**APPLICATION OF TEXT MINING FOR PREDICTING TITLE OF A DOCUMENT**



***by***

**Majid Khan**

A research submitted in partial fulfillment of the requirement for the  
degree of Master of Science in  
Computer Science

**Advisor: Mr. Imran Ali**

**Co-Advisor: Dr. Maheen Bakhtyar**

**Department of Computer Science  
University of Balochistan  
April 2019**

**Acknowledgment**

To one true Allah - To Him alone I owe all praise and glory.

I am grateful to my Supervisor, Mr. Imran Ali, Lecturer, Department of Computer Science, University of Balochistan for his valuable suggestion, support and ideas for this research study work. He arranged a meeting with me in every week, his precious advice and efforts helped me during this research study.

I am also a very grateful to my co-supervisor Dr. Maheen Bakhytar, Assistant Professor, University of Balochistan. She also made a very friendly atmosphere for asking a question or any kind of problem with her which I faced and overcome during the research study. She was very polite, kind and active for providing a useful guidelines at the time of need and made me able to finish this research synopsis.

I am grateful to my Computer Science Department’s Teachers and other faculty  
members for their valuable support and their precious guidelines.

**Abstract**

A huge quantity of textual data, documents and newspaper articles are generated on Internet from different publishing sites, blogs and news-media, etc. This leads to one of the major tasks in Text Mining i.e. effectively managing, searching and categorizing documents depending upon their Title themes. Typically, these text mining tasks will provide a platform for predicting useful Title name of any articles, documents and other textual form of data.

In this research proposal we are trying to find out new ways or methods for predicting a suitable Title topic name of any articles. To achieve this objective we are using mapping technique, wherein first, we will extract useful statistical feature from documents and then based on this statistical feature we will try to find probability of similarity between each sentence of documents to its corresponding title sentence using well known **cosine similarity algorithm.** After the completion of training phase our model will be able to apply on testing data set for prediction title of any unstructured documents, which is known as “Topic Modelling".

**Keyword:** Predicting title using cosine similarity, Mapping technique, Statistical based feature extraction

Table of Contents

|  |  |  |
| --- | --- | --- |
| Chapter | Title | Page |
|  | Title Page | i |
|  | Acknowledgement | ii |
|  | Abstract | iii |
|  | Table of Contents | iv |
|  | List of Figures | V |
|  |  |  |
| **1** | Introduction | 1 |
|  | 1.1 Background | 2 |
|  | 1.2 Problem Statement | 3 |
|  | 1.3 Objectives | 4 |
|  | 1.4 Significance of the study | 4 |
|  | 1.5 Limitation and Scope | 4 |
|  |  |  |
| **2** | Literature Review | 5 |
|  |  |  |
| **3** | Proposed Methodology | 8 |
|  | 3.1 Workflow | 8 |
|  | 3.2 Data Acquisition | 9 |
|  | 3.3 Data Preprocessing | 9 |
|  | 3.3 Mapping sentence using Cosine Similarity Measure | 9 |
|  | 3.4 Statistical Based Feature Extraction (SBFE) | 9 |
|  | 3.5 Result Formulation and Analysis | 10 |
|  | References | 11 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Figure** | **Title** | **Page** |
| 3.1 | Statistical based feature extraction from text document | 10 |

