\theta)\}, i = 1, \ldots, M$ with $\sum_1^M p(\theta_i) = 1$.

Because it is assumed the independence when generating each word, the probability of the document is the product of the probability θ_i of each word. Since some word might have repeated occurrences, the likelihood function is a product over all the unique words in the vocabulary. A counter function $\mathrm{c}(w, d)$ denotes the count of each word in document. So, the likelihood function is the probability of generating the whole document, given the model $_{p(d|\theta)=\prod_{i=1}^{M} \theta_i^{\mathrm{c}(w,d)}}$.

By maximizing the likelihood function $p(d|\theta)$, subjected to the constraint $\sum_{i=1}^{M} \theta_i = 1$, the $\hat{\theta}_i$'s for each word are found through the expression: $(\hat{\theta}_1,...,\hat{\theta}_M) = \arg\max_{\theta_1,...,\theta_M} p(d|\theta)$.

The analytical solution for the optimization problem is the normalized count of the words by the sum of all the counts of words in the document given by $\hat{\theta}_i = \frac{c(w_i,d)}{|d|}$.

3 Conclusions

Text mining, as an interdisciplinary field, benefits from contribution of several correlated disciplines. Text Mining allows the identification of patterns in large sets of data, by uncovering previously unknown, useful knowledge for decision making.

This work is concerned with overview of techniques for discovering the main topics in text documents. The surveyed techniques included a (1) simplistic approach, with a topic as a simple term, (2) a model where each topic is a word distribution with known probabilities, for all the words in the vocabulary, and (3) a model for mining a single topic from a document using the Unigram Language Model.

Acknowledgements. This work was supported by Portuguese funds through the *Center of Naval Research* (CINAV), Portuguese Naval Academy, Portugal.

References

1. Miner, G.: Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press, New York (2012)
2. Zhai, C., Massung, S.: Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. Morgan & Claypool, Williston (2016)
3. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
4. Lu, Y., Mei, Q., Zhai, C.: Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Inf. Retrieval **14**(2), 178–203 (2011)
5. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
6. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
7. Berry, M.W., Castellanos, M.: Survey of text mining: clustering, classification, and retrieval (2007)
8. Hotho, A., et al.: A brief survey of text mining. Proc. LDV Forum **20**, 19–62 (2005)
9. Tated, R.R., Ghonge, M.M.: A survey on text mining-techniques and application. Int. J. Res. Adv. Technol. (2015)
10. Berry, M.: Survey of text mining: clustering, classification, and retrieval (2003)
11. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. retrieval **2**(1–2), 1–135 (2008)
12. Patel, M.R., Sharma, M.G.: A survey on text mining techniques. Int. J. Eng. Comput. Sci. **3**(5), 5621–5625 (2014)

13. Inzalkar, S., Sharma, J.: A survey on text mining-techniques and application. Int. J. Res. Sci. Eng. (2015)
14. Jiang, S., Zhai, C.: Random walks on adjacency graphs for mining lexical relations from big text data. In: Proceedings of IEEE International Conference on Big Data (2014). doi:10.1109/BigData.2014.7004272
15. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, New Jersey (2012)
16. Zhai, C.: Exploiting context to identify lexical atoms-A statistical view of linguistic context (1997). arXiv preprint cmp-lg/9701001
17. Zhai, C.: Text Mining and Analytics, 25 May 2016. https://www.coursera.org/learn/text-mining
18. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2007)