Volume 4, Issue 3, March 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Classification of News and Articles Using Text Pattern Mining

J.P. Maurya, Sandeep Singh
A.P.(CSE)
ICOT, Bhopal, India

Harish Patil, Pinki Jain
M.Tech [CSE]
ICOT, Bhopal, India

Abstract— Text mining is nothing but the discovery of interesting knowledge in text documents. But there is a big challenging issue that how to guarantee the quality of discovered relevant features. And that are in the text documents for describing user preferences because of the large number of terms, patterns and noise. For text mining there are basically two types of approaches; one is term based approach and another is phrase based approach. But term based approach suffered with the problem of polysemy and synonymy. And phrase based approach suffered with low frequency occurrence. Disputant are use for categorization is one of the new application of text mining where document are arrange as per the different opponent.

Index Terms- Clustering analysis, document analysis, decision support systems, ontology, text mining.

I. INTRODUCTION

Text Mining [1] is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down.

Text mining offers a solution to this problem by replacing or supplementing the human reader with automatic systems Undeterred by the text explosion. It involves analyzing a large Collection of documents to discover previously unknown Information. The information might be relationships or Patterns that are buried in the document collection and which would otherwise be extremely difficult, if not impossible, to discover. Text mining can be used to analyze natural language documents about any subject, although much of the Interest at present is coming from the biological sciences. Originally, research in text categorization addressed the binary problem, where a document is either relevant or not. Text mining involves the application of techniques from are as such as information retrieval, natural language Processing, information extraction and data mining. Information Retrieval (IR) systems identify the documents in a collection which match a user's query. The most well known IR systems are search engines such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. This is however changing with the advent of digital libraries, where the documents being retrieved are digital versions of books and journals. IR systems allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis. For example, if we are interested in mining information only about protein interactions, we might restrict our analysis to documents that contain the name of a protein, or some form of the verb 'to interact' or one of its synonyms.

II. FEATURES FOR TEXT MINING

1) Title feature

The word in sentence that also occurs in title gives high score. This is determined by counting the number of matches between the content words in a sentence and the words in the title. We calculate the score for this feature which is the ratio of the number of words in the sentence [i]that occur in the title over the number of words in title.

2) Sentence Length

This feature is useful to filter out short sentences such as datelines and author names commonly found in news articles. The short sentences are not expected to belong to the summary. We use the length of the sentence, which is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

3) Term Weight

The frequency of term occurrences within a document has often been used for calculating the importance of sentence. The score of a sentence can be calculated as the sum of the score of words in the sentence. The score of important score *wi* of word *i* can be calculated by the traditional *tf.idf* method as follows [7].

4) *Sentence Position*

Whether it is the first 5 sentences in the paragraph, sentence position in text gives the importance of the sentences. This feature can involve several items such as the position of a sentence in the document, section, and paragraph, etc., proposed the first sentence is highest ranking. The score for this feature: we consider the first 5 sentences in the paragraph.

This feature score is calculated as the following equation (5).

5) *Sentence to Sentence Similarity*

This feature is a similarity between sentences. For each sentence *S*, the similarity between *S* and each other sentence is computed by the cosine similarity measure with a resulting value between 0 and 1 [2]. The term weight *wi* and *wj* of term t to n term in sentence *Si* and *Sj* are represented as the vectors. The similarity of each sentence pair is calculated based on similarity formula (6).

Proper Noun

The sentence that contains more proper nouns (name entity) is an important and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of proper nouns that occur in sentence over the sentence length.

S_F6(S) = No. Proper nouns in S / Sentence Length (S)

7) *Thematic Word*

The number of thematic word in sentence, this feature is important because terms that occur frequently in a document are probably related to topic. The number of thematic words indicates the words with maximum possible relativity. We used the top 10 most frequent content word for consideration as thematic. The score for this feature is calculated as the ratio of the number of thematic words that occur in the sentence over the maximum summary of thematic words in the sentence.

S_F7(S) = No. Thematic word in S/Max(No. Thematic word)

III. TEXT MINING ARCHITECTURE

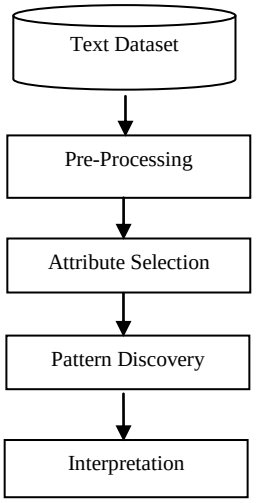

Fig.1 of Text Mining processing

1.5 Preprocessing:

All words passes to preprocessing level. Irrelevant terms are eliminated there. This process is also called as tokenization process. It consists of two kinds operations such as stop list removal, stem word removal. [8]

1.5.1 Stop List Removal: It saves the system resources. Stop word has list of words. That are deemed or irrelevant and then it is removing .It consists of articles (a, an, the), preposition (for, in, at, etc), and so on. A text document is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized representation is then used for further processing. The set of different words obtained by merging all text documents of a collection is called the dictionary of document collection shown in Fig.1.

