© 2019 IJSRSET | Volume 6 | Issue 3 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET21841135

Self-Tuned Descriptive Document Clustering using a Predictive Network

Dr. K. Syed Kousar Niasi[1], P. Sidheshwari[2]

[1]Associate Professor, Department of Computer Science, Jamal Mohamed College. Trichy, Tamil Nadu, India
[2]Research Scholar, Department of Computer Science, Jamal Mohamed College. Trichy, Tamil Nadu, India

ABSTRACT

Document network is defined as a collection of documents that are connected by links. Document clustering become ubiquitous nowadays due to the widespread use of online databases, such as academic search engines. Topic modeling has become a widely used tool for document management because of its superior performance. However, there are few topic models differentiate the importance of documents on different topics. In this survey, can implement text rank algorithms of documents to improve topic modeling and propose to incorporate link based ranking into topic modeling. Text summarization provides an important role in information retrieval. Snippets generated by web search engines for every query result is an application of text summarization. Existing text summarization techniques shows that the indexing is done on the basis of the words present in the document and consists of an array of the posting lists. Document features such as term frequency, text length are used to allocate indexing weight to words. Specifically, topical rank is used to compute the subject stage rating of files, which indicates the significance of documents on special topics. By taking flight the topical ranking of a file as the opportunity of the record concerned in corresponding subject matter, a generalized relation is created between ranking and subject matter modeling. In this thesis, can implement topic discovery model for large number of medical database. The datasets are trained and extract the key terms based text mining and fuzzy latent semantic analysis (FLSA), a novel approach in topic modeling using fuzzy perspective. FLSA can maintain health & medical corpora redundancy problem and provides a new method to estimate the number of topics.

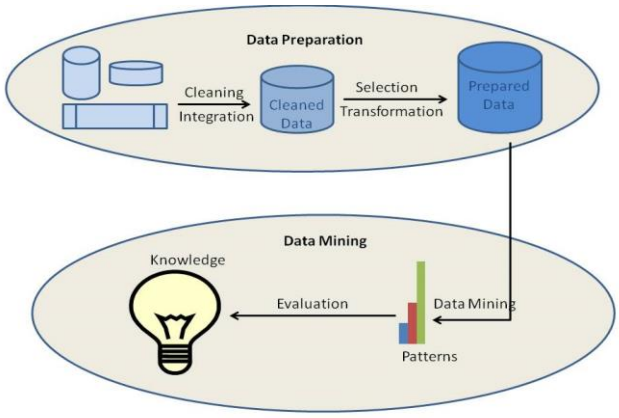
Keywords : Text Ranking, Text Mining, FLSA, Document Clustering, Text Summarization

I. INTRODUCTION

The real facts mining method is the semi-computerized or automated analysis of massive quantity of statistics to extract previously unknown, exciting styles like groups of information from facts (cluster analysis), unusual records (anomaly detection), and dependencies between records (association rule mining, sequential pattern mining). This basically involves using database techniques such as spatial indices. These patterns can then be seen as a type of summary of the input data, and may be used in further analysis, for example, in machine learning and predictive analytics. For example, the data mining method may pick out a couple of groups within the records, that could then be used to get extra correct prediction consequences by using a choice guide device. Data series, facts preparation, result interpretation and reporting are part of the

facts mining step, however do belong to the overall KDD manner as extra steps.

Different records mining processes may be classified into types: data preparation or facts pre-processing and facts mining. The first four processes involved in data cleaning, data integration, data selection and data transformation, are considered as data preparation processes. The last three processes including such as data mining, pattern evaluation and knowledge representation are combined into one process called data mining.



