A LITERATURE SURVEY ON TEXT DOCUMENT CLUSTERING TECHNIQUES

Mayank Sharma
Research Scholar
Department of Computer Science
OPJS University

Dr. Kalpana Midda
Research Guide
Department of Computer Science
OPJS University

Abstract:

Knowing something from a body of information is called knowledge discovery. This popular data mining method involves a number of steps, including data selection and preparation, data purification, the incorporation of previous information about the data sets, and the interpretation of precise answers from the observed findings. A subfield of knowledge finding from text data is text mining. The work that has been provided offers a thorough grasp of text mining and its uses in many real-time application domains. Text categorization and text clustering are processes included in text mining. On the other hand, unstructured, unlabeled data is used for the cluster analysis. In this paper, we examined a number of studies that look at Text Clustering methodologies across diverse genres.

Keywords: *Text Mining, Clustering technique, Domain, Data Mining,*

Introduction

Modern technological developments have made it necessary to translate all of the data into text format. This creates a demand for data to be saved and retrieved quickly and effectively. Text mining is the process of looking through a document for various keywords. The goal of this research is to investigate the field of text mining. In real-time applications, such as search engines, digital libraries, and other kinds of applications, text mining is a crucial field that is regularly employed. Most data resources in this age of artificial intelligence are provided in text format. The volume of data is too great for analysis, and it is too difficult to locate the key information.

As a result, it is a fascinating area for study and development. A broad variety of applications are included in the text classification. It is a component of artificial intelligence's knowledge process and natural language processing. The subfield of text and semantic analysis includes content mining, artificial intelligence, and content mining methods that cover numerous applications including semantic analysis, compiler design, document and large data analysis, among others. Here are some examples of comparisons between text mining and other forms of mining:

a) Data Mining

In Text Mining, patterns are extracted from natural language text rather than databases.

b) Web Mining

In Text Mining, the input is free unstructured text, whilst web sources are structured.

c) Information Retrieval

- No genuinely new information is found.
- The desired information merely coexists with other valid pieces of information.

d) Computation Linguistics (CPL) & Natural Language Processing (NLP)

- An extrapolation from Data Mining on numerical data to Data Mining from textual collections.
- CPL computes statistics across huge text collections to identify relevant patterns that are used to guide algorithms for different sub-problems in NLP, such as Parts Of Speech tagging and Word Sense Disambiguation.

The primary focus of the work provided in this study is the Text summarization and Text clustering technique. A technique for condensing a written material into a concise presentation of its most significant aspects. Essentially, to determine whether or not a certain content will meet the demands of the users. The software's inability to evaluate semantics and comprehend meaning is a significant obstacle. Extraction and abstraction are the two most prevalent techniques to automated summarization. The extractive technique picks a subset of existing words to compose an abstract. Abstractive approach creates an internal semantic representation and then generates a summary using natural language generation methods.

1. Text Mining History

It takes a lot of work to manually mine text for a certain kind of data. Since it was first launched in the middle of the 1980s, technology advancements have enabled the domain to advance from its prior problems. Text mining is a diverse field with several applications in data mining, statistics, computational semantics, machine learning, and information retrieval. Recently, text data has been used to store the majority of the data. Nowadays, the majority of research is moving in the direction of supporting several languages. This kind of system may gather information from numerous languages and can also organize related data from several language sources according to their original semantics [9].

Utilizing the vast quantity of company data that is accessible in a "unstructured" way is a hurdle. The main problem with text mining for unstructured data is addressed in an article written by H.P. Luhn in October 1958 titled "A Business Intelligence System," which describes a system that will use data-processing machines to automatically abstract and encode documents as well as create interest profiles for each of the "action points" in an organization. All written material, whether received or created internally, is automatically abstracted, given a word pattern, and routed to the proper action locations. [10]

Management information systems were first created in the 1960s, and business intelligence emerged as a software category and a field of study in the 1980s and 1990s. The focus was on relational databases' storage of numerical data. Text in "unstructured" papers is also difficult to use in practice. According to Prof. Marti A. Hearst in an essay published in Untangling Text Data Mining: A Practical Guide, the development of text analytics in its current form dates back to a shift in research focus from algorithm development to application in the 1990s.