Description: The description of the algorithms, we define first some terms and variables that will be frequently used in the following: Let D be the set of documents and T = f (t1, t2, t3…tm) be the dictionary, i.e. the set of all different terms occurring in D.

Tokenization is the process of splitting a text stream into symbols, words, phrases, or other meaningful elements called tokens. These tokens are used further text mining techniques. Word tokens are typically sent to preprocessing stages like stop-word removal and stemming, which are described later. They are also used as input for feature extraction processes.

There are many ways of tokenizing text streams into tokens. A simple method would be just to split the text on blank spaces, but better methods also takes punctuation and other sings into consideration.

The tokenizing method used in this thesis would tokenize the following text string:

"Hello! This is test number 11. It tests the word_punct-tokenizer!@ test66"
First by splitting it on blank space, then this is followed by splitting it on most special characters. The tokenized string would then consist of the following tokens.

['Hello', '!', 'This', 'is', 'test', 'number', '11', '.', 'It', 'tests', 'the', 'word_punct', '-', 'tokenizer', '!@', 'test66']

1.5.2 Stem Word Removal: The group of different words may share the same word is called as stem. For example drug, drugged, drugs, Different occurrences of the same word. Terms with a common stem would have same meaning. So it is filtering from the concern text documents. A stem is a natural group of words with equal (or very similar) meaning. After the stemming process, every word is represented by its stem. A well-known rule based stemming algorithm has been originally proposed by Porter [Por80]. He defined a set of production rules to iteratively transform (English) words into their stemming algorithm Every word is identified and the word co-occurrences are calculated with a score is calculated for each word.

Text Transformation
In order to analyze text data some separate presentation need to be developed for this that will evaluate the data effectively these is such as Bag of words, Binary Representation, TFIDF, etc. Here Bag of word is the collection of keyword for the categorization of the word. This can be understand as the BOG = {'India', 'Country', 'Production'} [6]. While in case of binary representation it shows that whether the word is present in the document or not, here if 0 represent the presence of word then 1 present word is not present in the document. TFIDF stands for Term Frequency inverse document frequency where tfidf is the product of two statistics, term frequency and inverse document frequency. Various ways for determining the exact values of both statistics exist. In the case of the **term frequency** tf, the simplest choice is to use the *raw frequency* of a term in a document, i.e. the number of times that term *t* occurs in document *d*.

The **inverse document frequency** is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient idf.

Then Tfidf = tf * idf

Feature/ Attribute Selection
Now selecting proper feature from the document is necessary for evaluation of the data, selection of representation come under this step. Here text in form of feature vector is organizing for data analysis in order to take decision. As some training is required for system to learn different feature so as per the training requirement feature need to be select.

Pattern Discovery:
On the basis of the obtained feature vector pattern are discovered for the knowledge generation[1]. This can be understand as let a document have words then how many exactly matching the feature vector. On the basis of some lower limit of the matched words in the document one can classify that whether that document is relevant or not. Mostly this is done by some kind of system that undergoes some training, then testing.

Interpretation evaluation
Once pattern are discover then the compared results of the system are need to be evaluate that either results obtain are correct or not. If the obtained results are not valid then training parameter need to be change, if results are still vary then the pattern discovered are also need to be reshuffled or change as per requirement. In order to evaluate results there are many parameter such as accuracy, precision, recall, F-score, etc. Obtaining values can be put in the mention parameter formula to get better results.

IV. CLASSIFICATION OF CONTENTIOUS NEWS ISSUES

Many topics in news may be considered controversial or have controversial parts or elements. This means that they include the study of issues which people may strongly disagree about. For example, controversial issues may include exploring why some people in the world are rich while others are poor, learning about the different sides of a conflict or even whether or not a local authority should build coastal defences to protect land from erosion.
The coverage of contentious issues of a community is an essential function of journalism. Contentious issues continuously arise in various domains, such as politics, economy, environment; each issue involves diverse participants and their different complex arguments. However, news articles are frequently biased and fail to fairly deliver conflicting arguments of the issue. It is difficult for ordinary readers to analyze the conflicting arguments and understand the contention they mostly perceive the issue passively, often through a single article.

