JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Enhancement of Topics Modeling using Probabilistic Latent Semantic Analysis Learning

G.Ragu
Assistant Professor
Department of Computer Science and Engineering
SRM Institute of science and Technology,

Srinivas S
Department of Computer Science and Engineering
SRM Institute of Science and Technology,
Ramapuram, Chennai, India.

Vishnu Priya
Department of computer Science and Engineering
SRM Institute of Science and Technology,
Ramapuram, Chennai, India.

***Abstract*- Over the last years topic modeling has emerged as a powerful technique for organizing and summarizing big collections of documents or searching for patterns in them. However, privacy concerns arise when cross-analyzing data from different sources is required. Federated topic modeling solves this issue by allowing multiple parties to jointly train a topic model without sharing their data. Topic models simultaneously model the latent topic structure of large collections of documents and a response variable associated with each document. It is popular to detect hot topics which can benefit many tasks including topic recommendations the guidance of public opinions and so on. However, in some cases people may want to know when to re-hot a topic i.e. make the topic popular again. Also, it considers the continuous temporal modeling of topics since topics are changing continuously over time. Furthermore, a weighting scheme is proposed to smooth the fluctuations in topic re-hotting prediction. In this proposed system we explore task-centric topic model comparison considering how we can both provide detail for a more nuanced understanding of differences and address the wealth of tasks for which topic models are used. We derive comparison tasks from single-model uses of topic models which predominantly fall into the categories of understanding topics understanding similarity and understanding change. We abstract the model complexity in an interactive visual workspace for exploring the automatic matching results of two models investigating topic summaries analyzing parameter distributions and reviewing documents. The main contribution of our work is an iterative decision-making technique in which users provide document-based relevance feedback that allows the framework to converge to a user-endorsed topic distribution.**

I. Introduction

The extensive growth of social media in the past few years has caused people to join social media websites and contribute to the increasing amount of content on the Internet by sharing their daily activities. The huge amount of

data shared on social media allows us to use this data for prediction in various tasks. Many people share their day to day activities on social media. Such a collection of information might report a specific event; e.g. a player might score a goal in a football match and people might report this event on their twitter account. Therefore analyzing tweets in a specific time might identify this event. This makes event detection one of the popular tasks among researchers. The event detection task can be more challenging than it looks and it could be different from other social media analysis tasks Event detection can be used in various fields such as medicine emergency politics. The necessity of event detection in these fields comes from the fact that an important event is usually followed by a set of other events. For instance, a car accident is normally followed by traffic jams and casualties. Therefore, if the rescue team is informed earlier and arrives on-time they might prevent casualties. This indicates the importance of accurately detecting events within a suitable time interval. Event detection is normally performed using task-based or similarity-based approaches. Task-based methods first describe the problem that the system wants to solve. Then the system gathers information as needed and the classifier must be trained based on these data. Assume that we want to use a method in order to report a car accident. For this matter data about the accident must be collected in a specific time interval and then according to machine learning algorithms the model must be trained. This allows us to have the ability to detect an event in a specific topic. Similarity-based methods use a set of algorithms that are placed in a stream of data and can detect events by recognizing structures and similar patterns. They can detect various events using specific settings.

II. Objective and scope

With the rapid development and popularization of Internet technology how to obtain useful information from massive data has become a common concerned problem. Text mining technology can extract effective useful and valuable information from a large number of texts and it has gradually become one of the key technologies to solve the problem of topic discovery. Effective analysis of massive information on the network has also become a key research content by researchers in the eld of machine learning and data mining. This motivates us.

Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be defined as the calculation of how relevant a word in a series or corpus is to a text. The meaning increases proportionally to the number of times in the text a word appears but is compensated by the word frequency in the corpus (data-set).

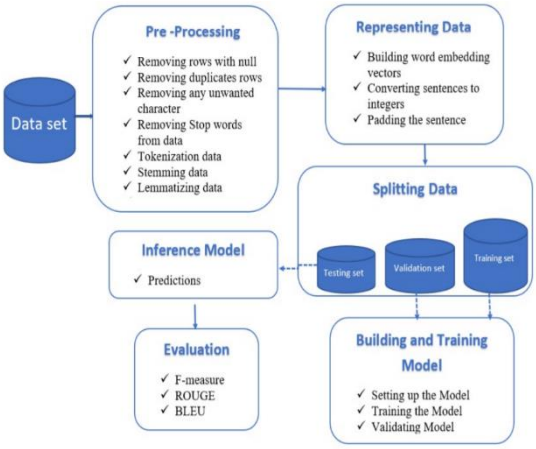
Term frequency is the number of instances of a term in a single document only; although the frequency of the document is the number of separate documents in which the term appears, it depends on the entire corpus. Now let’s look at the definition of the frequency of the inverse paper. The IDF of the word is the number of documents in the corpus separated by the frequency of the text.

