ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
IJARCCE
nCORETech
LBS College of Engineering, Kasaragod
Vol. 5, Special Issue 1, Feburary 2016

A New Way of Topic Modeling Using MALLET for Current Job Trends

Athira M[1], Bhavya K[2], Soorya K[3], Ajeesh Ramanujan[4], Anoop V.S[5]

Student, Computer Science and Engineering, Govt. Engineering College, Palakkad, India[1,2,3]

Assistant Professor, Computer Science and Engineering, Govt. Engineering College, Palakkad, India[4]

Research Scholar, Data Engineering Lab, IIITMK, Trivandrum, India[5]

Abstract: Topic modeling is the process of extracting topics from texts. A **topic** can be viewed as a collection or cluster of words that occur together and frequently. Latent Dirichlet Allocation(LDA), a statistical topic model is used to extract topics from the collected corpus which is a collection of job related data from LinkedIn. LDA is an unsupervised machine learning approach. We analyzed the interrelationship between topics and represented it graphically. The recent job trends in the industry can be interpreted easily using this representation.

Keywords: Topics, Topic Modeling, MALLET, Latent Dirichlet Allocation(LDA), Gephi.

I. INTRODUCTION

We may come across situations where we have to analyze the composition of the documents that we are dealing with. The document may consist of topics of different kind. Here comes the importance of topic modeling. It is the process of extracting topics from texts containing data from different domain. Topic models provide a simple way to analyze large volumes of unlabeled text [1].

A. Topic Modeling

Topic modeling works on the idea that documents are mixtures of different topics. The distribution of each topic inside each document may differ. Topic modeling is a set of methods that analyze the words in the original documents to find the topics and how those topics are related to each other. A “topic” can be understood as a collection of words that have different probabilities of appearance in passages discussing the topic.

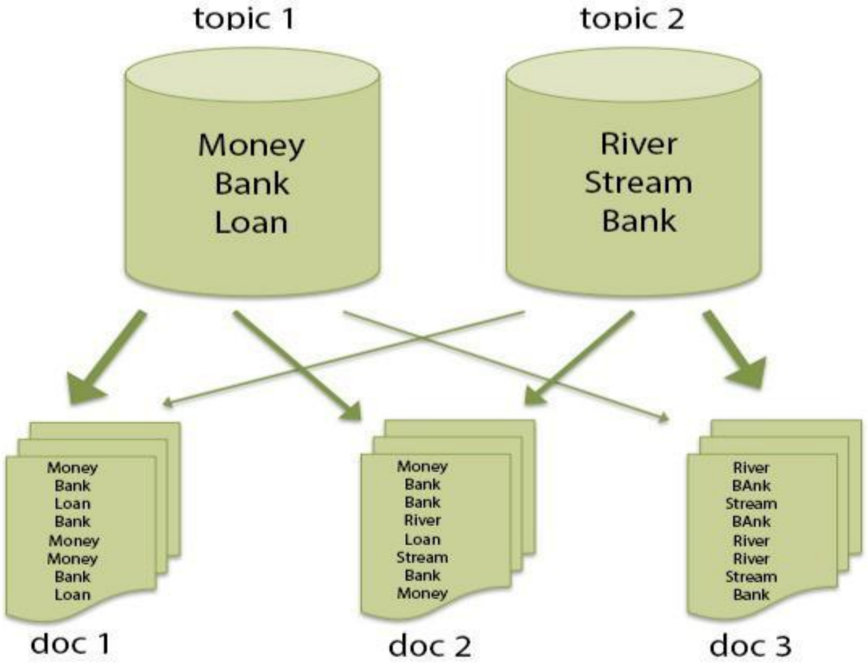


Fig 1. Illustration of topic modeling

DOI 10.17148/IJARCCE

Fig 1 illustrates the process of topic modeling. Consider three documents doc 1, doc 2 and doc 3, which are represented by frequently occurring words in them. It is clear from the figure that doc 1 comprises of words that are more related to the financial institution (bank) like money, bank, loan, etc. Whereas doc 3 contains words that are more related to river bank like river, stream, etc. doc 2 contains a mixture of words that are related to both financial institution and river. When these three documents are given as input for topic modeling, it extracts two topics. The topics obtained are related to financial institution and river. The directed edges shows the presence of the extracted topic in each of the document. The thickness of the edges represent the proportion with which the corresponding topic is present in the document(greater the thickness, higher the proportion).

The next two sections describe the related works and objective of our work. Section 4 discusses about our proposed system. The results are given in Section 5. Section 6 presents our conclusions and future works.

