।। विद्या परम् दैवतम्।।
Wainganga Bahu-Uddeshiya Vikas Sanstha's
WAINGANGA COLLEGE OF ENGINEERING & MANAGEMENT
Near Gumgaon Railway Station, Dongargaon, Wardha Road, Nagpur - 441 114. (M.S.) INDIA
Tel.:07103 - 202007, 203728 Email:wcem@rediffmail.com, wcem4145@gmail.com, Website-www.wcem.in
An ISO 9001 : 2008 Certified Institute, Accreditated by International Accreditation Organisation
Approved by AICTE, DTE, Govt. of Maharashtra & Affiliated to RTM Nagpur University, Nagpur.

Academic Year: 2018-19

Sr. No	Title of paper	Name of the author/s	Department of the teacher	Name of journal	Year of publication	ISSN number	Page No.
1	REVIEW OF ANALYSING AND VISUALIZATION OF WEB OPINION MINING FOR SOCIAL INTERACTION	RAHUL BHANDEKAR	CSE	INTERNATIONAL JOURNAL OF RESAERCH AND ANALYTICAL REVIEWS	2018	2349-5138	79
2	REVIEW ON FACE RECOGNITION AND CLASSIFICATION USING LDP	CHANCHALA TRIPATHI	CSE	INTERNATIONAL JOURNAL OF RESAERCH AND ANALYTICAL REVIEWS	01-Feb-19	2349-5138	80
3	SURVEY ON TERM WEIGHTING USING COHERENT CLUSTERING IN TOPIC MODELING	Manisha N. Amnerkar	CSE	INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY	2019	2349-6002	81
4	TERM WEIGHTING USING COHERENT CLUSTERING IN TOPIC MODELING	Manisha N. Amnerkar	CSE	INTERNATIONAL JOURNAL OF RESAERCH AND ANALYTICAL REVIEWS	2019	2349-5138	82

Wainganga College of Engineering & Management * Dongargaon * Nagpur *
Seal

Principal
Wainganga College of Engineering
& Management, Dongargaon, Nagpur

Review on Analyzing and Visualization of Web Opinion Mining For Social Interaction

1st Rahul Bhandekar
Assistant professor
Dept. of computer science and engineering
Wainganga college of engineering and management
Nagpur, India

2nd Monali Dhale
Dept. of Information Technology
Wainganga college of engineering and management
Nagpur, India

3rd Rohit Pohankar
Dept. of Information Technology
Wainganga college of engineering and management
Nagpur, India

4th Rohit Vishwakarma
Dept. of Information Technology
Wainganga college of engineering and management
Nagpur, India

Abstract-*Due to the improvement of technology, a large number of web opinions are available on the social media such as web forum and weblogs. This platform provide the communication between numbers of users and express their opinion on the particular topic such as social, political, sport, technical, entertainment etc. In our propose approach the web opinion clustering technique enable the identification of themes within discussion in web social network and there development, as well as the interaction of active participants. While discussing any topic, each and every participant can share his/her view on that topic and all the suggestion will accept. And on the sis of their opinion we will give the result by using grading system.*

Keywords: *Data mining, Clustering.*

I.INTODUCTION

Data mining**:** Data Mining is defined as extracting information from big sets of data. In other words, we can say that data mining is the strategy of mining knowledge fromss data. The information or knowledge extracted so can be used for any applications Clustering**:** Making a group of abstract objects into classes of similar objects is called clustering Clustering can be divided into two subgroup:

Hard Clustering: In hard clustering each data point put into single group out of the 10 groups. Soft Clustering: In soft clustering substitute of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario per customer is assigned a probability to be in either of 10 clusters of the retail store. Requirements of Clustering in Data Mining: The following sentences throw light on why clustering is required in data mining. Scalability- we need highly scalable clustering algorithms to deal with big databases. Ability to deal with different kinds of attributes − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary information. Discovery of clusters with attribute shape − the clustering algorithm should be capable of detecting clusters of arbitrary structure. They should not be restricted to only distance measures that tend to find spherical cluster of little sizes. Giant dimensionality − the clustering algorithm should not only be able to treat low-dimensional data but also the giant dimensional space. Capability to deal with noisy data − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to bad quality clusters. Interpretability − the clustering results should be interpretable, comprehensible, and utilizable.

