A SURVEY ON VARIOUS TECHNIQUES OF TOPIC MODELING

Jheel Bagani
Shri Shankaracharya group of Institution
Dept. of computer science and engineering
Bhilai, Chhattisgarh, India
jheelbagani22@gmail.com

Prof. Dr. Abha Choubey
Shri Shankaracharya Group of Institution
Dept. of computer science and engineering
Bhilai, Chhattisgarh, India
abha.is.shukla@gmail.com

***Abstract*—** Document clustering is the use of cluster investigation to textual documents. It is regularly utilized the system in information mining, information retrieval, learning revelation from information, design acknowledgment, and so on. In customary document clustering, a document is considered as a pack of words; where semantic importance of word is not taken into consideration. Be that as it may, to accomplish exact document clustering, feature, for example, implications of the words is imperative. Document clustering should be possible utilizing semantic approach since it considers semantic relationship among words. This paper features the issues in customary approach and in addition semantic approach. This paper distinguishes four noteworthy zones under semantic clustering and displays a survey that are studied, covering major critical work. The presented survey is utilized as a part of setting up the proposed work a similar way.

***Keywords*—** *Clustering, document clustering, semantic document clustering.*

I. INTRODUCTION

Clustering is an imperative unsupervised learning strategy. By and large, in clustering process, comparable articles are gathered into a single cluster. In this way, protests in a single cluster vary from the items in different groups. Document clustering is the way toward separating a collection of texts into little gatherings including the substance based on comparative ones [1]. The motivation behind report clustering is to help the people in information seeking and understanding [1].

In conventional document clustering strategy, the terms (words) of the documents are considered as features; notwithstanding, the semantic connections among these terms of reports aren't thought about. Because of this, issues like synonymy and polysemy, equivocalness, high dimensionality, and so forth occur. There are a few approaches to take care of this issue happens because of utilization of the customary approach. Distinctive approaches to take care of the issue incorporate the utilization of Latent Semantic Indexing (LSI), Lexical Chains, and Ontology.

Ontology can be utilized as a foundation information that can help in finding the related implications for the terms happening in reports. LSI can be utilized to tackle the issue like high dimensionality as LSI decreases the quantity of measurements in word vectors. In run of the mill use for text examination, LSI utilizes a client built corpus to make a term-by-report framework. Afterward, a technique called Singular Value Decomposition (SVD) will be connected to the term-document grid which will make a decreased term-document lattice. In the wake of making the lessened term-document framework, new report vectors will be gotten which will be utilized for processing the likeness between the query vectors and document vectors with a specific end goal to rank the reports based on the similarity. LSI helps in lessening the dimensionality of the data [1].

II. CLUSTERING

Clustering can be measured as a very important unsupervised learning issue. It comes down with searching a structure in a collection of unlabeled data.

In general, the definition of clustering could be "the process of organizing given objects into certain number of groups whose members are similar in some way". Therefore a cluster is a collection of objects which are "similar" and are "dissimilar" between them to the objects belonging to other clusters.

If the general query is given then it is extremely difficult to recognize the specific document which the user is interested in. The users are required to explore through a long list of off topics from the documents. Furthermore, internal relationships among all the documents in the search result are hardly ever presented and are left for the user.

.

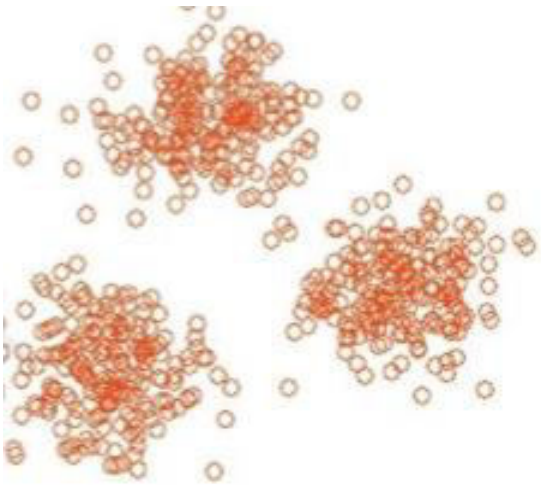

Fig. 1. Overview of Clustering

III. FUNDAMENTAL OF DOCUMENT CLUSTERING

In this area, depiction of document clustering and its two methodologies, conventional and semantic document clustering, are talked about. This segment features imperative difficulties and issues in report clustering. This segment likewise examines points of interest of semantic clustering over conventional clustering. The area quickly gives a comprehension of ontology that will be utilized as a part of our proposed fills in as foundation learning.