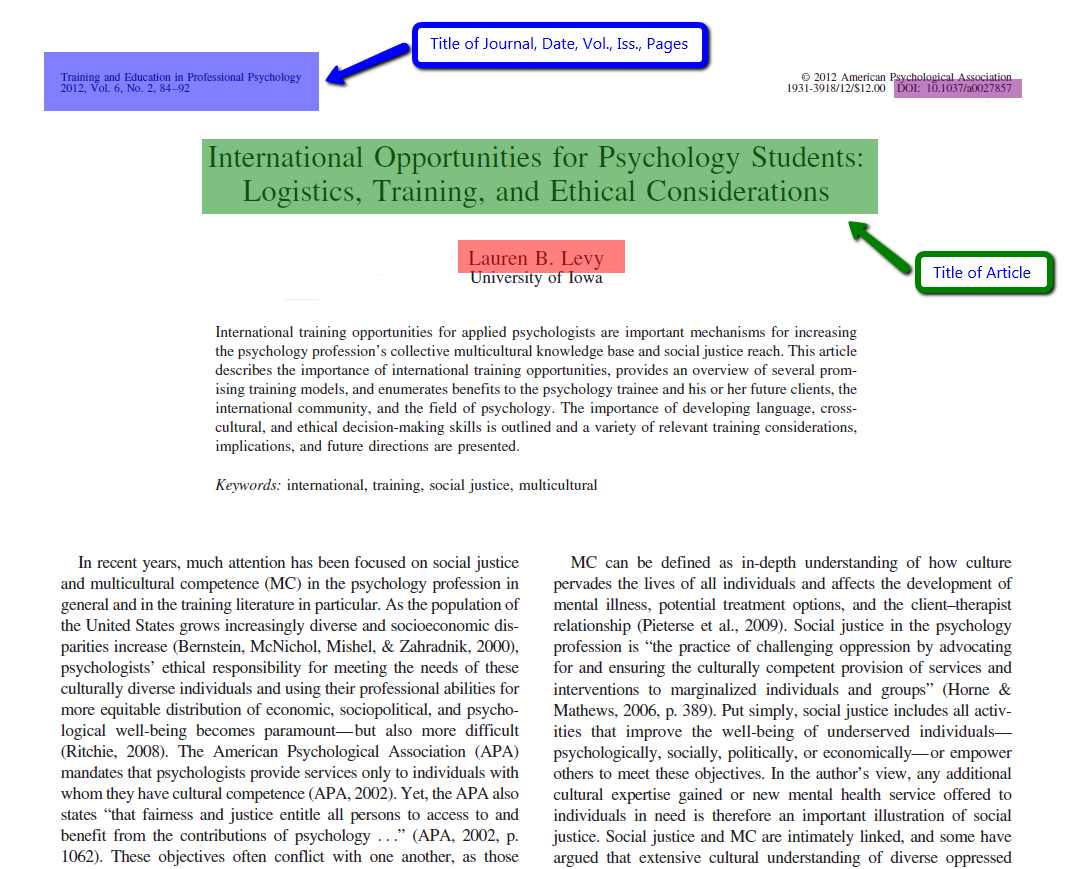
**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure** | **Title** | **Page** |
| 1.1 | Illustrate Importance of Title in a document | 1 |
| 1.2 | Analyzing Text Data | 2 |
| 1.3 | Unstructured Text Document | 3 |
| 1.4 | Objective to construct predictor to predict new case | 4 |
| 3.1 | Workflow of Proposed Model | 8 |

**Chapter 1**

**Introduction**

Choosing a good title for an article is an important step in the writing process. A  
good title catch the reader attention and offers readers to read the whole content of  
the paper and to have become more curious to find out more about it that what you  
have written, how you developed your contents and also showing wisdom, knowledge  
and writing skill of the writer in the article . The qualities of good titles contain  
fewest and relevant words but adequately describe the whole content of the article.



**Figure 1.1:** Illustrate Importance of Title in a document

Prior choosing a good title name is the only expertise of human being who has  
equipped with knowledge, wisdom and command on language which make easy  
for him to choose good title name, but now with the advent of computer and  
tremendous research work in the field of Text Mining make this problem easy to  
some extent toward choosing automatically title name without the intervention of  
human being. Lots of examples are there which shows the important of title like  
News agencies, News wires generate from news story/article, Breaking News ticker  
etc are of the few example which pave the way for further works.

**1.1 Background**

We are all surrounded by high tech gadgets that generate and consume data at enormous rate. It would not be wrong to say that we are drenched in data. Although  
advancement in technology has brought ease and comfort in our life, it has not come  
without cost.

