International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

A Probe on Document Clustering Methodologies and Its Performance Metrics

P.Kalpana, P. Tamije Selvy

Abstract: ***Due to the huge growth of internet usage, large volume of information flow has also been increased, which leads to the problem of information congestion. In unsupervised learning, clustering is consider as most important problem. Big quality, high dimensionality and complicated semantics are the difficult issue of document clustering.it focus on the way of identifying a structure from an unlabeled data collection. A cluster is a method in which the data items are identified and grouped based on the resemblance between the objects from a dissimilar object set. Decision of a good cluster, can be demonstrated that there is no absolute "best" criterion independent of the final objective of the clustering. A good document clustering scheme's primary objective is to minimize intra-cluster distance between papers while maximizing inter-cluster distance(using a suitable document distance measure).A distance measure(or, dually, measure of resemblance)is therefore at the core of document clustering. This assessment gives an implication about the different methods(Vector Space Model, Latent Sematic Indexing, Latent Dirichlet Allocation, Singular Value Decomposition, Doc2Vec Model, Graph model), distance measures(Euclidean Distance, Cosine Similarity, Jaccard Coefficient, Pearson Correlation Coefficient)and evaluation parameters of document clustering. This work is theoretical in nature and aims to corner the overall procedure of document clustering.***

Index Terms**: *Document clustering, Distance measure, unsupervised learning, intra-cluster.***

I. INTRODUCTION

Text clustering initially deals with characteristics or attributes for clustering. It falls under different categories,

i. ***Textbased*** -Clustering deals with the document's content
ii. ***Linkbased***-Clustering deals with Documents in the collection's link structure
iii. ***Hybrid based*** -Text & Link combination.

Clustering text at distinct document rates is well created in the Information retrieval literature, depicted in these papers as information points in a high-dimensional vector space in which every measurement relates to a distinctive keyword[1], resulting in a rectangular representation in which rows represents documents and columns represents attributes of those documents. This sort of information, referred to as "attribute information", can be clustered by wide range of algorithms.

The clustering approach began its era by defining a term (word / phrase) TF-IDF[28]. The weighting factor can be a ssessed for a feature or a particular term t in the input document. Weight of a term can be assessed using the formula given,

$$Wt,d=Tt,d \log(c/Dt) \quad (1)$$

In which: c-No.of Corpus in a document-term present in a document, Tt, d –Quantity of occurrences. .BoW(Bag of Word)Model counts the number of basic term, Regardless of grammar and word flow in a document. BoW portrayal experiences its natural extra ordinary sparsity, [29] impotence behind text information to capture high-level semantic consequences. To solve this & for improving the clustering process, synonyms &Polysemy of the term is also considered in Latent Semantic Analysis(LSA)[24] technique.

A Synonymous refers to, different words represent the same concept and Polysemous refers to the coexistence of many possible meanings for a word or phrase. Above coined problems can be solved by using LSA. It is used to represent a document by a high dimensionality space.

It additionally finds the calculated connections among the term as their semantic utilization patterns. LSA follows a dense representation. So, it is tough to index based on dimensions and to decide the quantity of topics based on heuristics. Fig. 1 Shows the Stages in Document Clustering.

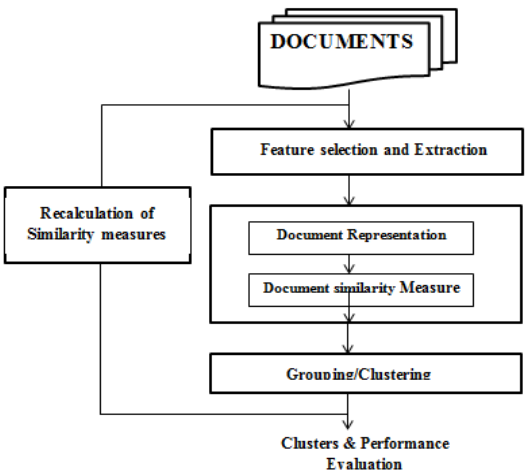

Fig. 1.Stages in Document Clustering

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Ms. P. Kalpana*, Assistant Professor, CSE, Sri Krishna College of technology, Coimbatore, Tamilnadu.
Dr. P. Tamije Selvy, Professor, CSE, Sri Krishna College of technology, Coimbatore, Tamilnadu.