Process of Datamining

Text Mining

Text mining, also called to as text data mining, it roughly equivalent to text analytics, is the process of deriving important information from text. High-quality information is commonly derived from the devising of patterns and trends through means such as statistical pattern learning. Text mining typically involves the technique of structuring the input text (normally parsing, along with the addition of a few derived linguistic features and the elimination of others, and subsequent insertion into a database), deriving patterns present in the structured data, at last evaluation and interpretation of the output. 'High quality' in text mining usually refers to some aggregation of relevance, novelty, and interestingness. The overarching purpose is, essentially, to convert text into data for analysis, via application of natural language processing (NLP) and analytical methods. A commonplace software is to scan a fixed of files written in a natural language and version the record set for predictive class functions or populate a database or seek index with the facts extracted. Text Analytics, also called as text mining, is the process of examining large collections of written resources to generate new data, and to transform the unstructured data into structured data for the purpose of further analysis. Text mining identifies facts, relationships and assertions that might otherwise remain buried in the mass of textual big data. These facts are collected and converted into structured data, for analysis, visualization (e.g. through html tables, mind maps, charts), integration with structured data in data warehouses, and further refinement using machine learning (ML) systems.

Text mining has become more practical for data scientists and other users due to the development of big data platforms and deep learning algorithms that can analyze massive sets of unstructured data. Mining and analyzing of text helps organizations to find potentially valuable business insights in corporate documents, customer emails, call center logs, verbatim survey comments, social network posts, medical records and more other sources of text-based data. Increasingly, text mining facilities are also being incorporated into AI chatbots and virtual agents that companies deploy to provide automated responses to customers as part of their marketing, sales and customer service operations. Text mining is also similar to data mining, but text mining focus on text instead of more structured forms of data. However, one of the first steps in the text mining process is to collect and structure the data in some fashion so it can be used to both qualitative and quantitative analysis. Doing so commonly involves the use of natural language processing (NLP) technology, which applies computational linguistics standards to parse and interpret data sets. The upfront work implements categorizing, clustering and tagging textual content; summarizing data units; developing taxonomies; and extracting information approximately things like word frequencies and relationships between facts entities.

II. RELATED WORK

G. Brown,*et al.* [1] proposed a technique instead of trying to define feature relevance indices,derive them starting from a clearly specified objective function. The objective chooses is a well-accepted statistical principle, the conditional likelihood of the class labels given the features. As a result we are able to provide deeper insight into the feature selection problem, and achieve precisely the goal above, to retrofit numerous hand-designed heuristics into a theoretical framework. In this section we empirically examine a number of the criteria within the literature in opposition to each other. Note that we aren't pursuing an exhaustive analysis, trying to identify criterion that provides excellent overall performance overall alternatively; we mainly observe how the theoretical properties of standards relate to the similarity of the returned feature sets. While these properties are interesting, we of course must acknowledge that classification performance is the ultimate evaluation of a criterion hence we also include here classification results on UCI data sets and on the well-known benchmark NIPS Feature Selection Challenge. These are chosen to have wide different example-feature ratios, and a range of multi-class problems. The features within each data set have a different of characteristics some binary/discrete, and some continuous.

M. T. Ribeiro,*et al.* [2] produced a solution to the “trusting a prediction" problem, and selecting multiple such predictions (and explanations) as a solution to the \trusting the model" problem. Our main contributions are summarized as follows. LIME, is an algorithm that can explain the predictions of any classier or regressor in a faithful way, by approximating it locally with an interpretable model. SP-LIME,is a method that selects a set of representative instances with explanations to address the \trusting the model" problem, via sub modular optimization. Comprehensive evaluation with simulated and human subjects used to measure the impact of explanations on trust and associated tasks. In this research, non-experts using LIME are able to identify which classier from a pair generalizes better in the real world. This also explain how understanding the predictions of a neural network on images facilitates practitioners know when and why they must no longer trust a model. The process of explaining individual predictions is illustrated. It is apparent that a doctor is plenty better positioned to decide with the assist of a model if intelligible factors are provided. In this situation, an explanation is a small listing of symptoms with relative weights symptoms that either contribute to the prediction (in green) or are proof towards it (in red). Humans commonly have prior know-how about the software domain, which they could use to accept (trust) or reject a prediction if they understand the reasoning in the back of it.It has been observed, for example, that providing explanations can increase the acceptance of movie recommendations and other automated systems.

