IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, February 2022
Impact Factor: 6.252

Enhanced Techniques of Text Mining for Improving Text Clustering

Prof. S. H. Sangale, Prof. P. S. Chavan, Prof. Y. U. Kadam, Prof. T. B. Gade
Department of Computer Technology
K. K. Wagh Polytechnic, Nashik, Maharashtra, India
shsangale@kkwagh.edu.in, pschavan@kkwagh.edu.in, yukadam@kkwagh.edu.in, tugade@kkwagh.edu.in

Abstract: *In traditional text mining many text documents are separated and clustered by considering similarity feature among text document. Various classification algorithms are used for text categorization, but image separation is not available considering text mining concept. In this paper, we have developed image clustering and categorization technics by using the concept of text mining. Here multiple images are uploaded .every images has given some meaningful text considering the contents in it. And our fuzzy algorithm is applied to images and images are separated by applying text mining techniques.*

Keywords: Feature Extraction, Concept Mining, Feature Clustering, Sentence Extractor.

I. INTRODUCTION

Text Technically, text mining is the use of automated methods for exploiting the enormous amount of knowledge available in text documents. Text Mining represents a step forward from text retrieval mining, sometimes alternately referred to as text data mining, refers generally to the process of deriving high quality information from text Researchers like [2], [3] and others pointed that text mining is also known as Text Data Mining (TDM) and knowledge Discovery in Textual Databases (KDT). According to [4] the boundaries between data mining and text mining are fuzzy.

The difference between regular data mining and text mining is that in text mining, the patterns are extracted from natural language texts rather than from structured databases of facts. Text mining is an interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics, and computational linguistics. Preprocessing of document collection (text categorization, information extraction, term extraction), storing the intermediate representations, analyzing the intermediate representation using a selected technique such as distribution analysis, Clustering, trend analysis, and association rules, and visualizing the results are considered necessary processes in designing and implementing a text mining tool. Among the features of text mining systems/tools are:

1. A user centric process which leverages analysis
2. Technologies and computing power to access Valuable information within unstructured text data sources
3. Text mining processes are driven by natural language processing and linguistic based algorithm
4. Eliminate the need to manually read Unstructured data sources.

1.1 Importance of Text Mining

Text mining is data mining which is applied to textual data. Text is "unstructured, vague and difficult to deal with but it is the most common method for formal exchange of information.

Whereas data mining belongs in the corporate world because that's where most databases are Text mining is nothing but "nontraditional information retrieval strategies." The goal of these strategies is to reduce the effort required of users to obtain useful information from large computerized text data sources

II. WORKING OF TEXT MINING

1. Traditional keyword search retrieves documents containing pre-defined keywords. Text mining extracts precise information based on much more than just keywords, such as entities or concepts, relationships, phrases, sentences and even numerical information in context.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, February 2022
Impact Factor: 6.252

2. Text mining software tools often use computational algorithms based on Natural Language Processing, or NLP, to enable a computer to read and analyze textual information. It interprets the meaning of the text and identifies extracts, synthesizes and analyzes relevant facts and relationships that directly answer the question.
3. Text can be mined in a systematic, comprehensive and reproducible way, and business critical information can be captured automatically.
4. Powerful NLP-based queries can be run in real time across millions of documents. These can be pre-written queries.
5. Using wildcards, one can ask questions without even having to know the keywords for which he is looking for and still get back high quality, structured results.
6. One can switch in any vocabularies or thesauri to take advantage of terminology used in its own specific domain.

2.1 Techniques

Researchers in the text mining community have been trying to apply many techniques or methods such as rule-based, knowledge based, statistical and machine-learning-based approaches. However, the fundamental methods for text mining are natural language processing (NLP) and information extraction (IE) techniques.

The former technique focuses on text processing while the latter focuses on extracting information from actual texts. Once extracted, the information can then be stored in databases to be queried, data mined, summarized in a natural language and so on. The use of natural language processing techniques enables text mining tools to get closer to the semantics of a text source [6]. This is important, especially when the text mining tool is expected to discover knowledge from texts.

NLP is a technology that concerns with natural language generation (NLG) and natural language understanding (NLU). NLG uses some level of underlying linguistic representation of text, to make sure that the generated text is grammatically correct and fluent. Most NLG systems include a syntactic reliazer to ensure that grammatical rules such as subject-verb agreement are obeyed, and text planner to decide how to arrange sentences, paragraph, and other parts coherently. NLU consists of at least of one the following components:

- Tokenization, morphological or lexical analysis, syntactic analysis and semantic analysis.

