International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Text Document Clustering Approach: A Brief Review of Literature

Ruchika Mavis Daniel[1], Arun Kumar Shukla[2]

[1]Department of CS&IT, SSET, SHIATS (Deemed-to-be-University), Allahabad, India

Abstract: ***Knowledge discovery is a process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. Text mining is a sub domain of knowledge discovery from the text data. The presented study provides a broad way of understanding the text mining and their applications in different domain of real time applications. The text mining includes the process of text classification and text clustering. On the other hand, the cluster analysis is performed on the un-labelled and unstructured data. In this paper, I have presented a study of various research papers that explore the area of Text Clustering approaches in various genres.***

Keywords: Text Mining, Clustering approach, Domain, Data Mining, Text Summarization.

1. Introduction

The advances of the modern world have forced all the data to be converted into text form. This provides a need to save and retrieve the data in a faster and efficient way. Text Mining means to search for different keywords in the given document. This study is intended to explore the domain of text mining. Text mining is an essential domain that is frequently used in real time applications, such as search engine, digital libraries, and other kind of applications. In this era of computational intelligence most of the data resources are available in text format. The amount of data is too big to analyse and finding the important knowledge is too complex.

Therefore, it is an interesting domain of research and development. The text categorization includes a wide range of applications. It is part of knowledge process and natural language processing in artificial intelligence. The content mining, artificial intelligence and content mining techniques that include various applications such as semantics analysis, compiler design, document and big data analysis are the sub domain of text and semantic analysis. Some of the comparison between Text mining and various other types of mining are shown here:

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 (Information Access)
- 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 over large text collections in order to discover useful patterns which are used to inform algorithms for various sub-problems within NLP, e.g. Parts Of Speech tagging, and Word Sense Disambiguation.

The work presented in this paper basically revolves around the Text summarization and Text clustering approach. A process of reducing a text document to create a summary that retains the most important points. Basically to find out whether a certain document will fulfill the users' needs or not. Major difficulty is for the software to analyze semantics and interpret meaning. Generally, there are two approaches to automatic summarization: *extraction* and *abstraction. Extractive method* selects a subset of existing words to form a summary. *Abstractive method* builds an internal semantic representation and then use natural language generation techniques to create a summary.

2. Text Mining History

Manual text mining approaches require much effort in order to find specific kind of data. First introduced in mid-1980s, technological progress has allowed the domain to improve from the previous issues. Text mining is a diver's domain that is having a wide range of applications on information retrieval, machine learning, data mining, statistics, and computational semantics. Most of the information is recently stored as text data. Now in these days most of the research is progressing in the direction of multiple language support. This kind of system is able to gain information across languages and also capable to grouping similar data from different kind of language sources according to their original semantics [9].

The challenge is to take advantage of the large amount of enterprise information that is available in "unstructured" manner. In Oct 1958, an article is given by H.P. Luhn for 'A Business Intelligence System' where text mining for unstructured data is addressed as the major issue, which describes a system that will Utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the 'action points' in an organization. Both incoming and internally generated documents are automatically abstracted,

characterized by a word pattern, and sent automatically to appropriate action points. [10]

In the 1960s, initially management information systems were developed and Business Intelligence appeared in the '80s and '90s as a software category and arena of preparation. The prominence was on numerical data stored in relational databases. Additionally, text in "unstructured" documents is complex to practice. The arrival of text analytics in its existing form stops from a redeploying of research in the 1990s from algorithm improvement to application, according to Prof. Marti A. Hearst in an article of Untangling Text Data Mining:

The computational linguistics community has found that large text information stores as a resource for producing better text analysis techniques. A new prominence is the use of large online text storages to find new facts and developments about the world itself. For making development we do not need complete automated text analysis, rather, a hybridization of computational and user-interactive analysis may open the door to get innovative results.