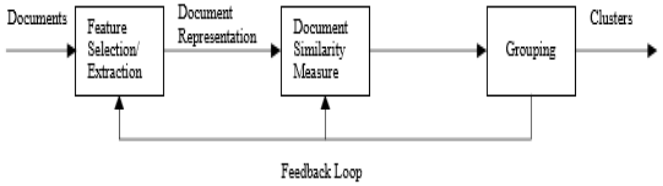

Fig. 2. A typical process of document clustering

A. *Document Clustering*

Document clustering is the assignment of consequently sorting out text documents into important clustering or gatherings. The documents in a single group share a similar point while the documents in another cluster speak to an alternate theme. Document clustering should be possible utilizing two methodologies, customary and semantic. Fig.2 demonstrates a typical procedure of document clustering.

B. *Traditional Document Clustering:*

Traditional document clustering approach utilizes "Bag of Words" show for an age of keywords that finds the recurrence of the words happening in the report. The real burden of this model is that it overlooks the semantic connection between the words. Conventional report clustering utilizes words and phrases as information features for clustering. The downside of the conventional approach is that it may not discover significant clusters for documents and furthermore now and again it can't segregate two unique groups.

C. *Semantic Document Clustering:*

Semantic document clustering is a method to cluster the documents into significant groups. In this approach, the semantic relations between the words are contemplated. The documents that are semantically identified with each other are gathered into a similar cluster and reports that are semantically inconsequential are gathered into another cluster. The semantic approach can likewise help in recognizing the subject of a cluster. The semantic approach concentrates on implications of the words and in this manner semantic approach by and large uses a lexicon to discover the implications or connection among terms for producing the keywords.

D. *Challenges in implementing Document Clustering:*

- Selection of suitable document features that are to be utilized for clustering of reports.
- Selection of suitable likeness measure to figure term-document and document report comparability.
- Selection of suitable clustering strategy for better cluster development based on likeness comes about.
- Finding suitable assessment measures to assess the nature of the clusters.
- Selection of suitable devices for executing the report clustering system.

E. *Issues in Document Clustering:*

1) *Synonymy and Polysemy:*

Synonymy is the condition of being about the same. Polysemy is the limit with respect to a word, phrase or an image to have different implications, generally related by contiguousness of significance inside a semantic field. English has numerous synonymous words, for example, "top" rather than "summit", "tiny" for "minute", and so forth. English has numerous words which are polysemous. For instance, the verb "to get" can signify "procure", "understand" (I get it), and so on.

2) *High dimensionality:*

There are different words in a document, and each word has a few implications in context of different sentences. Hence, a substantial number of words are utilized as a part of the development of feature space Moreover, each word may have a few implications brought about expansive feature space. This causes the issue of high dimensionality.

3) *Cluster labeling:*

At the point when clusters are framed without thinking about the semantic relations between the words, the subsequent groups are not precise. It ends up noticeably hard to distinguish the substance of the groups i.e. what sort of reports are available in the cluster. Therefore finding the subject of a cluster isn't simple. This issue can be alluded as cluster labeling.

F. *Advantages of Semantic Document Clustering over Customary Document Clustering:*

- Semantic approach helps in information and relationship disclosure among terms of the documents.
- Semantic approach helps in recovering the significant information as indicated by client query.
- Semantic approach can help in semantically relating the groups to each other.
- Helps in creating important groups.

- Helps in giving names to the clusters as per the substance of the groups.