J. Chorowski, *et al.* [3] studied a new neural networks designed for classification and trained in a discriminative manner. We assume that the input data has nonnegative values. This condition is often satisfied in practice. For example, text documents in bag-of-words format or pixel intensities in images are naturally nonnegative. Categorical data encoded a using the 1-hot or thermometer-scale encoding is also nonnegative and can be used. The specified output for each sample must be a unique class label. The number of hidden neurons was chosen to yield good classification accuracy while keeping the network reasonably small. For the reduced MNIST and Reuters data the networks have 10 and 15 hidden neurons, respectively, which allow easy inspection. For the full MNIST data, the number of hidden neurons had to be increased to 150, which hinders network interpretability. However, the hidden weights can be effortlessly inspected visually when they're supplied as pictures. For most interpretability, the hidden neurons must resemble threshold gates with simplest two states: ON and OFF. In reality, their output is

squashed through the logistic sigmoid into the range (0, 1). To force the output of hidden neurons to be close to the limits of this range, the parameter λ changed into in all cases steadily extended. To determine the value of other parameters, we have first trained the network without regularization.

J. H. Lau,*et al.* [4]proposed approach is to first generate a topic label candidate set by: (1) sourcing topic label candidates from Wikipedia by querying with the top-*N* topic terms; (2) identifying the top-ranked documents with titles; and (3) further post processing the document titles to extract sub-strings. Here convert each topic label into features extracted from Wikipedia, lexical association with the topic terms in Wikipedia documents, and also lexical features for the component terms. This is used support vector regression model, which ranks each topic label candidate. The contributions in this work are: (1) the creation of a novel evaluation framework and dataset for topic label evaluation; (2) the implementation of a method for both generating and scoring topic label candidates; and (3) strong in- and cross-domain results across four identical document collections and associated topic models, demonstrating the ability of proposed method to automatically label topics with remarkable success. The processof automatic labelling of documents is a natural progression from the best topic term selection task. In that work, the authors use a re-ranking framework to produce a ranking of the top-10 topic terms based on how well each term in isolation represents a topic.

N. Aletras,*et al.* [5] implemented variations of topic representation modalities in finding relevant documents for a given query, and also measure the level of difficulty in interpreting the same topics through different representation modalities. Introduces an experiment in which three approaches to topic labeling are applied and evaluated within an exploratory search interface. The aim of the task was to identify as many documents relevant to a set of queries as possible. Each participant had to retrieve documents for 20 queries, with 3 minutes allocated for each query. In addition to the query, participants were also provided with a short description of documents that would be considered relevant for the query (e.g. News articles related to the travel and tourism industries, including articles about tourist destinations.) to assist them in identifying relevant documents. Subjects were asked to perform the document retrieval task as a two-step procedure. They were first provided with the list of LDA topics represented by a given modality (keywords, textual label or image), and a query. They were then asked to find all topics that were potentially relevant to the query. Here the topic browser interfaces for the three different modalities. In step two, the participant was presented with a list of documents related with the selected topics. Documents were presented in random order. Each document was represented using by its title, and users were able to read its content in a pop-up window. Here a subset of the documents that are associated with the topics selected in the first step.

J.T. Chien, *et al.* [6]created a new topic version to represent a bag of sentences in addition to the corresponding words. The idea of subject matter is properly understood within the community. Here, use another associated idea topic. Themes are the latent variables, which occur in one-of-a-kind level of grouped statistics, e.g., sentences, and so the concepts of subject matters and topics are exclusive. Here model the topics and topics one after the other and require the estimation of them mutually. The hierarchical subject matter and topic model is constructed. This explores a semantic tree structure of sentence-stage latent variables from a bag of sentences, whilst the phrase-level latent variables are learned from a bag of grouped phrases allocated in individual tree nodes. Build a two-degree topic model via a compound technique. The procedure of generating phrases conditions at the subject assigned to the sentence. The motivation of this paper goal to move past the phrase stage and upgrade the topic model through discovering the hierarchical family

members between the latent variables in phrase and sentence degrees. The benefit of this model is to establish a hierarchical latent variable version, that's feasible to characterize the heterogeneous documents with a couple of degrees of abstraction in different facts groupings. This model is widespread and can be applied for record summarization and many different facts systems.