If making effort for finding the definition of text mining, then that can be nearer to related research domain's or application's specific. At this point each of them can provide a different meaning of text mining, which is inspired by the specific viewpoint of the application area [9]:

- Text Mining = Information Extraction.
  In this context the text mining is essentially parallel to information extraction, which means the information extraction from texts.
- Text Mining = Text Data Mining.
  Text mining can be also defined by approximating to data mining as the application of algorithms and technique from the domain of machine learning and statistics analysis over text documents with the aim of discovering fruitful patterns. For that motive it is required to pre-process the text documents more appropriately. Different researchers use natural language processing, information extraction techniques or other simple pre-processing algorithms to find meaningful information from texts.
- Text Mining = KDD Process.
  In the knowledge discovery process model, it is commonly found that the text mining is a process with a series of fractional steps, among the use of data mining or statistical analysis. In general way, the extraction of information that is not yet discovered in collection of texts documents also text mining as procedure orientated methodology on texts.

3. A Brief Review of Various Research Papers

There are a large amount of text mining techniques available that are recently developed. These methods are promising to provide the accurate text categorization with efficient manner. In this section we investigate different techniques and tools that are efficiently providing the solution for document categorization. In addition to that, this section includes the understanding of the domain under which the study is conducted. This section introduces the recent studies in the domain of text mining and essential contribution on text mining applications.

Various data mining methods are proposed for extracting the fruitful patterns from text documents. On the other hand, how to use effectively and enhance the discovered patterns is an interesting domain of research. Most of the current text mining techniques involve term-based methodologies and among most of them are suffer from the issues of synonymy and polysemy. In recent years, researches also work on the hypothesis based pattern or phrase based techniques and obtains better outcomes than the term-based techniques, but different other experiments do not support hypothesis based techniques.

Zhong *et al.* [17] presents an advanced and operational pattern discovery method that involve the processes of pattern deploying and pattern evolving, to enhance the effectiveness of consuming them and enhancing the discovered patterns for obtaining relevant and interesting facts from data. Significant experimentations on RCV1 data assortment and TREC topics exhibit that the given solution provides improved performance.

Cai *et al.* [5] proposed a novel learning technique that performed in the data manifold adaptive kernel space. In various information processing methodologies, labels are expensive and the un-labeled data points are not uniform. Therefore to reduce the cost of labelling, it is much complex to discover which unlabeled sample is the most essential, for instance, the classifier performance can improve more if they contains labels. Various active learning techniques have been developed for text categorization, like SVM and Transductive Experimental Design. On the other hand, most of recent techniques are trying to find the discriminate structure of the search space, whereas the geometrical structure is not effective for such situations. The diverse structure is combined into the kernel space using Laplacian graph methodology. This way, the manifold adaptive kernel space simulates the fundamental geometry of the data distribution. By minimizing the expected error in comparison with the optimal classifier, author selects the most illustrative and discriminative data points for labelling. Experimental evaluation over text categorization demonstrates the efficiency of given approach.

Zhuang *et al.* [6] presented his work in which he reduced the stable combinations of word clusters and document classes that may persist unchanged over different domains as the bridge of knowledge transformation from the source to target domain by a non-negative matrix using tri-factorization methodology. Cross-domain text classification aims on adapting the recoverable knowledge from a labeled source of training data to an unlabeled testing data, in this context the documents from the source and target domains are demonstrated from different distributions. Specifically, author developed a hybrid optimization framework for the two matrix tri-factorizations for the source- and target-domain data, correspondingly, in which the associations among word clusters and document classes are collective between them. Then, they provided an iterative algorithm for optimization and theoretically demonstrate its effect. The experiments exhibit the effectiveness of the given technique.

Particularly, they demonstrated that the given methodology can handle some complex scenarios where baseline techniques basically do not perform effectively.

Nguyen *et al.* [8], in his article, proposed a text summarization technique and a detailed design of the system. The system is able to preserve the document's semantics in order to reduce the amount of text in concerning document. On the other hand, the text summarization techniques are much helpful in document categorization and cluster analysis. Therefore, the proposed study work is intended to find an optimum way by which the text summarization can be implemented for categorization by which the huge text documents are analysed in efficient and effective manner.

Navigli, in his paper, presented a new technique to learn semantic models for more than one domain. Here author consumed Wikipedia pages to categorize additionally for performing domain Word Sense Disambiguation. For appropriate learning a semantic model for each domain, first relevant terms are extracted from the texts from the specific domain and then these terms are utilized to start a random walk process over the WordNet graph. For individual input text, they first checked the semantic model, to select the effective domain and use the suitable-matching technique to develop Word Sense Disambiguation. The obtained results demonstrate adoptable enhancement on text categorization and domain WSD process [15].