II. RELATED WORK

The paper by David M. Blei et al. [1] deals with Latent Dirichlet Allocation (LDA). The involvement of LDA in topic modeling is based on the property of exchangeability of both words and documents. It also deals with the relationship between LDA and other latent variable models such as unigram model, a mixture of unigrams, and the pLSI model. All four models, unigram, mixture of unigrams, pLSI, and LDA, operate in the space of distributions over words. LDA has conceptual advantages over other latent topic models because LDA posits that each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter [2]. The paper also compares the performance of LDA with other models. For this purpose they trained the models on C. Elegans community and a subset of the TREC AP corpus and they found that LDA outperforms other models. They also found that both pLSI model and the mixture of unigrams suffer from serious overfitting issues.

In [2] Jie Tang et al. proposed a system which integrates topic modeling approach into the random walk framework for academic search. The proposed model is called Author-Conference-Topic (ACT) model which can also represent the inter-dependencies among authors, papers and publication venues. The paper proposes three different types of possible implementations - ACT model 1, model 2 and model 3 and the possible combinations of these three models and random walks.

Fei Xia (2007) [3] discusses the different guidelines for using MALLET which is an integrated collection of Java code useful for statistical natural language processing, document classification, clustering, information extraction, and other machine learning applications to text. Compared to many other NLP packages, MALLET code is well written and well organized and hence is easy to use.

The paper by Jeffrey C. Reynar [4] deals with Statistical Models for Topic Segmentation and introduces new concepts that mark the point of shifts within topics a document have. It is assumed that most documents consist of more than one topic. By considering so, one can improve the efficiency of an IR task since many NLP and IR techniques implicitly assumes documents have just one topic. The paper describes novel algorithms for identifying topic boundaries where shifting occurs and the uses of the same once identified. The paper illustrates topic segmentation performance on several corpora and report improvement on an IR task that benefits from good segmentation. The process of segmentation is viewed as a labeling task and perform this labeling using statistical algorithms to determine the likelihood of a topic boundary.

III. OBJECTIVE

Our objective is to find out current job trends in IT sector by extracting available information from business-oriented social networking service, LinkedIn.

IV. OUR APPROACH

Our system goes through three phases.

A. Creation of Corpus

This is the initial phase of our work. Here the corpus is a collection of documents, where each document contains information about a particular job (job description) in IT industry. These job descriptions were extracted from LinkedIn using Crawler, implemented by regular expressions. The data of our interest is the job descriptions that comes under the Trending Searches link. The seed URL for the crawler program is the URL that points to IT related jobs from LinkedIn. Beginning with the seed URL, crawl through each link upto Trending Searches. Then each link under Trending Searches are crawled and the data is extracted recursively. This process is continued until the link Browse by Company is reached. As job trends varies from time to time due to whatever reasons, we can get the updated job descriptions by running the Crawler program on LinkedIn.

B. Topic Modeling

Once the corpus is created, now it is ready for topic modeling. We have used MALLET (MAchine Learning for LanguagE Toolkit) for topic modeling. MALLET works on the basis of LDA. Let the number of topics of our interest be **K**. The LDA algorithm works as follows :

Randomly assign each of the words in the document to any of the **K** topics which will give the initial topic representation of all the words and the documents.

- To improve this representation, for each word in each document repeat the following until optimal condition is reached.
- For each topic, compute the probabilities P(topic|document) and P(word|topic).
- Assign the word to the topic whose probability P(topic|document) * P(word|topic) is higher.

The documents created in section A are imported into MALLET which then gets converted into a MALLET processable form. Then the topic model routine runs on this MALLET representation to produce the possible topics. As it iterates through the routine, the best classification of words into topics are produced [6]. We will get a keys file representing top keywords for each topic, a composition file representing the breakdown, by percentage, of each topic within each original text file we imported and a state file that outputs every word in the corpus and the topic it belongs to, as the output. Next we have to represent the obtained results graphically which is given in the next section.

C. Gephi: A Graphical Representation

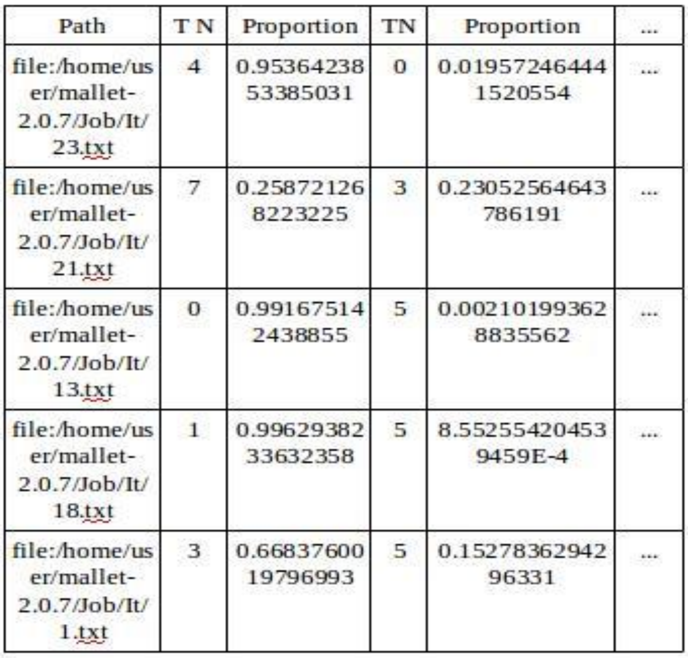
We have used the tool Gephi for representing our results graphically. Gephi is a program designed for exploring and visualizing all kinds of networks and complex systems, dynamic and hierarchical graphs [7].