© 2019 IJRAR February 2019, Volume 6, Issue 1 www.ijrar.org (i -1269, P- ISSN 2349-5138)

Review on Face Recognition and cl: ation using LDA

1ChanchlaTripathi, 2 Harshali Ragite, 3 Shekhar Ranjan, 4 Rajshree Raut 5. Premlata Barewar
1 Assistant Professor, 2,3,4,5 B.E Student,
1, 2, 3, 4,5Department of Computer Science & Engineering
1, 2, 3, 4,5Wainganga Collage of Engineering & Management, Nagpur, India

Abstract-As we know that handling the large dataset is most critical problem. And for the handling of image our problem increases twice. So here we are going to propose a system in which we generate the pattern of the image according to their properties and features. And that properties of the image by which generate the pattern that's help us to classify the image from the large dataset of the image. For the classification we compare the image properties to the properties of the image which is stored in the dataset. The properties are going to match maximum properties of the dataset or above the minimum limit of the properties, we classify that sets of the image as the output of the system. In this we also add the functionality to classify the image according to the emotion of the image. For the classification here we are using the Class-Specific-Simplex Latent Dirichlet Allocation Model (CSS LDA). It is a topic model of the classification.

Keywords: Face Recognition, classification

Introduction- As we know our basic problem is to handle the large dataset. And that large dataset is of image. For this we first process the image then after doing the classification process. This dataset contain the large amount of mathematical calculation. And that comes under the machine learning. Here we trend our system by giving the number of possibilities. That makes our system smarter and make our manual work easy.

We are going to classify the large dataset of the image in which firstly we have to categories the properties of the image, for taking care of the properties of image, we have apply the image processing method that are the completely depend upon the mathematical calculation. For this we have to extract all the feature of the image like the input of the image, conversion of the color image into the gray scale and so many properties. This is basically work on the face of the human image. Properties of the face is consider as the key to classify the image from that dataset.

Coming to the point of classification there is number of ways by which we classify the image but here we are considering the topic model of classification. In which we are categories the property according to the topic and the value of that topic. As we know that this is full of mathematical calculation that's why we prefer the topic model. Topic model of classification is a form of unsupervised learning (akin to clustering), so the set of possible topics are unknown apriori. They're defined as part of generating the topic models. With a non-deterministic algorithm like LDA, we'll get different topics each time as we run the algorithm.

For the completion, first we have to generate the pattern by the sets of image and when we have to classify the image, the recognition of the pattern from the image dataset.

Literature Review:

There is lots of process by which we can generate and classify the image. Here we are focusing on the time complexity and accuracy of the classifying the image.

For the feature extraction firstly we are discuss about the input of the image or which type of the image we are going to classify from the dataset. As we know that there is different types of image of with different view some are of the image with different facial expression like the sad, happy, etc. some of the view of the different facial motion i.e multiview of the face. In simple language we can say that some image of the face is in left side some of the in the right and so on. Also we are working on the image who is surgically altered face, it cannot be easy to find the accurate but we are trying to classify that image also. This is all about, which type of image we are going in surveillance. Apart from this we have to generate

© June 2019 | IJIRT | Volume 6 Issue 1 | ISSN: 2349-6002

Survey on Term Weighting Using Coherent Clustering in Topic Modeling

Manisha N. Amnerkar[1], Ashwini Tikle[2]
[1]*M-Tech Student, Department of Computer Science & Engineering, Wainganga Collage of Engineering & Management, Nagpur, India*
[1]*Assistant Professor, Department of Computer Science & Engineering, Wainganga Collage of Engineering & Management, Nagpur, India*

Abstract- **Topic models often produce uncountable topics that are filled with noisy words. The reason is that words in topic modelling have same weights. More frequency words dominate the top topic word lists, but most of them are meaningless words, e.g., domain-specific stopwords. To address this issue, in this paper we aim to investigate how to weight words, and then develop a straightforward but effective term weighting scheme, namely entropy weighting (EW). The proposed EW scheme is based on conditional entropy measured by word co-occurrences. Compared with existing term weighting schemes, the highlight of EW is that it can automatically reward informative words. For more robust word weight, we further suggest a integrated form of EW (CEW) with two existing weighting schemes. Basically, our CEW assigns unmeaning words lower weights and informative words higher weights, leading to more coherent topics during topic modelling inference. We apply CEW to DMM and LDA, and evaluate it by topic quality, document clustering and classification tasks on 8 real world data sets. Exploratory results show that weighting words can effectively improve the topic modelling performance over both short texts and normal long texts. More importantly, the proposed CEW significantly outperforms the existing term weighting schemes, since it further considers which words are informative.**