Two of the most difficult to find assets nowadays are spare time and uninterrupted  
focus or concentration towards a specific task. Creative tasks like writing and reading are no exception and require complete attention.

Scarcity of time and focus that the essence of information to be conveyed to the  
reader is provided in relatively fewer words and in the bearing and in the beginning  
of an article. Title of an article serve this purpose.

The title of an article is thus is very important, and comprising an informative,  
descriptive and concise, title is crucial for conveying information in least of time and  
words.

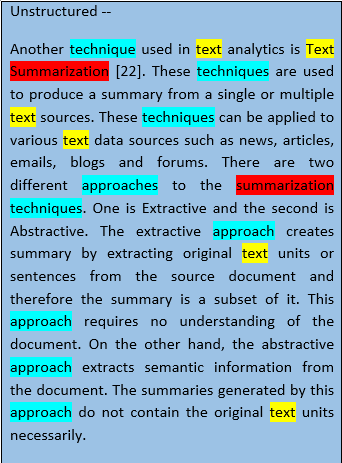


**Figure 1.2:** Analyzing Text Data

**1.2 Problem Statement**

The main focus of this research study is to design a model which can discover the  
mapping between the content of articles and its corresponding title by learning from  
existing examples (i.e pair of articles corresponding titles).

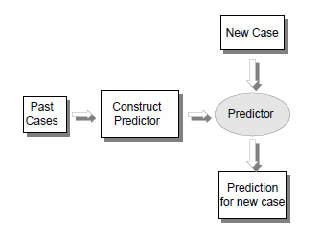
Automatically choosing or predicting a relevant good title from unstructured text of  
an article/document is the main problem of this research study. As far as concern  
for human being using or predicting title of any document is a not big deal or problem whereas for computer to predict automatically relevant and good title of any  
articles or documents involved such a complicated task and mechanism where lots  
of Text Mining feature extraction techniques required. For solving and ascertaining  
this problem answer how to predict relevant title of an articles will discuss partially  
in next chapter of methodology.



**Figure 1.3:** Unstructured Text Document

**1.3 Objectives**

The main objective of this research study is that learning the relationship between  
article and its title and develop a model that can predict title of any article not yet  
seen in training examples.



**Figure 1.4:** Objective to construct predictor to predict new case

**1.4 Significance of the study**

The proposed approach intends to provide systematic way that exhibits good predictive accuracy in even challenging and unseen documents. While extracting the  
feature of documents for mapping there is chance that some irrelevant feature may  
also be extracted which need state-of-the-art algorithm to reduce this problem and  
this approach will lead toward developing such model.

**1.5 Limitation and Scope**

This research study involves data set of a single language i.e English. While prediction of title of any documents involves different kind of language but in this research study we focus only on English unstructured data set. The scope and future of this research work will focus on multi-language domain where to develop such a model which can predict any kind of unstructured document for any language.

**Chapter 2**

**Literature Review**

Huge work has been carried out so far in the field of text mining and discovered  
new methods and techniques for various problems. Toward prediction of title of any  
unstructured document is a very active research problem of these days especially in the field of text mining, lot of work has been done for keyword extraction but to the best of my knowledge there is minimal work done so far for prediction of title. I have not yet fond any research work for predicting title. So therefore, few closest related research study work has been citing here which for the purpose of further enhancement and elaboration of this research work.

(**Chen, Wang, Li, & Sun**,2018), represented EmbedRank and EmbedRank++ methods, the two simple and scale able method for keyphrase extraction from a single documents. Both method are unsupervised, corpus independent and only require current document instead rather than the entire corpus. Both method depart traditional graph base method for keyphrase extraction.

EmbedRank can be implement on top of underline document embeding, these embedding can encode document of arbitrarily length. Finally, these methods used with Doc2Vec and Sent2Vec, EmbedRank based on sen2vec is much faster as compared with doc2vec.