In order to use our topic modeling result into Gephi, it has to be converted into XML format by identifying the nodes and edges in the graph. The nodes represent the particular words from the topics and edges denote the link between the topic and the words. The inter-relationship between topics can also be obtained by using Gephi.

V. RESULTS

Table I shows a portion of the composition of the input documents indicating the proportion of topics contained in them i.e., the output is a text file indicating the breakdown, by percentage, of each topic within each original text file imported.

TABLE I: COMPOSITION OF TOPICS IN THE CORPUS



Path	T N	Proportion	TN	Proportion	...
file:/home/user/mallet-2.0.7/Job/It/23.txt	4	0.9536423853385031	0	0.019572464441520554	...
file:/home/user/mallet-2.0.7/Job/It/21.txt	7	0.258721268223225	3	0.23052564643786191	...
file:/home/user/mallet-2.0.7/Job/It/13.txt	0	0.991675142438855	5	0.002101993628835562	...
file:/home/user/mallet-2.0.7/Job/It/18.txt	1	0.9962938233632358	5	8.552554204539459E-4	...
file:/home/user/mallet-2.0.7/Job/It/1.txt	3	0.6683760019796993	5	0.1527836294296331	...

The first column shows the path of the input documents. All the following columns shows the topic number (TN) and its corresponding proportion in that document. It is clear from Table I that the first text file (23.txt) has topic 4 (“management”) as its principal topic, at about 95%, topic 0 (“sales”) at about 1% and so on. The text file (21.txt) contains topic 7 (“Budgeting”) at about 25%, topic 3 (“Information-Security”) at about 23%. Similarly for each text document we obtained the proportion of each of the 10 topics described in Table II.

TABLE II : KEYS AND THEIR PROPORTIONS

Topic No. (TN)	Proportion	Keys
0	0.11006	strong experience track record candidate trading sales description high history firm account plan commission team recruiting work seeking geneva
1	0.05305	audit job accounting exciting processes opportunities finance information work opportunity authorized openings applying applicants apply provide world half robert
2	0.06757	security information controls risk microsoft job data standards firewall pci group travel develop policy compliance requirements control skills atum
3	0.02686	project work management brewery beer deschutes play strong identify system erp bend implementation effectively managing stakeholders goals define friendly
4	0.08621	services experience jde amp skills application business years consultants consulting oracle solutions areas developer functional full erp seeking polaris
5	0.15285	experience management support systems network technology knowledge administration operations skills description networks provide responsible including configuration degree recovery position
6	0.03299	prepares status orders purchase filing administrative reports departments asset mail calendars compiling expense payment invoices requisitions notes transcribes memos
7	0.12413	support development manage work routeone planning internal product customer team required project business skills software ability external activities delivery
8	0.03371	oversee system bal computer access management information disaster related business questionnaire validation lead company laboratory lims desk department set
9	0.02797	college university nursing technology information job wayne learning educational staff support state technologies distance academic department faculty health equal

Table II shows the obtained keys that characterize the different topics present in the input document. By analyzing Table II, we can derive the topics as Sales, Finance, Information Security, Management, Consulting, Network Administration, Budget, Customer Support, Research and Education respectively. The keys that determine the topics were represented graphically using the tool Gephi. Using this graph, we were able to find the inter-relationship between different topics. The output of Gephi for first topic is shown in Fig 3. We can see that the linked words are highly related to the derived topic 'Sales' (an area under IT). Similarly, we created graphs for all the 10 topics. Then we combined all these graphs to form the larger graph that illustrates the association between the topics which is shown in Fig 2.