Index Terms- Topic modeling, Term weighting, Informative word, Conditional entropy

1. INTRODUCTION

With the rapid increase in the amount of electronic information on internet web pages and modern applications, text analysis in the domain of text mining requires complex techniques to deal with numerous text documents. Topic models are widely used to uncover the latent semantic structure from text corpus. The effort of mining the semantic structure in a text collection can be dated from latent semantic analysis (LSA) [17], which employs the singular value decomposition to project documents into a lower dimensional space, called latent semantic space. Probabilistic latent semantic analysis (PLSA) [6] improves LSA with a sound probabilistic model based on a mixture decomposition derived from a latent class model. In PLSA, a document is represented as a mixture of topics, while a topic is a probability distribution over words. Extending PLSA, Latent Dirichlet Allocation (LDA)[7] adds Dirichlet priors for the document-specific topic mixtures, making it possible to generate unseen documents. Due to its nice generalization ability and extensibility, LDA has achieved huge success in text mining. In the last decade, topic models have been extensively studied. Many complicated variants and extensions of the standard LDA model have been proposed, which can be found in the comprehensive survey [18]. Here we only list some work closely related to us. Wallach [19] proposed the bigram topic model extending LDA by incorporating bigram statistics into topic modeling, but its detail is quite different from ours. The bigram topic model aims to capture ordinal dependencies between words (in normal texts) by exploiting document-level sequential patterns, while our model is designed specifically for short texts and tries to capture the semantic dependencies between words by exploiting corpus-level word co-occurrence patterns. Besides, two recently proposed models, i.e., the regularized topic model [20] and the generalized P_olya model [21], share the same idea of utilizing word co-occurrence (i.e., biterm) statistics to enhance topic learning. However, both of them only use word co-

IJARSCT

ISSN (Online) 2581-9429
ISSN (Print) ABCD-PQRS

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 9, Issue 5, September 2020

Term Weighting Using Coherent Clustering In Topic Modeling

Manisha N. Amnerkar

Masters of Technology

Computer Science & Engineering

Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur

Abstract: *Term weighting is known as a text presentation strategy to assign appropriate value to each term to improve the performance of text classification in the task of transforming the content of textual document into a vector in the term space. The reason is that words in topic modelling have equal weights. High frequency words dominate the top topic word lists, but most of them are meaningless words, e.g., domain-specific stop words. To address this issue, in this paper we aim to investigate how to weight words, and then develop a straightforward but effective term weighting scheme, namely entropy weighting (EW). The proposed EW scheme is based on conditional entropy measured by word co-occurrences. In this paper, a new weighting scheme is proposed via a matching score function based on a probabilistic model. We introduce a latent variable to indicate whether a term contains text classification information or not, specify conjugate priors and exploit the conjugacy by integrating out the latent indicator and the parameters. Then the non-discriminating terms can be assigned weights close to 0. Experimental results using KNN and SVM classifiers illustrate the effectiveness of the proposed approach on both small and large text data sets.*

Keywords: Machine learning, KNN & SVM classifiers, Data mining.

I. INTRODUCTION

Term weighting is a procedure that takes place during the text indexing process in order to assess the value of each term to the document. Term weighting is the assignment of numerical values to terms that represent their importance in a document in order to improve retrieval effectiveness.

Topic models often produce unexplainable topics that are filled with noisy words. The reason is that words in topic modelling have equal weights. High frequency words dominate the top topic word lists, but most of them are meaningless words, e.g., domain-specific stop words. To address this issue, in this paper we aim to investigate how to weight words, and then develop a straight forward but effective term weighting scheme, namely entropy weighting (EW).

Text classification (TC) is the task of automatically classifying unlabeled electronic documents, such as news articles, advertisements, e-mails, call records and so forth, into a predefined set of classes. It is a supervised