Large text information warehouses have been discovered by the computational linguistics community to be a resource for creating better text analysis methods. Utilizing sizable internet text repositories to discover fresh information and advancements regarding the planet itself has gained recent significance. We don't need fully automated text analysis to make progress; instead, a combination of computational and user-interactive analysis may pave the way for creative outcomes.

The definition of text mining may be closer to application- or research-domain-specific if search efforts are made for it. At this point, each of them might provide a distinctive definition of text mining, motivated by the particular perspective of the application field [9]:

- Text Mining = Information Extraction.
- In this sense, text mining is comparable to information extraction, which refers to the extraction of information from texts.
- Text Mining = Text Data Mining.
- Text mining is another name for data mining, which is described as the process of applying machine learning and statistical analysis techniques and algorithms to text documents in order to find useful patterns. For that reason, it is necessary to more adequately pre-process the text documents. To extract valuable information from texts, several researchers use natural language processing, information extraction methods, or other straightforward pre-processing procedures.
- Text Mining = KDD Process.

Text mining is often found to be a procedure with a succession of fractional phases in the knowledge discovery process model, along with the usage of data mining or statistical analysis. In general, text mining is a process-oriented approach on texts that extracts information from collections of texts that has not yet been found.

2. A Brief Review of Various Research Papers

There are several newly developed text mining methods that are now accessible. These techniques seem to be able to efficiently and accurately categorize text. This section looks at several methods and equipment that effectively provide the means of categorizing documents. Additionally, this component offers knowledge about the field in which the research is being undertaken. The most significant contributions to text mining applications are presented in this part, along with current work in the field.

For extracting the profitable patterns from text sources, many data mining techniques are suggested. On the other side, an intriguing area of study is how to use and improve the patterns that have been found. The majority of the existing text mining approaches use term-based methodology, and most of them have problems with polysemy and synonymy. In recent years, research has focused on hypothesis-based pattern or phrase-based strategies, which provide better results than term-based techniques but are not supported by additional studies.

Zhong *et al.* [17] the use of pattern deployment and pattern evolution to boost the efficacy of consuming and improving the found patterns for collecting pertinent and fascinating information from data is shown as a sophisticated and operational pattern discovery approach. Numerous tests on the RCV1 data collection and TREC subjects show that the proposed method performs better.

Cai *et al.* [5] provided a cutting-edge learning method that worked in the space of the data manifold adaptive kernel. The unlabeled data points are not homogeneous and labels are costly in a variety of information processing approaches. Therefore, it is difficult to determine which unlabeled sample is the most important in order to lower the cost of labeling; for instance, the performance of the classifier may be increased if the sample has labels. For text classification, a number of active learning approaches, including SVM and Transductive Experimental Design, have been developed. The geometrical structure is ineffective in these circumstances, whereas the majority of contemporary strategies aim to identify the discriminating structure of the search space. Utilizing Laplacian graph theory, the various structure is unified into the kernel space. The underlying geometry of the data distribution is simulated by the manifold adaptive kernel space in this manner. The author chooses the most representative and discriminative data points for labeling by reducing the predicted error in comparison with the best classifier. The effectiveness of the suggested technique is shown through experimental assessment of text classification.

Zhuang *et al.* [6] Tri- factoring approach was used to limit the stable combinations of word clusters and document classes that may endure unmodified across multiple domains as the bridge of knowledge transition from source to destination domain by a non-negative matrix. The objective of cross-domain text classification is to adapt the recovered information from a labeled source of training data to unlabeled testing data; in this context, documents from the source and target domains are presented from separate distributions. In particular, the author designed a hybrid optimization framework for the two matrix tri-factorizations for the source- and target-domain data, in which the correlations between word clusters and document classes are shared. Then, they presented an iterative technique for optimization and demonstrated its theoretical efficacy. The experiments demonstrate the efficacy of the offered method. Specifically, they proved that the offered approach is capable of handling difficult cases in which standard procedures perform poorly.

Nguyen *et al*. [8], In his work, he suggested a text summarizing approach and system architecture. The system is capable of preserving the semantics of the document in order to decrease the quantity of text in a given document. Text summarization approaches, on the other hand, are very useful for document classification and cluster analysis. Therefore, the purpose of the proposed research is to determine the best technique to execute text summarization for classification, so that massive text volumes may be analyzed in an efficient and effective manner.