Advanced news delivery models are required to increase awareness on conflicting views. In this paper, we present disputant relation based method for classifying news articles on contentious issues. We observe that the disputants of a contention, i.e., people who take a position and participate in the contention such as politicians, companies, stake holders, civic groups, experts, commentators, etc., are an important feature for understanding the discourse. News producers primarily shape an article on a contention by selecting and covering specific disputants (Baker. 1994). Readers also intuitively understand the contention by identifying who the opposing disputants are. The method helps readers intuitively view the news articles through the opponent based frame. It performs classification in an unsupervised manner: it dynamically identifies opposing disputant groups and classifies the articles according to their positions.
As such, it effectively helps readers contrast articles of a contention and attain balanced understanding, free from specific biased view points

Contrasting Opposing Views

Online web forums discussing ideological and political hot-topics are popular. In this work, we are interested in dual-sided opposes (there are two possible polarizing sides that the participants can take). For example, in a healthcare oppose, participants can take a for-healthcare stance or an against-healthcare stance. Participants generally pick a side (the websites provide a way for users to tag their stance) and post an argument/justification supporting their instance.

The discourse of contentious issues in news articles shows different characteristics from that studied in the sentiment classification tasks. First, the opponents of a contentious issue often discuss different topics, as discussed in the example above. Research in mass communication has showed that opposing disputants talk across each other, not by dialogue, i.e., they martial different facts and interpretations rather than to give different answers to the same topics [3].
Second, the frame of argument is not fixed as "positive versus negative." We frequently observed both sides of a contention articulating negative arguments attacking each other. The forms of arguments are also complex and diverse to classify them as positive or negative; for example, an argument may just neglect the opponent's argument without positive or negative expressions, or emphasize a different discussion point. In addition, a position of a contention can be communicated without explicit expression of opinion or sentiment. It is often conveyed through objective sentences that include carefully selected facts. For example, a news article can cast a negative light on a government program simply by covering the increase of deficit caused by it.
A number of recent works deal with debate stance recognition, which is a closely related task. They attempt to identify the side of a debate (e.g., ideological debates [4). They deal with debate posts that are all on one coherent topic, for example, Iphone versus Blackberry, and explicitly express arguments for or against to the topic, for example, for or against Iphone or Blackberry. Among these work, Somasundaram and Wiebe's work [4], [6] is similar to our work as it does not assume a fixed classification frame nor require pre specifying the discussed topics.

V. CONCLUSION

As text mining is done by different way such as term, phrase and pattern base. In this paper one concept of pattern base is explain with a mix of keywords of that document, this term has give an effective results that are highly dynamic, as it was acceptable for paper of different field. As the writing work of different articles from laboratory, organizations, press media, institutes are increasing day by day then publishing of their work is also increases which is done by most of the journals, news paper, organization only by reading their title Terms which is not the sufficient method for classifying the articles to the particular group, instead of this whole article need to be scanned with the opposing based approach in order to improve the method of classifying the into particular. For deciding specific class this work make disputant cluster of two field.

REFERENCES

1. D.A. Schon and M. Rien, Frame Reflection: Toward the Resolution of Intractable Policy Controversies. BasicBooks, 1994.
2. S. Somasundaram and J. Wiebe, "Recognizing Stances in Ideological Online Debates," Proc. NAACL HLT Workshop Computational Approaches Analysis and Generation Emotion in Text (CAAGET '10), pp. 116-124, 2010.
3. G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval" *Information Processing and Management* 24, 1988. 513-523. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) Readings in I.Retrieval. Morgan Kaufmann. pp.323-328.1997.
4. G. Salton, "Automatic Text Processing: The Transfor-mation, Analysis, and Retrieval of Information by Computer" Addison-Wesley Publishing Company.1989.
5. M. Wasson, "Using leading text for news summaries: Evaluation results and implications for commercial summarization applications" In Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL. Pp.1364-1368. 1998.
6. Disputant Relation-Based Classification for Contrasting opposing Views of Contentious News Issues Souneil park, Jungil Kim, Kyung Soon Lee, and Junehwa Song, IEEE, Dec'13.
7. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu "Effective Pattern Discovery for Text Mining". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

8. Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu. "An Ontology-Based Text-Mining Method to Cluster"
9. "Proposals for Research Project Selection". IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 42, NO. 3, MAY 2012