U. Scaiella,*et al.* [7]proposed bag-of-words paradigm toward a more ambitious graph-of- subjects paradigm derived by using using the above subject matter-annotators, and expand a novel labeled-clustering set of rules primarily based at the spectral homes of that graph. Our solution to the SRC trouble then includes 4 fundamental steps: 1. Deploy Tagme1, a modern-day subject matter annotator for short texts, to system on-the-fly and with high accuracy the snippets back by means of a search engine. 2. Represent every snippet as a richly based graph of subjects, in which the nodes are the topics annotated via Tagme, and the edges among topics are weighted through the relatedness measure brought. 3. Then version SRC as a categorized clustering trouble over a graph along with two kinds of nodes: topics and snippets. Edges on this graph are weighted to denote either subject topic-to-topic similarities or subject matter-to-snippet memberships. The former are computed via the Wikipedia connected-structure, the latter are observed with the aid of Tagme and weighted thru right statistics. 4. Finally, layout a unique algorithm that exploits the spectral properties of the above graph to construct a terrific categorized clustering in terms of diversification and insurance of the snippet topics, coherence of clusters content material, meaningfulness of the cluster labels, and small range of balanced clusters. The final result could be a topical decomposition of the quest results lower back for a user query via one or extra search engines.

Y.-H. Tseng, *et al.* [8] carried out to mechanically create regular labels which do not always exist in the clustered files for simpler cluster interpretation. As an example, if the documents in a cluster were talking about tables, chairs, and beds, then a title labeled “furniture” would be perfect for this cluster, especially when this hypernym does not occur in it (or occurs rarely). This kind of problem was often solved by human experts, such as those, where cluster titles were given manually. To make the automatic methodology feasible, external resources such as WordNet or other hierarchical knowledge structures are used. The proposed title mapping algorithm was applied to the final-stage results of the document clustering and term clustering described above. The first set consists 6 clusters and the second has 10. Their best 5 descriptors selected by CC x TFC are shown in the second column. The proposed method was compared to a similar tool called InfoMap which is developed by the Computational Semantics Laboratory at Stanford University. This online tool identifies a set of taxonomic classes for a given set of words. It shows that WordNet is also used as its reference system, because the output classes are mostly present in WordNet’s terms. Therefore, an agent program was written to send the descriptors to Info Map and collect the results that it returns.

K. Kummamuru, *et al.* [9]proposed a novel algorithm for post-retrieval hierarchical monothetic clustering of search results to generate concept hierarchies. As the algorithm progressively identifies clusters it tries to maximize the distinctiveness of the monothetic features describing the clusters while at the same time maximizing the number of documents that can be described by the monothetic features. Compare the overall performance of DisCover with that of of other acknowledged monothetic algorithms. This assessment is primarily based on certain goal measures and it indicates that DisCover results in hierarchies with superior insurance and attain time (defined later). DisCover takes slightly greater time (19ms) than CAARD to generate hierarchies, however it takes plenty much less time than DSP. In addition to comparison based on objective measures, have also conducted user studies evaluate the performance of

the algorithms subjectively. The user studies reveal that the hierarchies obtained using DisCover are more meaningful than to those obtained by CAARD and DSP.Evaluation of the quality of taxonomies generated by a particular algorithm is an important and non-trivial task. Here briefly review some of the relevant evaluation measures used in the literature.