IV. LITERATURE SURVEY

Tamara G. Kolda et al. [1], the huge measure of textual information accessible today is pointless unless it can be successfully and productively sought. The objective in information retrieval is to discover documents that are pertinent to a given client query. Author can speak to and document collection by a lattice whose (I, j) section is nonzero just if the ith term shows up in the jth report; hence each document relates to a column vector. The query is additionally spoken to as a column vector whose ith term is nonzero just if the ith term shows up in the query. Author score each report for pertinence by taking its inward item with the query. The most elevated scoring documents are viewed as the most applicable. Shockingly, this technique does not really recover every single important document since it is based on exacting term coordinating.

M. Andrian et al. [2], an information retrieval strategy, latent semantic indexing, is utilized to naturally distinguish traceability joins from framework documentation to program source code. The consequences of two investigations to recognize interfaces in existing programming frameworks (i.e., the LEDA library, and Albergate) are displayed. These outcomes are contrasted and other comparative compose test consequences of traceability interface recognizable proof utilizing distinctive kinds of information retrieval methods. The technique displayed demonstrates to give great outcomes by correlation and also it is an ease, profoundly adaptable strategy to apply concerning preprocessing and additionally parsing of the source code and documentation.

Jen-Yuan Yeh et al. [3], this paper proposes two ways to deal with address text summarization: changed corpus-based approach (MCBA) and LSA-based T.R.M. approach (LSA + T.R.M.). The first is a trainable summarizer, which considers a few features, including position, positive keyword, negative keyword, centrality, and the likeness to the title, to produce outlines. Two new thoughts are abused: (1) sentence positions are positioned to accentuate the significances of various sentence positions, and (2) the score work is prepared by the hereditary algorithm (GA) to get an appropriate blend of feature weights. The second uses latent semantic examination (LSA) to determine the semantic network of a report or a corpus and utilizations semantic sentence portrayal to develop a semantic text relationship outline.

Yoshihiko Gotoh et al. [4], in this paper, an approach for building blend dialect models (LMs) based on some thought of semantics is examined. To this end, a strategy known as latent semantic examination (LSA) is utilized. The approach epitomizes corpus determined semantic information and can show the shifting style of the text. Utilizing such information, the corpus texts are grouped in an unsupervised way and blend LMs are consequently made. The foremost commitment of this work is to describe the report space coming about because of the LSA displaying and to show the approach for blend LM application. Correlation is made amongst manual and programmed clustering with a specific end goal to clarify how the semantic information is communicated in the space. It is demonstrated that, utilizing semantic information, blend LMs performs superior to anything a traditional single LM with slight increment of calculation cost.

Chun-Ling Chen et al. [5], with the quick development of text documents, document clustering has turned out to be one of the fundamental systems for sorting out vast measure of reports into few significant clusters. Be that as it may, there still exist a few difficulties for report clustering, for example, high dimensionality, scalability, exactness, and significant group names, covering clusters, and separating semantics from texts. Keeping in mind the end goal to enhance the nature of document clustering comes about, we propose a compelling Fuzzy-based Multi-mark Document Clustering (FMDC) approach that integrates fuzzy affiliation manage mining with a current metaphysics WordNet to lighten these issues. In our approach, the key terms will be removed from the report set, and the underlying portrayal of all documents is additionally advanced by utilizing hypernyms of WordNet with a specific end goal to abuse the semantic relations between terms.

Jeroen De Knijff et al. [6], this paper proposes a structure to consequently develop scientific categorizations from a corpus of text reports. This system first concentrates terms from reports utilizing a grammatical form parser. These terms are then sifted utilizing area congruity, space agreement, lexical union, and auxiliary significance. The rest of the terms speak to ideas in the scientific classification. These ideas are organized in a chain of importance with either the broadened subsumption technique that documents for idea precursors in deciding the parent of an idea or a progressive clustering algorithm that utilizations different text-based window and report scopes for idea co-events.