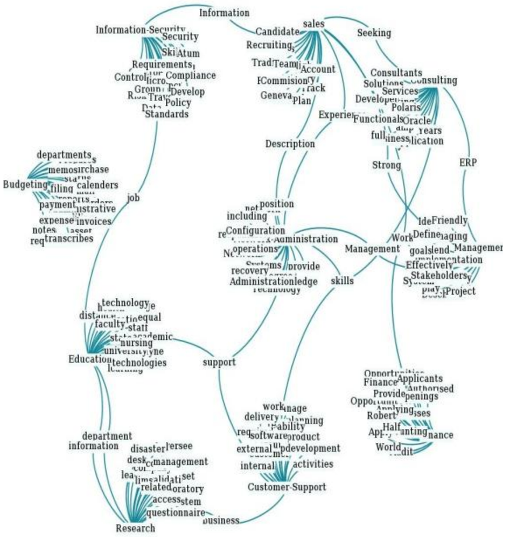


Fig 2. Inter-relationship between topics

DOI 10.17148/IJARCCE

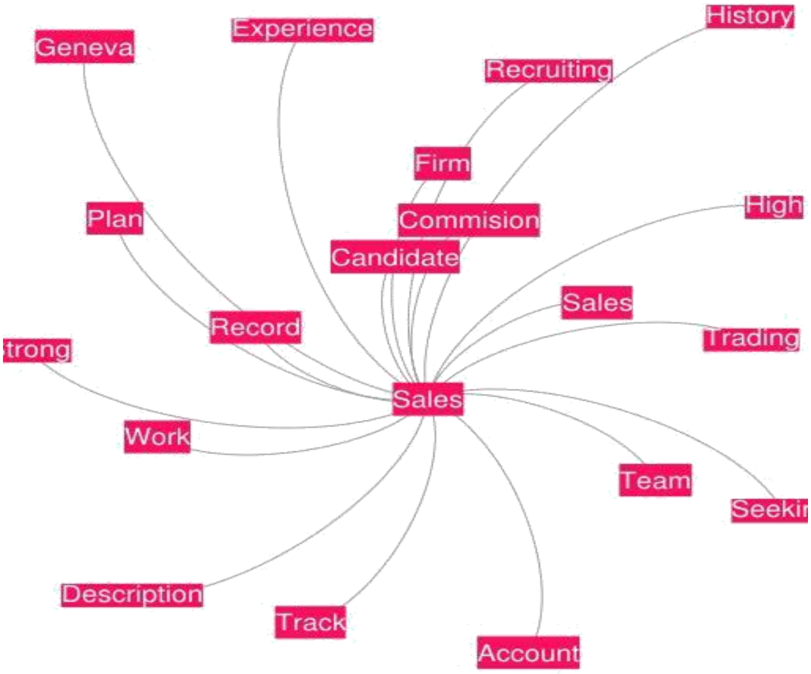


Fig 3. Graphical representation of first topic('Sales')

Fig 2 covers all the ten topics of which Fig 3 is also a part. The words or the qualities that link different topics were analyzed from Fig 2. For example, the area 'Budgeting' is not directly related to any of the other areas. For the remaining areas the inter-relationships are represented by their connected edges, like 'Education' and 'Research' are connected by the links 'information' and 'Department'. The frequent words that characterize the topic 'sales' are clear from Fig 3. By analyzing the graph, we can infer the current trend towards sales department. For eg. a person willing to work in such a department should be strong, have the ability to work in team, should be experienced and so on.

CONCLUSION AND FUTURE WORK

Topic Modeling has greater importance in this era of electronically generated documents. In this paper, we have presented topic modeling for LinkedIn data using MALLET. Using the information about IT related jobs extracted from LinkedIn, we found out the keys that represent the topics in the given input using MALLET. By analyzing the keys, we found out the topics to which the keys are related to. The results shows that some keys occur in more than one topic which indicates that they are related. This relationship is clearly shown in the graphical representation of the topics using the tool Gephi. Better results can be obtained by doing some preprocessing operations like removal of citations, any foreign-languages within the document, metadata and so on from the corpus. This work can also be extended to many domain to find out the possible topics and their relationship.

ACKNOWLEDGMENT

We are highly thankful to TEQIP Phase II for providing us financial support.

REFERENCES

[1]. David M. Blei, Andrew Y. Ng and Michael I. Jordan, Latent Dirichlet Allocation, in Journal of Machine Learning Research 3, 2003, pp.993-1022.

[2]. Jie Tang, Ruoming Jin and Jing Zhang, A topic modeling approach and its integration into the random walk framework for academia research, Data Mining, ICDM '08. Eighth IEEE International Conference, 2008, pp.1055-1060.

[3]. Jeffrey C. Reynar, Statistical models for topic segmentation, work done as a part of Ph.D. thesis work at the University of Pennsylvania.

[4]. Fei Xia, Mallet guides for LING 572, 2007.

[5]. MAchine Learning for LanguagE Toolkit [online]. Available: http://mallet.cs.umass.edu/. [Accessed]:02/09/2015Getting Started with Topic Modeling and MALLET [online]. Available: http://programminghistorian.org/lessons/topic-modeling-and-mallet. [Accessed]:08/09/2015

[6]. Topic Modeling and Gephi: A work in progress [online]. Available: http://dig-eh.org/topic-modeling-and-gephi-a-work-in-progress/. [Accessed]:10/11/2015