P. Xie, *et al.* [10] analyzed a generative model which integrates document clustering and topic modeling together. Each group possesses a fixed of neighborhood topics that capture the specific semantics of documents in this organization and a Dirichlet earlier expressing alternatives over neighborhood subjects. Assume there exist a number of global topics shared by all groups to capture the common semantics of the whole collection and a common Dirichlet prior governing the sampling of proportion vectors over global topics for all documents. Each document is a combination of local topics and global topics. Words in a document can be either generated from a global or local topic of the group to which the document belongs. In our version, the latent variables of cluster membership, file topic distribution and topics are mutually inferred. Clustering and modeling are seamlessly coupled and jointly promoted. The major contribution of this paper can be summarized as follows we propose a unified model to integrate document clustering and topic modeling together. We derive variation inference for posterior inference and parameter learning. Through experiments on two datasets, we reveal the functionality of our model in simultaneously clustering document and extracting local and international subjects. In these experiments, the input cluster number needed by clustering algorithms is set to the ground truth number of categories in corpus.

III. EXISTING WORK

Document clustering basically involves the use of descriptors and descriptor extraction. Descriptor is defined as sets of words that describe the contents within the cluster. Document clustering is commonly considered to be a centralized process. Tf-Idf based frequent candidate item-set generation has been used in the proposed work whose aim is to remove those item-set whose values are more than the pre-calculated threshold value. This process will continue until one not able to generate any more frequent candidate item sets. Finally one can get the clusters with relevant documents.Text Summarization has become grown from the early sixties but the need was different in those days, as the amount of storage capacity was very limited.

IV. METHODOLOGY

Principle Component Analysis

In PCA, high dimensional facts is converted into linearly uncorrelated variables, known as fundamental components, the use of orthogonal transformation. This transformation is done in this sort of way that the primary components are ordered inside the order of lowering variance. Even though the spectral signatures of substances resident in the document are correlated however the Eigen vectors used to derive the essential components are orthogonal to every different. Thus, the wide variety of principal additives is significantly reduced whilst as compared to unique function size, that is, the variety of bands. Even though PCA is extensively used it has the downside of excessive computational fee, huge reminiscence requirement in managing high dimensional data. Principal Components Analysis (PCA) has been one of the most commonly applied method for reducing the size of multi-dimensional data sets. It has been applied for many purposes including feature extraction and data compression. PCA method is applied with the intention of removing the redundancy existing in the data set.

Components are calculated by ranking them in their importance order. Thus, data sets can be described or visualised by a smaller number of components with limited loss of information. Component loadings show the relative positions of the variables along the new component axes. The trendy steps involved in PCA are:

1) In x-space, the mean vector is obtained.

2) Covariance matrix is computed in x-area.

3) Calculate the Eigen values and the subsequent Eigen vectors.

4) In y-space, the components are formed. Additives in the direction of maximum variation are considered as fundamental additives in this technique. Thus most effective first few additives comprise the desired records

Linear Discriminate Analysis

By generalizing fisher's linear discriminant,Linear discriminant analysis (LDA), regular discriminant analysis (NDA), or discriminant function assessment, a method applied to data, sample popularity and analyzing gadget to discover a characteristic capabilities of linear combination or separation of items or activities into two or more classes. In advance, for dimensionality discount than the later type, the linear classifier may use the ensuing mixture. In the evaluation (PDA) of the critical component, LDA is also related carefully and element analysis gives the first-rate data explanation of variables looking for linear combinations. The distinct classes of explicit data are tried by the LDA version. PCA however does now not preserve in thoughts any difference in beauty, and issue evaluation builds the function mixtures based mostly on differences in choice to similarities. Discriminate assessment is likewise unique from component evaluation in that it isn't always an interdependence approach: a distinction among independent variables and mounted variables (additionally referred to as criterion variables) needs to be made. For every elegance, a calculation is done which includes statistical homes of information. Variable (x) is entered for an unmarried and for every elegance; the variance of that variable is proposed. Over the multivariate Gaussian, the identical homes are calculated based on the specific manner and covariance matrix for more than one variable. A new brand input sets belonging to each beauty are predicted by LDA with the aid risk estimation. A prediction is done by the class which gets an output magnificence opportunity in a pleasant way. The possibilities are estimated by the Bayes' Theorem version. The output class (okay) opportunity and the possible information which belongs to every class is estimated from the given input (x) using Bayes' Theorem briefly for every elegance.