Navigli [15] In his article, he developed a novel method for learning semantic models for several domains. Here, the author used Wikipedia articles to further classify domain Word Sense Disambiguation terms. To appropriately develop a semantic model for each domain, first relevant words are retrieved from the domain-specific texts and then used to initiate a random walk across the

Word Net graph. For each input text, they inspected the semantic model in order to identify the appropriate domain and use the suitable-matching approach to produce Word Sense Disambiguation. The observed findings indicate an adoptable improvement in text classification and domain WSD procedure.

Moreira-Matias *et al.* [12] MECAC is an approach for building ensemble classifiers. In recent years, Text Categorization has captured the interest of the scientific community. Frequent usage of machine learning algorithms, such as Nav Bayes, Support Vector Machines, and k Nearest Neighbours, is substantiated by a number of comparison studies. In Text Classification, various ensemble classifiers have been introduced recently. In contrast, the majority of them merely offer a classification for a certain test sample. In its place, this article provides two benefits above other ensemble techniques: 1) reducing processing time since it may be executed utilizing parallel computing, and 2) discovering crucial statistics from clusters formed. It utilizes the mean co-association matrix to tackle binary Text Categorization problems. Given experiments are, on average, 2.04% more accurate than the best individual classifier on the examined datasets. Using the Friedman Test, a significance level of 0.05 was proven for these data.

Ramage *et al.* [3] According to his definition, offer a classification application for partly labelled text data. The majority of accessible data is in electronic text format, tagged with human-recognizable domains such as topic codes on academic articles and tags on web sites. Original text mining in this context requires data models, which may be used flexibly for the finding of textual patterns. This results in the detection of labels while searching for unlabeled subjects. In this circumstance, the supervised classification and unsupervised learning techniques are not suited for label prediction. These methods do not clearly display the labels. In this article, the author introduces two novel semi-supervised models for labeled text: Partially Labeled Dirichlet Allocation and the Partially Labeled Dirichlet Process. These models employ unsupervised learning to uncover the hidden domains in each labeled and unlabeled topic for demonstration purposes. The author identifies applications for qualitative case studies of online pages using tags. The presented prototype enhances the interpretability of models in comparison to conventional topic categorization algorithms. The author consumes many tags from the del.icio.us dataset. It is shown numerically that the new models have a stronger correlation with human comprehension scores than numerous strong baselines.

Sunikka *et al.* [2] developed a text-mining approach for research on personalization and customisation utilizing a traditional literature review. In order to discern between two distinct study areas, the primary characteristics of each will be examined. The Web of Science literature database creates a profile of search terms to facilitate customisation and modification. Personalization and customisation have several definitions that are often modified in scholarly collections. This article identifies the qualities for personalisation and customisation purposes.

Personalization relies heavily on approaches and online data navigational patterns; it also stresses the behavior and preferences of clients. Similarly, information collecting for modeling user behavior and customisation of recommender systems is of relevance.

Crammer *et al.* [9] investigated a number of confidence-weighted learning strategies that use a Gaussian distribution with weight vectors and are modified with each observed sample to achieve high classification accuracy rates. Margin-based learning for linear classifiers is an extension of confidence-weighted online learning. A probabilistic constraint employing distribution across classifier weights is used to replace the margin constraint. Online improvements are made when new examples are found. The distribution reveals a notion of confidence in the classifier weights, and under some circumstances it may also be seen as replacing an adaptive per-weight rate for a single learning rate. The statistical properties of the natural language learning process, where the majority of the crucial qualities are very few, served as the inspiration for confidence-weighted learning. Experiential evaluation on a variety of text categorization processes shows that the proposed method outperforms other state-of-the-art online and batch learning techniques. The online environment allows for faster learning and a better classifier arrangement for the type of distributed training that is primarily used in cloud computing environments.