The following is the structure of this document.Section I I. deals with the techniques of document representation ; Section III focuses on Motivation Challenges in document clustering. Section IV provides an ideology of document clustering methods. In Section V, Web Document Clustering Approach is investigated. Section VI exclusively speaks about Similarity measure & evaluation methods and In Section VII conclusion is made.

II. MOTIVATION & CHALLENGES IN DOCUMENT CLUSTERING

Document clustering is evolved a long before, still it is in shaping face to fine tune it. The challenges [22] involved are

1. Identifying a correct distance measure among documents by using standard equations like Euclidian, Manhattan, and maximum distance measure.
2. Identifying proper feature of documents for clustering.
3. Choosing the Initial cluster
4. Identifying Clustering objects needed & suitable algorithm for document clustering which optimize CPU usage & Memory.
5. Data abstraction
6. Evaluation ways to identify performance of clustering.
7. Problem representation about feature extraction, selection.
8. Selection of proximity measure.

III. REPRESENTATION METHODS

A. Vector Space Model

One of a traditional technique VSM [26],[5] represents the records as vectors.In this model, the conditions are autonomous from each other and dimensions in vector are free of other dimensions. It cannot work on the term based on (synonymy and polysemy) semantic association.

Dimensionality reduction needs statistical measurement of term to be done for it will be a curse of dimensionality when faced with large documents and some words have little effect on the classification &clustering of documents . VSM doesn't concentrate on semantic relations of terms. Vector public space model [21] is a extension of VSM, It introduces the similarity metric based on weight of contact.

B. Latent Sematic Indexing

It utilizes singular value decomposition (SVD) and generates latent concept space to represent documents[25] [8].The LSI model utilizes numerical algorithms to decrease sizes.The concept not only captures the meaning of individual word but also identifies short essays, sentence, and paragraphs. It provides a set of mutual constraints for finding the term presence and its count.

In k-dimensional semantic space[20], the document can be represented as the value of jth term.TF*IDF is an input matrix. To minimize the error documents are represented in several dimensions.

Words occurring in lesser counts are removed & elevated characteristics are selected from terms ordered depending on the frequency of the term.

C. Latent Dirichlet Allocation

LDA shows latent topics by using random mixtures based on probabilistic model. Likewise LSI, LDA utilizes word patterns for co-occurrence,,representation of documents. 3-Level Bayesian hierarchical model [27],[9],[15]used in LDA that models every individual data item in a given input document collection.. Gidds sampling [28] uses to weight the value of each feature. The document can be represent as [L*i*1, L*i*2,… L*ij*, …L*ik*]where L*ij* is the jth value function in the frequency space of the k-dimensional subject.

D. Singular Value Decomposition

SVM construct a subspace by using term-by-document marix.in this each are spanned by terms [13]. It is a rectangular matrix W which is broken down into three other matrices.X denotes entities of row ,Y Denotes entities of a column and Z denotes scaling values of diagonal matrix such that the original matrix (i.e., X = UV) is reconstructed by product of three components.

E. Doc2Vec Model

A doc2vec model [23],[18] is a word embedding model. It is an extension of word2vec and achieved substantial results in NLP and ML tasks. This model uses unsupervised training to obtain word vectors in the initial stage. In the second stage, it considers paragraph vector, predicts the labels using a standard classifier.

The document can be represented as [D*i*1, D*i*2, ⋯ D*ij*, ⋯ ,D*ik*]where D*ij* is the estimation of jth term in the k-dimensional semantics pace. The vector of representation predicts paragraphs in sampled cases from the collection of documents. It ranks the resemblance of the document based on the vector of text and query vector.