$$P(X=x|Y=x) = (fk(x) * PIk) / sum(fl(x) * PIl)$$

Where For each class, k , the base probability is referred as PIk in training data (e.g. In a two class problem, 0.5 is the probability which splits in to two halves). It is known as prior probability in Bayes' Theorem.

$$PIk = nk/n$$

The magnificence which belongs to the estimated opportunity of x is termed as f(x). To compute this f(x), Gaussian distribution characteristic has been used. The simplification of the equation is done by inserting the Gaussian into the above equation. This is known as a discriminate characteristic and the class having the highest cost is calculated and the output classification (y) can be framed:

$$Dk(x) = x * (muk/sigma^2) - (muk^2/(2*sigma^2)) + ln(PIk)$$

Where, the discriminate function is mentioned as Dk(x) for class k of the given input x. Then from the given data, the muk, sigma and PIk are anticipated.

V. RESULTS AND DISCUSSION

Author name	Title	Year	Methodology	Performance	Advantages	Disadvantages
Brown, Gavin, et al. [1]	Conditional likelihood maximisation : a unifying framework for information theoretic feature selection	2012	NIPS Feature Selection	Attempting to identify the 'winning' criterion that provides best performance.	Effectively ranks the features in descending order of their individual mutual information content	This do slightly obscure the strong link to our framework.
Ribeiro, et al. [2]	Why should i trust you?: Explaining the predictions of any classifier	2016	Random forests) And neural networks	Explains the predictions of any classifier in an interpretable and faithful manner.	To faithfully explain the predictions of any model in an interpretable manner.	In raw data is not enough to understand predictions
Chorowski, Jan, et al. [3]	Learning understandable neural networks with nonnegative weight constraints	2015	Neural Network with Pattern analysis	Constraining neurons' weights to be nonnegative improves the interpretability	Reuters text corpus the network's understandability has improved	Do not guarantee that hidden units of pruned network will have any identifiable meaning.
Lau, Jey Han, et al. [4]	Automatic labelling of topic models	2011	LDA Analysis	Rank the label candidates using a combination of association measures and lexical features.	Perform strongly over four independent sets of topics, significantly better than a benchmark method.	The method is only applicable to a hierarchical topic model.
Aletras, Nikolaos, et al. [5]	Representing topics labels for exploring digital libraries	2014	Topic Modeling and topic browsing system	Compare different topic representations, i.e. sets of topic words, textual phrases and images, in a document	Labeling methods are an effective alternative topic representation.	The time required to interpret topic representations has a direct impact on the number of retrieved

				retrieval task.		documents.
Chien, Jen-Tzung, et al. [6]	Hierarchical theme and topic modeling	2016	Bayesian non-parametrics and document summarization	A central database (Central DB), located in the wide-area network, hosts all the data required by the cloud applications.	Effective to build semantic tree structure for sentences and the correspondin g words.	Unsupervised learning beyond word level is required in many information systems.
Scaiella, Ugo, et al. [7]	Topical clustering of search results	2012	Snippets clustering and labeling	Labeled clustering of the nodes of a newly introduced graph of topics.	Graph of topics in order to enhance ad-searches or ad-page matches.	It is not always useful in discriminating topics and not always effective in describing clusters.
Tseng, Yuen-Hsien, et al. [8]	Generic title labeling for clustered documents	2010	Cluster labeling algorithm	Creating generictitles, based on external resources such as WordNet.	It can be easily extended to use other hierarchical resources for adaptable label generation.	Clustered documents require post-assignment of descriptive titles
Kummamu ru, Krishna, et al. [9]	A hierarchical monothetic document clustering algorithm for summarizatio n and browsing search results	2004	CAARD and DSP	Identifies clusters it tries to maximize the distinctiveness of the monothetic features	Hierarchies with superior coverage and reach time	This approach requires the use of ground truth
Xie, Pengtao, and Eric P. Xing.	Integrating document clustering and topic modeling	2013	Multi-grain clustering topic model (MGCTM)	Integrates document clustering and topic modeling into a unified framework	Simultaneous ly perform document clustering and modeling.	They lack the mechanism to identify local topics specific to each cluster and global topics shared by all clusters.