Moreira-Matias *et al.* [12] proposed a methodology – MECAC – to construct ensemble classifiers. Text Categorization has concerned the attention of the research community in the recent years. Machine learning techniques such as Naïv Bayes, Support Vector Machines, or k Nearest Neighbours are frequently used for optimum performance that is justified in several comparative analyses. Recently, numerous ensemble classifiers are also presented in Text Classification. On the other hand, most of them only provide a category for a given test sample. In its place, in this paper, two advantages were provided over other ensemble techniques: 1) saving processing time because it can be run using parallel computing, and 2) it is able to discover essential statistics from the generated clusters. It consumes the mean co-association matrix to resolve binary Text Categorization issues. Given experiments provides on average, 2.04% better than the superlative individual classifier on the tested datasets. These results were statistically demonstrated for a significance level of 0.05 using the Friedman Test.

Ramage *et al.* [3] provide an application for classification for partially labelled text data, according to his description: most of available data in electronic text format, which is annotated with human recognizable domains, like subject codes on academic publications and tags on web pages. Original text mining in this context need to models the data, by which can flexibly use for the textual patterns discovery. That causes the detected labels while finding un-labelled subjects. In this situation supervised classification and unsupervised learning technique is not much appropriate for label prediction. These techniques do not demonstrate the labels clearly. In this paper, author presents two new semi-supervised models for labelled text, namely Partially Labeled Dirichlet Allocation and the Partially Labeled Dirichlet Process. These models use the unsupervised learning of subject for demonstration to discover the hidden domains in each label, and in similar way un-labelled. Author discovers applications for qualitative case studies of tagged web pages. The given prototype improves model interpretability over classical topic classification techniques. Author consumes various tags exist in the del.icio.us dataset. In order to demonstrate quantitatively the new models' higher association with human understanding scores over several strong baselines is provided.

Sunikka *et al.* [2] organized a text-mining technique for personalization and customization research using a classical literature evaluation. In order to differentiate, the main features of two different research domains. Making a profile of search words for personalization and customization is performed by the Web of Science literature database. Personalization and customization have various different definitions that are most of the time changed in the literature collections. The attributes for the personalization and customization purpose are identified in this paper. Personalization is strongly focuses on techniques and the web data navigational patterns; additionally it emphasizes customers' behaviour and preferences. In the same way information collection for user behaviour modelling and recommender systems personalization is an area of interest.

Crammer *et al.* [9] examined several techniques of confidence-weighted learning that consumes a Gaussian distribution with weight vectors, additionally that are updated with every observed sample to obtain high probability of accurate classification. Margin-based learning is a generalization of Confidence-weighted online learning for linear classifiers. The margin restraint is substituted by a probabilistic constraint using distribution over classifier weights. That is enhanced online as examples are detected. The distribution finds a perception of confidence on classifier weights, and in some conditions it can also be understood as substituting a single learning rate via adaptive per-weight rates. Confidence-weighted learning was motivated by the statistical characteristics of natural-language learning process; here most of the essential features are relatively limited. Experiential assessment on a range of text categorization process demonstrates that given technique improve over different state-of-the-art for online and batch learning methods, the online setting includes learn faster, and to better classifier arrangement for a kind of distributed training basically utilized in cloud computing environment.

Daniel R.M. et al. [4] proposed an algorithm that is intended to provide efficient text categorization technique using the text summarization technique and cluster analysis technique. Therefore, in this study a hybrid text clustering technique is developed for categorizing the text in a given domain. In text mining and text categorization the resource consumption is the major issue, therefore feature extraction technique is utilized to reduce the amount of text. This reduced text represents the whole text document. Additionally, the k-mean clustering algorithm's Euclidian distance based approach is utilized for finding similarity

between domain knowledge and available document. The proposed technique is implemented and demonstrated using the visual studio environment and for performance analysis the well-known data mining validation method namely N-cross validation process is consumed.

4. Conclusion

The above study shows various techniques that were implemented by different Research Scholars in the field of Text Mining and Text summarization. Text Mining is a vast Area of study. I have just presented few of the aspects of text mining. There are various algorithms that explore this field.

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", accepted 17 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 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, (2010)