F. Graph Model

In graph model G={V,E) uses vertices and edges. An edge represents the relation between vertices. Document Index Graph [14] indexes phrase structure documents. It provides an informative phrase matching. It provides different level of sentence significance from original sentences. DIG is a directed graph in which each word is characterized as corpus. It scans in a sequential fashion ie.one document at a time. Degrees of overlap between documents are identified. The graph building process occupies less memory and also works in offline.

In this ,various papers are referenced and indicated as (see Table I)paper representation.

TABLE I. PAPER REPRESENTATION

Title	Indexing
A Visual Approach for Interactive Key term-Based Clustering [6]	A
Hybrid clustering analysis using improved krill herd algorithm [5]	B

A Novel Weighting Scheme Applied to Improve the Text Document Clustering Techniques [4]	C
Determining the Number of Clusters using Neural Network and Max Stable Set Problem [3]	D
The contribution of the lexical component in hybrid clustering, the case of four decades of ‘‘Scientometrics’’ [2]	E
Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering [9]	F
Large scale document categorization with fuzzy clustering. [12]	G
An adaptive version of k-medoids to deal with the uncertainty in clustering heterogeneous data using an intermediary fusion approach [11]	H
Concept Factorization With Adaptive Neighborsfor Document Clustering [7]	I
Sparse Poisson Latent Block Model for Document Clustering [8]	J
Fuzzy Bag-of-Words Model for Document Representation [10]	K
A Similarity measure for text classification and clustering [18]	L
A fragment based iterative consensus clustering algorithm with a robust similarity [16]	M

TABLE II. SIMILARITY REPRESENTATION

Paper Representation	Year	Method
A	2018	Latent Dirichlet Allocation
B	2018	VSM
C	2018	Vector Space Model
D	2018	Graph Model
E	2018	Graph Model
F	2017	Latent Semantic Indexing, graph Model
G	2017	Vector Space Model
H	2016	Similarity Matrix Fusion(SMF), Latent Semantic Indexing
I	2016	Graph Model-neighbors graph regularizer
J	2015	Vector Space Model
K	2015	DOC2VEC Model
L	2013	Vector Space Model
M	2013	Singular value decomposition

From the TableII , it is evident that in several document clustering approaches vector space is being used and Document to vector model has become almost obsolete.

IV. DOCUMENT CLUSTERING APPROACHES/ ALGORITHM

Document clustering mainly falls under two categories Traditional approach and Ontological approach are discussed in this section.

A. Traditional Approach

Initial clustering starts with data representation that uses syntax in a document clustering by using vector space model, Bag of words. As clarified in [21], this methodology depicts ventures semantic comprehension for mining documents.In this methodology depends on identifying [21] the content in documents as outlined in the accompanying Fig. 2.

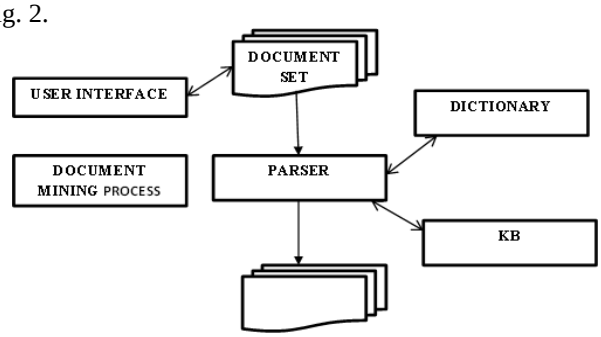

Fig. 2 Conventional Document Clustering

In case of conventional text clustering approach, parsing, similarity estimator &mining process are three major components are connected to each other. Some of the problems faced by using this approach are polysemy, ambiguity, synonymy and semantic similarities.