In tokenization, a sentence is segmented into a list of tokens. The token represents a word or a special symbol such an exclamation mark. The complexity arises in this process when it is possible to tag a word with more than one part of speech. Syntactic analysis is a process of assigning a syntactic structure or a parse tree, to a given natural language sentence.

It determines, for instance, how a sentence is broken down into phrases, how the phrases are broken down into sub-phrases, and all the way down to the actual structure of the words used Semantic analysis is a process of translating a syntactic structure of a sentence into a semantic representation that

2.2 Information Extraction (IE)

IE involves directly with text mining process by extracting useful information from the texts. IE deals with the extraction of specified entities, events and relationships from unrestricted text sources. IE can be described as the creation of a structured representation of selected information drawn from texts. In IE natural language texts are mapped to be predefine, structured representation, or templates, which, when it is filled, represent an extract of key information from the original text [8], [9]. The goal is to find specific data or information in natural language texts.

The input can be unstructured documents like free texts that are written in natural language or the semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists. Using IE approach, events, facts and entities are extracted and stored into a structured database. Then data mining techniques can be applied to the data for discovering new knowledge. Unlike information retrieval (IR), which concerns how to identify relevant documents from a document collection, IE produces structured data ready for post-processing, which is crucial to many text mining applications.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, February 2022
Impact Factor: 6.252

According to [11] and [12] typical IE are developed using the following three steps:-

1. Text pre-processing; whose level ranges from text
2. Segmentation into sentences and sentences into tokens,
3. from tokens into full syntactic analysis;

2.3 Methods

A. Mining Plain Text This section describes the major ways in which text is mined when the input is plain natural language, rather than partially- structured Web documents. We begin with problems that involve extracting information for human consumption. Here are the various techniques which mine the plain text like text summarization, document retrieval, Information retrieval, Assessing document similarity and Text categorization

A. Text Summarization

A text summarizer produces a compressed representation of its input, which specifies human Consumption. It also contains individual documents or groups of documents. Text Compression is a related area but the output of text summarization is specific to be human-readable. The output of text compression algorithms is definitely not human-readable and it is also not actionable, It only supports decompression, that is, automatic reconstruction of the original text

B. Document Retrieval

Document retrieval is the task of identifying and returning the most relevant documents. Traditional libraries provide catalogues that allow users to identify documents based on resources which consist of metadata. Metadata is a highly structured document for summary, and successful methodologies have been developed for manually extracting metadata and for identifying relevant documents based on it, methodologies that are widely taught in library school. Automatic extraction of metadata (e.g. subjects, language, author, key-phrases) is a prime application of text mining techniques. The idea is to index every individual word in the document collection. It specifies many effective and popular document retrieval techniques.

C. Information retrieval

Information retrieval is considered as an extension to document retrieval where the documents that are returned are processed to condense or extract the particular information sought by the user. Thus document retrieval is followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage.

D. Assessing Document Similarity

Many text mining problems involve assessing the similarity between different documents; for example, assigning documents to pre-defined categories and grouping documents into natural clusters. These are the basic problems in data mining too, and have been a focus for research in text mining, perhaps because the success of different techniques can be evaluated and compared using standard, objective, measures of success.

Text categorization is the assignment of natural language documents to predefined categories according to their content. The set of categories is often called a “controlled vocabulary.” Document categorization is a long-standing traditional technique for information retrieval in libraries, where subjects rival authors as the predominant gateway to library contents—although they are far harder to assign objectively than authorship. Automatic text categorization has many practical applications, including indexing for document retrieval, automatically extracting metadata, word sense disambiguation by detecting the topics a document covers, and organizing and maintaining large catalogues of Web resources.

III. Various Terminologies of Text Mining

1. Text Mining Vs. Data Mining In Text Mining, patterns are extracted from natural language text but in Data Mining patters are extracted from databases.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, February 2022
Impact Factor: 6.252

2. Text Mining Vs. Web Mining In Text Mining, the input is free unstructured text, but in Web Mining web sources are structured.