VI. PROPOSED WORK

Automatic keyword extraction is an important research area in text mining, natural language processing and information retrieval. Keyword extraction enables us to make text documents in a condensed way. The compact representation of files can be useful in several packages, including automated indexing, automated summarization, and automated type, clustering and filtering. For instance, text class is a website consist high dimensional feature space assignment. Hence, extracting the most important or relevant words about the content present in document and using these keywords as the features can be extremely useful. In this regard, this proposed machine examines the predictive overall performance of key-word extraction strategies (maximum common measure based totally key-word extraction, time period frequency-inverse sentence frequency primarily based key-word extraction, and TextRank algorithm) on class algorithms. In proposed system we can implement the text mining approach which contains pre-processing steps such as stop words removal, stemming words analysis. Then using automatic key word extraction contains text rank algorithm and TF-IDF calculation. After that calculate the document similarity based in word move distance and Word embedding similarity measurements. Finally predict rank of the text based longest common sequence to identify the relevant topics from large datasets.A topic model is an algorithm that aims to discover latent structures in large document collections. A topic model defines how words in a document are generated through the control of latent topics. A document corresponds to a healthcare record that contains these tokens. The means of automatic summarizationis an automatically summarized output is given when an input is applied. Remember that input is well structured document. For this there are initially pre-processes such as Sentence Segmentation, Tokenization, Removing stop words and Word Stemming. Sentence Segmentation is separating document into sentences. Tokenization means separating sentences into words. The means of removing stop words is removing frequently occurring words such as a, an, the etc. And word stemming means removing suffixes and prefixes. After pre-processing each sentence is represented by attribute of vector of features. Auto encoder approach is a statistical model of word usage that permits comparisons of semantic similarity between pieces of textual information. Encoder approach was originally designed to improve the effectiveness of information retrieval methods by performing retrieval based on the derived "semantic" content of words in a query as opposed to performing direct word matching. This approach avoids the problems of synonym, in which different words can be used to describe the same semantic concept.

Methodology

The proposed work includes the text mining algorithm to extract the key terms and cluster the data using enhanced auto encoder approach

Text Mining Algorithm

Text mining is the process of extracting meaningful information or knowledge or patterns from the available text documents from various sources. The pattern discovery from the text and document organization is a well-known problem in data mining. At current world, the quantity of stored facts has been particularly increasing each day that is generally inside the unstructured form and cannot be used for any processing to extract beneficial facts, so distinct strategies which include class, clustering and information extraction are available below the category of text mining. It contains following steps as follows

- Step 1: Choosing the scope of document
- Step 2: Tokenization
- Step 3: Token Normalization
- Step 4: Stop words removal
- Step 5: Stemming the words
- Step 6: Remove special characters

Enhanced Auto Encoder Approach:

It includes the term frequency and inverse document frequency. Then implement fuzzy c means clustering algorithm with construct document term matrix.

TF(t) = (Number of times the term t presents in a document) / (Total number of terms present in the document).

IDF(t) = log_e(Total number of documents present / Number of documents present with term t in it).