B. Ontological Approach

Ontology approach clusters[19] the input text files based on idea and relations to express learning and goal without semantic vagueness. It centers around the semantics of the language structure which gives quality groups. Problems like synonymy, polysemy &ambiguity problem can solved by using LSI and word sense technique. Fig. 3 shows Concept Weight Document Clustering approach.

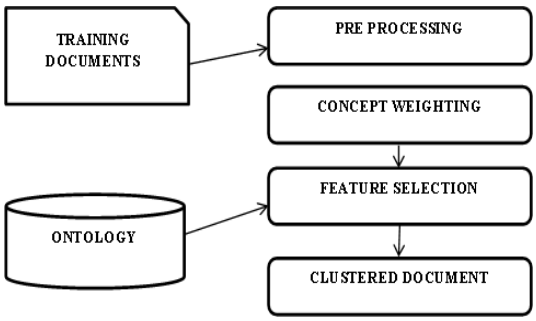

Fig. 3 Ontological approach of document clustering using concept weight.

Ontology-based approach is promising as a natural evolution of existing technologies to cope with the information onslaught. The idea behind this approach is that various kinds of people have diverse requirements with respect to text clustering. Several analyses have proved that ontological approach is highly preferable while clustering the documents.

V. SIMILARITY MEASURE & EVALUATION METHODS

I. Document Clustering Similarity Measure

Document similarities measures [31],[16] must be analyzed in the clustering.

A. Euclidean Distance

Where T = { t1,. is the word set, tm}. Tfidf as the weights of the term = tfidf(da, t).

$$D_E(\vec{t_a},\vec{t_b})(\sum_{t=1}^{m}|w_{t,a}-w_{t,b}|^2)^{1/2} \quad (3)$$

B. Cosine Similarity

The Jaccard coefficient is a metric of resemblance and range s from 0 to 1.

$$SIM_c(\vec{t_a},\vec{t_b}) = \frac{\vec{t_a}.\vec{t_b}}{|\vec{t_a}|*|\vec{t_b}|} \quad (4)$$

C. Jaccard Coefficient

$$SIM_j(\vec{t_a},\vec{t_b}) = \frac{\vec{t_a}.\vec{t_b}}{|\vec{t_a}|^2+|\vec{t_b}|^2-\vec{t_a}.\vec{t_b}} \quad (5)$$

D. Pearson Correlation Coefficient

$$|SIM_p(\vec{t_a},\vec{t_b}) = \frac{m\sum_{a=1}^{m}w_{ta}Xw_{tb}-TF_a*TF_b}{\sqrt{[m\sum_{a=1}^{m}w_{ta}^2-TF_a^2]-[m\sum_{a=1}^{m}w_{tb}^2-TF_b^2}} \quad (6)$$

E. Averaged Kullback-Leibler Divergence

$$D_{KL}(t_a \| t_b) = \sum_{t=1}^{m} w_{t,a} * \log\left(\frac{w_{t,a}}{w_{t,b}}\right) \quad (7)$$

TABLE III. DISTANCE MEASURE

Paper Representation	Year	Distance measure
A	2018	Cosine Similarity
B	2018	Euclidean Distance, Cosine Similarity
C	2018	Cosine Similarity, Jaccard Coefficient
D	2018	Euclidean Distance, Cosine Similarity
E	2018	Cosine similarity, Averaged Kullback-Leibler Divergence
F	2017	Cosine Similarity
G	2017	Euclidean Distance, Cosine distance
H	2016	Euclidean Distance, Jaccard Coefficient
I	2016	Pearson Correlation Coefficient
J	2015	Cosine Similarity
K	2015	Cosine Similarity
L	2013	Euclidean Distance, Jaccard Coefficient
M	2013	Pearson Correlation Coefficient

Table III denotes the similarity metrics used for document clustering. Cosine Similarity is fine for pattern matching while clustering the documents. If the focus is on semantics, meaning of text, then LDA (Latent Dirichlet Allocation) would be the better choice.