(**Bennani-Smires, Musat, Hossmann, Baeriswyl, & Jaggi**,2018), proposed a correlated neural influence model to study the intrinsic mechanism and predict trending research topic, measuring the scientific influence to allocate and efforts and resources in science. It has the ability to capture the sequential properties of the research evolution and correlate with different conference under the cross conference influence. Demonstrate the effectiveness of the proposed model by conducting experiments on a scientific data and proposed model shows the best result as compared with the other existing method.

(**Zhai et al.**,2018), the proposed method used bilingual text data from Chinese and Vietnamese events where word used as vertexes form and similar semantic word as hyper edges and then calculates the weights of the hyper edges which showed high importance in the sentence.

Experimental results shows the effectiveness of the method. Multivariate relations  
between two language have a good and supportive role for the keyword extraction.

(**Das, Pal, Mondal, Dalui, & Shome**,2013), Fuzzy based algorithm used to extract keyword from any text. The strength of this algorithm does not requires corpus  
nor depend on the size of the document. This algorithm is very effective and  
useful for keyword from a document. Fuzzy algorithm is very useful in numerous  
application where corpus is not available.

(**Dostal & Je\_zek**,2011), The proposed method for extraction of keyphrase based on statistical and word-net based on pattern was derived from combination of graph method (TextRank) and statistical method (TFIDF). Keyword candidate are merged with named entity and the stop word to NL POS tagged pattern. Automatic keyphrase are generated as weighted TF-IDF unigram.

The proposed method are efficient as compared to with other automatic keyword  
extraction system. For example, the RAKE system achieved 33.7 percent precision  
and 41.5 percent recall and the un-directed TextRank achieved 31.2 percent precision and 43.1 percent recall whereas in this paper achieved 37.4 percent precision  
and 54.6 percent recall for a smaller corpus with expanded number of annotations.  
The limitation of this paper is assume that precision and recall will be a little lower  
for a bigger corpus.

(**Rose, Engel, Cramer, & Cowley**,2010), The proposed method is used RAKE (Rapid Automatic Keyword Extraction) from individual document. RAKE take a simple set of input parameter and automatically extract keyword in a single pass RAKE achieve higher precision with similar recall in comparison to existing techniques. RAKE its simplicity and efficiency enable its use in many application where keyword can be leveraged.

(**Bosni\_c, Verbert, & Duval**,2010), describe a use case for an application that recommended learning object for reuse and integrated in the authoring environment.

(**Z. Liu, Huang, Zheng, & Sun**,2010), the proposed method is graph based  
Topical PangeRank. This method is very useful for keyword extraction. Experiment  
results taken from two datasets proves that TPR method achieved better result as  
compare to other base line method, besides this TPR also indicates its effectiveness  
and roubstness.

(**F. Liu, Liu, & Liu**,2008), investigated the problem of automatic keyword extraction in the meeting domain. Adopted supervised framework and leveraged feature extracted from meeting specific characteristic such as decision making sentence and system generated summaries. Feature selection and different candidate word resampling technique prove to be helpful in the supervised method. In addition, this paper introduced bigram expansion module which leverage both web resources and confidence scores from the classifier. Experiment result demonstrate that both supervised approach and the bigram module improved keyword extraction performance.

(**Meyerzon, Cao, Li, Zheng, & Hu**,2005), proposed method for title extraction from general document is based on machine learning approach where  
general documents indicate any kind of documents from any domain which includes power point presentations, research papers, letters, reports, book chapters  
and brochures. The techniques applied on this research method is very unique in a  
sense that used formatting information of text font and size as a features.  
In this proposed method the experiment results depict precision and recall for Word  
are 0.810 and 0.837 respectively, and for PowerPoint text are 0.875 and 0.895, respectively. Other features of this models are that domain independent and language  
independent this is the specific usefulness of this model.