The clustering algorithm steps as follows

The n sample of the input data points is expressed as $X = \{x_1, x_2, \ldots, x_n\}$ while the corresponding cluster centres of the data points is expressed as $V = \{v_1 v_2, \ldots, v_c\}$, where c is the number of clusters. μ_{ij} is the membership degree of the data point x_i to the cluster centre v_j Fuzzy clustering computes the optimum partition based on the minimization of the objective function given that μ_{ij} satisfies

$$\sum_{i=1}^{n} \mu_{ij} = 1, 1 \leq j \leq n$$

The cluster center (i.e centroid)V_j is computed as

$$V_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}}$$

Where m is defined as the fuzziness index parameter and $m \in [1, \infty]$

Given that

$$d_{ij} = \|x_i - v_j\|$$

The dissimilarity between the centroids v_j and the data x_i is computed as

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m d_{ij}$$

Such that d_{ij} is the Euclidean distance between the i^{th}data point and the j^{th}centroid while $\mu_{ij} \in [0,1]$ and the fuzziness index parameter $m \in [1, \epsilon]$

And finally construct the document topic matrix as follows:

$$P(D_j, T_k) = P(T_k | D_j) * P(D_j)$$

D named as Document and T named as Topic. Finally provide the topics automatically based on key terms extraction

VII. CONCLUSION

Document summarization provides an instrument for efficient understanding the collection of text documents and has a number of real life applications. Semantic similarity and clustering can be performed efficiently for generating effective summary of large text collections. Summarizing large amount of textual content is a completely difficult and time consuming issue in particular while considering the semantic similarity computation in summarization method. Summarization of text collection carried out intensive text processing and computations to generate the summary of the text. In this project, we have studied text ranking and word similarity in text summarization. Intuitively, TextRank with auto encoder approach works well because it does not only rely on the local context of a text unit (vertex), but rather it takes account into information recursively drawn from the entire text (graph). Through the graphs it builds on texts, TextRank identifies connections among numerous entities in a text, and implements the idea of recommendation. A textual content unit can be recommends other related text units, and the electricity of the advice is recursively computed based on the importance of the gadgets making the recommendation. In the process of identifying valuable sentences in a text, a sentence recommends another sentence that addresses similar concepts as being useful for the overall understanding of the text. Sentences that are mostly recommended by other sentences are likely to be more informative for the given text, and will be therefore given a higher score. In future we can extend framework to implement with various algorithms in terms of accuracy. And also implement in various applications.

VIII. REFERENCES

[1]. Brown, Gavin, et al. "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection." Journal of machine learning research 13.Jan (2012): 27-66.

[2]. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

[3]. Chorowski, Jan, and Jacek M. Zurada. "Learning understandable neural networks with nonnegative weight constraints." IEEE transactions on neural networks and learning systems 26.1 (2015): 62-69.

[4]. Lau, Jey Han, et al. "Automatic labelling of topic models." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[5]. Aletras, Nikolaos, et al. "Representing topics labels for exploring digital libraries." Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries. IEEE Press, 2014.

[6]. Chien, Jen-Tzung. "Hierarchical theme and topic modeling." IEEE transactions on neural networks and learning systems27.3 (2016): 565-578.

[7]. Scaiella, Ugo, et al. "Topical clustering of search results." Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012.

[8]. Tseng, Yuen-Hsien. "Generic title labeling for clustered documents." Expert Systems with Applications 37.3 (2010): 2247-2254.

[9]. Kummamuru, Krishna, et al. "A hierarchical monothetic document clustering algorithm for summarization and browsing search results." Proceedings of the 13th international conference on World Wide Web. ACM, 2004.

[10]. Xie, Pengtao, and Eric P. Xing. "Integrating document clustering and topic modeling." arXiv preprint arXiv:1309.6874(2013).

Cite this article as :

Dr. K. Syed Kousar Niasi, P. Sidheshwari, "Self-Tuned Descriptive Document Clustering using a Predictive Network", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 6 Issue 3, pp. 320-331, May-June 2019. Available at doi : https://doi.org/10.32628/IJSRSET21841135
Journal URL : http://ijsrset.com/IJSRSET21841135