II. Evaluation Methods

Many evaluation steps are available to assess precision &effectiveness of the results of the clustering. Entropy,Recall, Precision, Silhouette, Co-efficient measure, Purity[28] & Inverse Purity are some of the remarkable methods.

TABLE IV. PERFORMANCE METRICS

Paper Representation	Year	Performance Metrics
A	2018	LDA,LDC, NMF
B	2018	ADDC, entropy,precision, , F-measure,recall, purity, and accuracy
C	2018	F-measure, precision, and recall.
D	2018	NMI,KMCost , Time cost , Space cost
E	2018	Precision , Recall, F-Measure
F	2017	Accuracy, F-Measure
G	2017	Precision , Recall, F-Measure
H	2016	Partition -entropy coefficient, Purity, entropy, V index, Rand Index, F-measure,Jaccard index calculations
I	2016	Entropy,Purity
J	2015	NMI,Acc and ARI
K	2015	BoW, FBoW, Cmean, BoWfull , LDA, FBoW, Cmax, LSA ,AE, WMD, FBoW, FBoW,Cmin.
L	2013	Accuracy, Entropy
M	2013	Normalize MI , Random Distance, VI distance

(See Table IV)denotes the performance metrics adopted for certain methodologies of text clustering. Different metrics have been evolved for document clustering. Precision and Recall are conventional yet powerful measures to validate the obtained results.

VI. CONCLUSION

A complete probe of Document clustering techniques has been reviewed. It also includes a motivation & challenges in document clustering. In general, this paper focuses on four different factors of document clustering.

Document representation, Clustering approach, Distance Measures and performance metrics are probed. In all these factors, the efficient one is identified and concluded.

In future, the efficient factors can be adopted to achieve better clustering results.