III. Existing system

In natural language processing, Latent Dirichlet Allocation (LDA) is a generative statistical model that explains a set of observations through unobserved groups, and each group explains why some parts of the data are similar. The LDA is an example of a topic model. In this, observations (e.g., words) are collected into documents, and each word's presence is attributable to one of the document's topics. Each document will contain a small number of topics.

IV. Proposed system

Term frequency is the number of instances of a term in a single document only; although the frequency of the document is the number of separate documents in which the term appears, it depends on the entire corpus. Now let’s look at the definition of the frequency of the inverse paper. The IDF of the word is the number of documents in the corpus separated by the frequency of the text.



V. Working

1.TF stands for term frequency, which means that the formula will calculate the frequency of each word in a given document.

TF= Number of times a word occurs
Total number of words in a document

2.Let's say I want to know the frequency of the phrase (let's say w1) in document 1.(D1). TF(w1,D1). Let's count the instances of W1 in D1: that's 3. D1 contains six words in total. So, TF(w1, D1)=0.5.

3.A given word's frequency within the corpus will be measured by IDFs (Inverse document frequency).
IDF=-Log(N/n)
n=Number of occurrences of a word in documents
N=Total number of documents
IDF= Log(N/n)

Why Log?

This is a hack to normalize the Number after division; there is no fix theory behind it. I may be mistaken in this case (just a thought).
Let's imagine we have a corpus of 10,000 papers, and only 5 of those documents include a specific word.
IDF = (10,000) /5 = 2000 will result from applying the IDF formula without utilizing the log. (Remember without log)

If you look at this amount, it's a significant quantity. Keep in mind that there could be 1 million documents, therefore this number will be excessive.
Now, if we use log, log(2000) = 3.30 (roughly), a negligible value.

4.Both TF and IDF will calculate over a certain word, which means that w1 will be used in both TF and IDF.
Now, let's use a sample from our unique corpus. TFIDF of w1 in D1 in our bespoke corpus is what we're after.

VI. Advantages

- Enable Topic Modelling naturally
- Reduce Costs and Inefficiencies
- Benefit from Market Research and Analysis
- Ability to Deliver High Quality Results

- Automatically detects the important features Constantly evolves with new information

VII. Conclusion

We have offered visual techniques for making comparisons associated with the tasks of using topic models to understand topics similarity and change within text corpora. Among these techniques we have introduced buddy plots a novel visualization for viewing changes across two sets of distances by restricting our vantage point to that of a single reference document. Finally, we have described a number of use cases that exhibit the variety of comparisons that can be made with these methods.

VIII. References

1)X. Cheng, X. Yan, Y. Lan, and J. Guo, BTM: Topic modeling over short, Dec.

2)M. Chen, Machine learning for wireless networks with articial intelli- gence: A tutorial on neural networks, , arXiv:,Online

3)Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang, A hybrid deep learning based trafc ow prediction method and its understanding, Transp.

4) G. P. Zhang, Time series forecasting using a hybrid ARIMA and neural, J. Wang, Q. Gu, J. Wu, G. Liu, and Z. Xiong, Trafc speed prediction and congestion source exploration: A deep learning method, in Proc., B. Lu, F. Meng, Y. Zhao, X. Qi, B. Lu, K. Yang, and X. Yan A linear regression-based prediction method to trafc ow for low-power WAN with smart electric power allocations, in Proc.

5)th Int. Conf. Simulation Tools Techn. Cham, Switzerland: Springer, Jul p.

6)S. Khorsandroo, N. R. Md, and S. Khorsandroo, A generic quantitative relationship to assess interdependency of QoE and QoS, KSII Trans.

7)Y. Zhu, G. Zhang, and J. Qiu, Network trafc prediction based on Oct., . , Y. Gu, W. Lu, X. Xu, L. Qin, Z. Shao, and H. Zhang, An improved Bayesian combination model for short-term trafc prediction with deep Mar.

8)D. Zhang, L. Liu, C. Xie, B. Yang, and Q. Liu, Citywide cellular trafc no., p., Jan. J. Li.

9) Y. Zhang, W. Liang, T. Cui, and J. Li, A spatial information network trafc prediction method based on hybrid model, Int. J. Electron

10) B. C. Pijanowski, D. G. Brown, B. A. Shellito, and G. A. Manik, "Using neural networks and GIS to forecast land use changes: A land transforma-