3.1 Mining Plain Text

This section describes the major ways in which text is mined when the input is plain natural language, rather than partially- structured Web documents. We begin with problems that involve extracting information for human consumption. Here are the various techniques which mine the plain text like text summarization, document retrieval, Information retrieval, Assessing document similarity and Text categorization.

IV. APPLICATIONS

Text mining application uses unstructured textual information and examines it in attempt to discover structure and implicit meanings hidden within the text [9]. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with. Through text mining, we can uncover hidden patterns, relationships, and trends in text. [14] argued that the benefits of using text mining is to get to decision points more quickly, at least 10x speedup over previous methods, and find information which is hidden.

Reference [23] addressed that text mining enables organizations to explore interesting patterns, models, directions, trends, rules, contained in text in much the same way that data mining explores tabular or “structured” data.

4.1 Bioinformatics

There are various applications of Text mining like automatic processing of messages and emails. For example, it is possible to "filter" out automatically "junk email" based on certain terms, such messages can automatically be discarded. Such automatic systems for classifying electronic messages can also be useful in applications where messages need to be routed automatically to the most appropriate department. Another application is Analyzing warranty or insurance claims, diagnostic interviews. In some business domains, the majority of information is collected in textual form At the same time, most of the text-mining tools in biomedical domain have focused on test collections developed by individual research groups.

4.2 Business Intelligence

Of the major concerns in any business is to minimize the amount of guessing work involved in decision making. The risk of making wrong prediction should be reduced. Most of the data mining techniques are created to deal with prediction. The problem with data mining is that it can help only up to a certain point, since most of data are available in texts (reports, memos, emails, planning document, etc). Data mining and text mining techniques can complement each other The use of text mining tool in national defence security domain has become an important issue. Government agencies are investing considerable resources in the surveillance of all kinds of communication, such as email, chats in chat rooms. Email is used in many legitimate activities such as messages and documents exchange.

Unfortunately, it can also be misused, for example in the distribution of unsolicited junk mail, mailing offensive or threatening materials. Since time is critical and given the scale of the problem, it is infeasible to monitor emails or chat rooms normally. Thus automatic text mining tools offer a considerable promise in this area. text mining technology is becoming an emergence technology for national security defense. The work of [28] particularly focuses on investigating and determining the gender of the author based on the gender preferential language used by the author. They claimed that men and women use language and converse differently even though they speak the same language. The work has been conducted by using the Support Vector Machine

4.3. Challenging Issue

The major challenging issue in text mining arise from the complexity of a natural language itself. The natural language is not free from the ambiguity problem. Ambiguity means the capability of being understood in two or more possible senses or ways. Ambiguity gives a natural language its flexibility and usability, and consequently, therefore it cannot be entirely eliminated from the natural language. One word may have multiple meanings. One phrase or sentence can be

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, February 2022
Impact Factor: 6.252

interpreted in various ways, thus various meanings can be obtained. Although a number of researches have been conducted in resolving the ambiguity problem, the work is still immature and the proposed approach has been dedicated. .

V. Conclusion and Future Scope

Thus fuzzy similarity algorithm of text mining can be applied for making clusters of various images. Thus image categorization is done by using the technic of text mining .In future fuzzy algorithm can also be applied for audio, video data set classification.

References

[1]. Marcus Vinicius C. Guelpeli Ana Cristina Bicharra, "Constructed Categories for Textual Classification Using Fuzzy Similarity and Agglomerative Hierarchical Methods"
[2]. L Choochart," Web Document Classification Based on Fuzzy Association"
[3]. Shalini Puri1 and Sona Kaushik. "A Technical Study And Analysis On Fuzzy Similarity Based Models For Text Classification"
[4]. Ahmad T. Al-Taani, and Noor Aldeen K. Al-Awad "A Comparative Study of Web-pages Classification Methods using Fuzzy Operators Applied to Arabic Web-pages"
[5]. Shady Shehata, and Fakhri Karray, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering"
[6]. Shalini Puri and Sona Kaushik "An Enhanced Fuzzy Similarity Based Concept-Based Mining Model for Enhancing Text Clustering"
[7]. Shalini Puri and Sona Kaushik, "Sensitive Information on Move"
[8]. Fathi H. Saad, Omer I. E. Mohamed, and Rafa E. Al-Qutaish, "comparison of hierarchical agglomerative Algorithms for clustering medical Documents