REFERENCES

1. Clustering: An Introduction, Available on [URL:http://home.deib.polimi.itlmatteucc/Clustering/tutorial_html/], Accessed on: 15 Aug.,2018
2. Thijs, Bart, and Wolfgang Glänzel. "The contribution of the lexical component in hybrid clustering, the case of four decades of "Scientometrics"." Scientometrics 115.1 (2018): 21-33.
3. Karim, Awatif, Chakir Loqman, and Jaouad Boumhidi. "Determining the Number of Clusters using Neural Network and Max Stable Set Problem." Procedia Computer Science127 (2018): 16-25.
4. Abualigah, Laith Mohammad, Ahamad Tajudin Khader, and Essam Said Hanandeh. "A novel weighting scheme applied to improve the text document clustering techniques." Innovative Computing, Optimization and Its Applications. Springer, Cham, 2018. 305-320.
5. Abualigah, Laith Mohammad, Ahamad Tajudin Khader, and Essam Said Hanandeh. "Hybrid clustering analysis using improved krill herd algorithm." Applied Intelligence (2018): 1-25.
6. Nourashrafeddin, Seyednaser, et al. "A Visual Approach for Interactive Keyterm-Based Clustering." ACM Transactions on Interactive Intelligent Systems (TiiS) 8.1 (2018): 6.
7. Pei, Xiaobing, Chuanbo Chen, and Weihua Gong. "Concept factorization with adaptive neighbors for document clustering." IEEE transactions on neural networks and learning systems29.2 (2018): 343-352.
8. "Sparse poisson latent block model for document clustering." IEEE Transactions on Knowledge and Data Engineering 29.7 (2017): 1563-1576.
9. Abualigah, Laith Mohammad, and Ahamad Tajudin Khader. "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering." The Journal of Supercomputing 73.11 (2017): 4773-4795.
10. Zhao, Rui, and Kezhi Mao. "Fuzzy bag-of-words model for document representation." IEEE Transactions on Fuzzy Systems (2017).794 – 804.
11. Mojahed, Aalaa, and Beatriz de la Iglesia. "An adaptive version of k-medoids to deal with the uncertainty in clustering heterogeneous data using an intermediary fusion approach." Knowledge and Information Systems 50.1 (2017): 27-52.
12. Mei, Jian-Ping, et al. "Large scale document categorization with fuzzy clustering." IEEE Transactions on Fuzzy Systems25.5 (2017): 1239-1251.
13. Gupta, Shashank, and Vasudeva Varma. "Scientific Article Recommendation by using Distributed Representations of Text and Graph." Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017.
14. Siamala Devi S, Shanmugam A., "An Integrated Harmony Search Method for Text Clustering using a Constraint based Approach", Indian Journal of Science and Technology, Vol 8(29), 73986, 2015.
15. Blomstedt, Paul, et al. "A Bayesian predictive model for clustering data of mixed discrete and continuous type." IEEE transactions on pattern analysis and machine intelligence 37.3 (2015): 489-498.
16. Chung, Chih-Heng, and Bi-Ru Dai. "A fragment-based iterative consensus clustering algorithm with a robust similarity." Knowledge and information systems 41.3 (2014): 591-609.
17. Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International Conference on Machine Learning. 2014.
18. Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee. "A similarity measure for text classification and clustering." IEEE transactions on knowledge and data engineering 26.7 (2014): 1575-1590.
19. Bharathi, G., and D. Venkatesan. "Study of ontology or thesaurus based document clustering and information retrieval." Journal of Engineering and Applied Sciences 4 (2012): 342-347.
20. Shaban, Khaled Bashir. "A Semantic Approach for Document Clustering." JSW 4.5 (2009): 391-404.
21. Pankaj Jajoo, "Document clustering ", M.Tech thesis, IIT, Kharagpur,2008
22. Hammouda, Khaled M., and Mohamed S. Kamel. "Efficient phrase-based document indexing for web document clustering." IEEE Transactions on Knowledge & Data Engineering 10 (2004): 1279-1296.
23. Ozcan, Rifat, and Y. A. Aslangdogan. "Concept based information access using ontologies and latent semantic analysis." Dept. of Computer Science and Engineering 8 (2004): 2004.
24. Marcus, Andrian, and Jonathan I. Maletic. "Recovering documentation-to-source-code traceability links using latent semantic indexing." Proceedings of the 25th international conference on software engineering. IEEE Computer Society, 2003.
25. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research3.Jan (2003): 993-1022.
26. Lee, Dik L., Huei Chuang, and Kent Seamons. "Document ranking and the vector-space model." IEEE software 14.2 (1997): 67-75.
27. Chib, Siddhartha. "Bayes regression with autoregressive errors: A Gibbs sampling approach." Journal of Econometrics58.3 (1993): 275-294.
28. Deerwester, Scott, et al. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391-407.
29. Salton, Gerard. "Automatic text processing: The transformation, analysis, and retrieval of." Reading: Addison-Wesley (1989).

AUTHORS PROFILE

Ms.P.Kalpana received her B.E(CSE)degree under Anna university, India in the year 2011 and M.E(CSE) degree from Sri Krishna College of Engineering and Technology, India in the year 2013.she is pursing her Ph.D under Anna university, India. she is currently working as a Assistant professor in the Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, India with work experience of 6 years. She has participated in various National and International conferences. she actively publishing lot of research papers in her thirst. Her areas of interests are, data mining, machine learning & image processing. Her research works include Data mining.

Dr.P.Tamijiselvy obtained her B.Tech (CSE) degree from Pondicherry Engineering College, India in the year 1996 and M.Tech (CSE) degree from Pondicherry University, India in the year 1998. She obtained her Doctorate from Anna University in the year 2013. At present she is working as a professor in the Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore with work experience of 21 years. She has participated in various National and International conferences. she actively publishing lot of research papers in her thirst. Her areas of interests are image processing, data mining and artificial intelligence. Her research works include medical imaging.