Andreas Hotho et al. [7], Text document clustering assumes a critical part in giving natural route and perusing instruments by sorting out huge arrangements of reports into few important clusters. The sack of words portrayal utilized for these clustering strategies is regularly inadmissible as it overlooks connections between imperative terms that don't occur truly. Keeping in mind the end goal to manage the issue, we incorporate center ontologies as foundation learning into the way toward clustering text documents. Our trial assessments look at clustering systems based on categorizations of texts from Reuter's newsfeeds and on a littler space of an eLearning course about Java. In the tests, changes of results by foundation information contrasted with a pattern without foundation learning can be appeared in numerous fascinating blends.

Tingting Wei et al. [8], customary clustering algorithms don't consider the semantic connections among words so that can't precisely speak to the significance of reports. To defeat this issue, presenting semantic information from cosmology, for example, WordNet has been broadly used to enhance the nature of text clustering. Notwithstanding, there still exist a few difficulties, for example, equivalent word and polysemy, high dimensionality, separating center semantics from texts, and relegating suitable depiction for the produced clusters. In this paper, Author report our endeavor towards coordinating WordNet with lexical chains to lighten these issues. The proposed approach misuses cosmology progressive structure and relations to give a more precise evaluation of the closeness between terms of word sense disambiguation.

Guoyu Tang et al. [9], Cross-lingual document clustering is the undertaking of consequently arranging a substantial collection of multi-lingual documents into a couple of clusters, contingent upon their substance or theme. It is notable that dialect boundary and interpretation vagueness are two testing issues for cross-lingual document portrayal. To this end, Author propose to speak to cross-lingual documents through factual word senses, which are consequently found from a parallel corpus through a novel cross-lingual word sense acceptance show and a sense clustering technique.

Shibamouli Lahiri et al. [10], Keyword and key phrase extraction is an imperative issue in regular dialect preparing, with applications extending from summarization to semantic hunt to document clustering. Diagram based ways to deal with keyword and key phrase extraction evade the issue of gaining a substantial in-area preparing corpus by applying variations of PageRank algorithm on a system of words. Despite the fact that diagram based methodologies are learning lean and effectively adoptable in online frameworks, it remains to a great extent open whether they can profit by centrality measures other than PageRank. In this paper, author try different things with a variety of centrality measures on word and thing phrase collocation arranges, and examine their execution on four benchmark datasets.

TABLE I. Comparisons of various techniques and method used in existing system

Author(s) References	Data Set	Feature vector	Similarity measure	Clustering algorithm	Evaluation measures	Strength	Weakness
Tamara G. Kolda et al. 1998 [1]	MEDLINE, CRANFIELD, CISI	Word	NIL	O'Leary and Peleg	Precision, recall	Takes very less storage, queries are faster	Time to form decomposition is large.
Andrian Marcus & Jonathan I. Maletic 2003[2]	LEDA (library of efficient data types and algorithm)	Document, terms	Cosine	LSI approach, probabilistic and VSM IR methods	Precision, recall	Low cost, highly flexible methods, provides good results	Does not rely on a predefined vocabulary
Jen-Yuan Yeh et al. 2004 [3]	Data corpus (Political articles)	feature weights, document	NIL	MCBA, GA, LSA+TRM	f-measure, precision, recall	Provides more precise semantic meanings from text	Not able to explicitly capture multiple senses of a word
Yoshihiko Gotoh and Steve Renals 2007 [4]	British national corpus (BNC)	Term	Euclidean distance, cosine	k- means clustering algorithm	Entropy	Performs as lower perplexity	Difficulty in comparison of different document space
Chun-Ling Chen et al. 2010 [5]	Classic, Re0, R8, WebKB, Reuters-21578	Term	Inter similarity	k-means, bisecting k means	F-measure	Help in identifying content of cluster	Inefficient to reform cluster tree for each new insertion
Jeroen de Knijff et al. 2012 [6]	RePEc, RePub documents	Term	document co-occurrence, window-based	Hierarchical clustering algorithm	F-measure, precision, recall	NIL	Difficulty in labeling clusters