**Chapter 3**

**Proposed Methodology**

**3.1 Work Flow**

In this chapter, I intend to propose methodology workflow using a very different approach as compared to its previous cited research work in this study for predicting a title theme from unstructured text data set applying statistical feature of the data to develop a probability based model. Based on idea of statistical feature base model we are developing such technique wherein introducing mapping techniques to map each word of title to its corresponding each sentence of a document using cosine similarity measure for extracting higher similarity feature to find a title. The proposed model can be summarized into following steps describe in Figure 3.1:

**Data Acquisition**

**Data Preprocessing**

**Mapping sentences using Cosine Similarity Similarity**

**Statistical Based Feature Extraction**

**Analysis**

**Test data**

**Output**

**Result Formulation**

**Figure 3.1:** Workflow of Proposed Model

**3.2 Data Acquisition**

The first phase of proposed method is data acquisition here we choose bytecup2018 Chinese oriented website free competition of machine learning problems data sets of approximately 3000 unstructured JSON file (Java Son Oriented Annotation) documents.

**3.3 Data Prepossessing**

The Second phase of this research methodology are data cleaning. After taking data from above mentioned website, then it need to be clean and make readable form, remove all the irrelevant character, punctuation, stop word, lemmatization etc and then convert JSON file into vectorize form i.e ID, CONTENT AND TITLE for further processing and extracting some useful statistical feature.

**3.4 Mapping Sentence using Cosine Similarity Measure**

In this phase we are describing Mapping Techniques between (known input and known output) that is title to its corresponding document sentence for the purpose to find higher similar position of a sentence in a document for predicting title suing **Cosine Similarity Measure**. Cosine is a similarity metric used to measure similarity between documents and measure angle between two vectors or sentence in a multi-dimensional space. Cosine similarity even if two documents are far apart by the Euclidean distance it will close together. Higher the angle smaller the cosine similarity and smaller the angle, higher the cosine similarity.

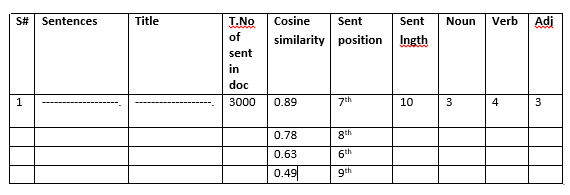
cos(a,b) = a\*b

||a|\*|b||

As we already mentioned in previous phase that all the unstructured data set of text placed into vector form for using only two columns of data set i.e Title column and Contents Column for matching purpose. After the successful completion of mapping and cosine measure technique, now we will proceed the next and very important phase of this proposed research study work i.e Extraction of Statistical Based Feature (ESBF).

**3.5 Statistical Based Feature Extracting (SBFE)**

After preprocessing all the data set into vectorize form, next phase is being i.e feature extraction based statistical. The third and very most important phase of this proposed methodology is extracting some hidden features from a text documents for further mapping and finding the cosine similarities. The statistical feature which we will extract from a document are of title total number of sentence in a document, cosine similarity value, matched sentence position, sentence length value, total number of (noun, pronoun and verb) in a sentence using cosine similarity measure. To find higher similarity between two sentences either can lead to predict a title or be capable our model to from new acquired information for new instances.

****

**Table 3.1:** Statistical based feature extraction from text document

**3.6 Result Formulation and Analysis**

Finally after the finding the cosine similarity between title and sentences, and extracting some useful statistical feature the last phase will commence that is evaluation of result and Analysis where we can be able to predict most relevant sentences as a title or near to the title. At last but not least, this evaluation process will be conducted through survey where documents against its predicted titles list of each document will be presented to expert of English language for choosing title for further measurement of proposed model accuracy.

To predict title from unstructured text data based on mapping techniques using SBFE, where set of labeled documents and its corresponding set of labeled titles are an essential part of input. Aim to create such a model where we can establish relationship between set of document to its corresponding titles for prediction of unseen test data set.

This model can be considered a good model to achieve a reasonable accuracy for the proposed goal. We can make further reformation, modification, to achieve better result, but now it seems to be a good model for initial proposed work.