Daniel R.M. et al. [4] Using text summarization and cluster analysis techniques, a suggested method aims to give an effective text classification approach. In order to categorize the text in a specific domain, a hybrid text clustering approach is proposed in this paper. The main problem with text mining and text classification is resource consumption, hence a method called feature extraction is used to decrease the quantity of text. The whole text document is represented by this condensed version. Additionally, the Euclidian distance based technique of the k-mean clustering algorithm is used to determine the degree of similarity between the domain knowledge and the accessible content. The suggested approach is put into practice and illustrated utilizing the visual studio environment, and N-cross validation procedure, a well-known data mining validation method, is used for performance analysis.

3. Conclusion

The aforementioned research demonstrates numerous methods that have been used by various Research Scholars in the fields of Text Mining and Text Summarization. There is a lot to learn about text mining. We merely touched on a few text mining-related topics. Different algorithms investigate this area.

References

[1] Andreas H., Andreas N., Gerhard P, Fraunhofer A, "A Brief Survey of Text Mining", Knowledge Discovery Group Sankt Augustin, May 13, 2005

[2] Anne S, Johanna B, "Applying text-mining to personalization and customization research literature – Who, what and where", 2012 Elsevier Ltd. All rights reserved

[3] Daniel R, Christopher D. M, Susan D, "Partially Labeled Topic Models for Interpretable Text Mining", KDD'11, August 21–24, 2011, Copyright 2011 ACM 978-1-4503-0813-7/11/08.

[4] Daniel, R.M. Shukla, A.K., “Improving Text Search Process using Text Document Clustering Approach”, ISSN 2319-7064, International Journal of Science and Research (IJSR), Volume 3 Issue 5, Page 1424 (2014)

[5] Deng C and Xiaofei H, “Manifold Adaptive Experimental Design for Text Categorization”, accepted17 Sep. 2010

[6] Fuzhen Z, Ping L, Hui X, Qing H, Yuhong X and Zhongzhi S, “Exploiting Associations between Word Clusters and Document Classes for Cross-domain Text Categorization”,27 October 2010, DOI:10.1002/sam.10099, Wiley Online Library.

[7] G. Koteswara R and Shubhamoy D, “DECISION SUPPORT FOR E-GOVERNANCE: A TEXT MINING APPROACH”, International Journal of Managing Information Technology, Vol.3, No.3,
August 2011

[8] Hien Nguyen, Eugene Santos, and Jacob Russell, “Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization”, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 41, NO. 6, NOVEMBER 2011

[9] Koby C, Mark D, Fernando P, “Confidence-Weighted Linear Classification for Text Categorization”, Journal of Machine Learning Research 13 (2012) 1891-1926

[10] Krishna, B.V.R., B. Sushma, “Novel Approach to Museums Development & \Emergence of Text Mining”, ISSN 2249-6343, International Journal of Computer Technology and Electronics Engineering (IJCTEE),
Volume 2, Issue 2

[11] Luhn, H.P. “A Business Intelligence (2010) System”, Volume 2, Number 4, Page 314 (1958), Nontopical Issue, IBM Research Journals

[12] Luís Moreira-M, João Mendes-M, João G, and Pavel B, “Text Categorization Using an Ensemble Classifier Based on a Mean Co-association Matrix”, MLDM 2012, pp. 525–539, 2012. © Springer.

[13] Miloš R, Mirjana I, “Text Mining: Approaches and Applications”, Abstract Methods and Applications in Computer Science, Vol. 38, No. 3, 2008, 227-234

[14] P. Bhargavi, B. Jyothi, S. Jyothi, K. Sekar, “Knowledge Extraction Using Rule Based Decision Tree Approach”, International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008

[15] Roberto N, Stefano F, Aitor S, Oier de L, Eneko A, “Two Birds with One Stone: Learning Semantic Models for Text Categorization and Word Sense Disambiguation”, CIKM’11, Copyright 2011 ACM 978-1-4503-0717-8/11/10

[16] Umajancy. S, Dr. Antony S T, “An Analysis on Text Mining –Text Retrieval and Text Extraction”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013.

[17] Vishal G., Gurpreet S. L, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in web Intelligence, VOL. 1, NO. 1,
AUGUST 2009

[18] Zhong, N, Li, Y, & Wu, Sheng-T, “Effective pattern discovery for text mining”. IEEE Transactions on Knowledge and Data Engineering