Andreas Hotho et al. 2003 [7]	Reuters-21578 newsgroup dataset	Word	Cosine	k-means, bisecting k means	Precision, purity, inverse purity	Improves performance compared to best baseline.	Beneficial effects of background knowledge may require some care
Tingting Wei et al. 2014 [8]	Reuters- 21578	Word	Modified similarity measure	K-means	Purity, entropy	Solves synonymy and polysemy, high dimensionality, cluster labelling	Do not consider implicit and explicit relationship between words
Guoyu Tang and Yunqing Xia 2015 [9]	LD corpora, TDT4, CLTC	Word	Cosine	Bisecting k means	Precision, recall, f measure	NIL	Language barrier and translation ambiguity
Shibamouli Lahiri et al. 2014 [10]	Benchmark dataset (ICSI, NUS, INSPEC, SemEval)	Word	Centrality measures	Page rank algo, graph based approach	Precision, recall, f score	Avoids the problem of acquiring a large in-domaincorpus	Requires large amount of in domain labeled data that are often expensive.

V. CONCLUSION

This paper helps to examine the noteworthy part of document for compelling clustering and mining. There is assortment of text mining applications, which contains information inside them; this information might be of number of sorts, for example, beginning of information of the documents, web logs, the connections in the documents which contain user access behavior.

A considerable measure of work has been done in display days on the issue of clustering in text collections in the database and information retrieval. All things considered, this work is generally intended for issue of unadulterated text clustering in the absence of different sort of properties. These characteristics may likewise having a ton of information for clustering aims. In this paper, we examined different diverse procedures, algorithm for powerful text clustering and mining, in the wake of concentrate these systems we reached the conclusion that, considering information for text information clustering and mining is an exceptionally astounding choice in light of the fact that if the information is connected then it give to a great degree awesome outcomes and if the information is boisterous it can be perilous to merge information into the mining procedure, since it can add clamor to the procedure. So by expelling this sort of noise information we can enhance the nature of clustering.

REFERENCES

[1] Tamara G. Kolda, and Dianne P. O'leary. "A semidiscrete matrix decomposition for latent semantic indexing information retrieval." ACM Transactions on Information Systems (TOIS) 16, no. 4 (1998): 322-346.

[2] Andrian Marcus, and Jonathan Maletic. "Recovering documentation-tosource-code traceability links using latent semantic indexing." In Software Engineering, 2003. Proceedings. 25th International Conference on, pp. 125-135. IEEE, 2003.

[3] Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, and I-Heng Meng. "Text summarization using a trainable summarizer and latent semantic analysis."Information processing & management 41, no. 1, Elsevier (2005): 75-95

[4] Yoshihiko Gotoh, and Steve Renals. "Document space models using latent semantic analysis." (1997).

[5] Chun-Ling Chen, Frank SC Tseng, and Tyne Liang. "An integration of WordNet and fuzzy association rule mining for multi-label document clustering." Data & Knowledge Engineering 69, no. 11, Elsevier (2010): 1208-1226.

[6] Jeroen De Knijff, Flavius Frasincar, and Frederik Hogenboom. "Domain taxonomy learning from text:

The subsumption method versus hierarchical clustering." Data & Knowledge Engineering 83, Elsevier (2013): 54-69.

[7] Andreas Hotho, Steffen Staab, and Gerd Stumme. "Ontologies improve text document clustering." In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, pp. 541-544. IEEE, 2003.

[8] Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, and Xianyu Bao. "A semantic approach for text clustering using WordNet and Lexical chains." Expert Systems with Applications 42, no. 4, Elsevier (2015): 2264-2275.

[9] Guoyu Tang, Yunqing Xia, Erik Cambria, Peng Jin, and Thomas Fang Zheng. "Document representation with statistical word senses in crosslingual document clustering." International Journal of Pattern Recognition and Artificial Intelligence 29, no. 02, World Scientific (2015): 1559003.

[10] Shibamouli Lahiri, Sagnik Ray Choudhury, and Cornelia Caragea. "Keyword and keyphrase extraction using centrality measures on collocation networks." arXiv preprint arXiv: 1401.6571 (2014).