Survey on Document Network Identifying Similar Cases

Pradnya Kadam[1], G.S.Mate[2]
Department of Computer Engineering
[1, 2]Rajashri Shahu College of Engineering ,Savitribai Phule Pune University, Pune

Abstract-*In the big word of internet, most of the things are become online. Now days, most of the various organizations such as hospitals, education, government offices, makes their all formalities online. To provide the easy and reliable access, such organizations, keep their all paper documents in an electronic form. To get relevant documents in such a vast database of documents is very difficult. Therefore it is necessary to develop a system, that automatically find out the relevant documents in human readable form. To achieve this system, document clustering plays very important role. In this document clustering process, numbers of documents are segmented into particular number of subgroups according to specific topic. The documents belongs to same cluster contain, similar topic wise documents. It means that heterogeneous documents are analyzed and represent uniformly. This will make easy and understandable searching of documents to humans. Even though, there are some challenging problems related to this document clustering such as huge amount of documents, its high dimensionality and sometimes complex meaning of words in particular document. Till now, various clustering approaches are presented by several authors to solve these problems of document clustering. This paper makes the survey of various recent document clustering and topic modeling approaches and compares them on the basis of technique used and their respected advantages and disadvantages, which will helpful for further research.*

Keywords-Document networks, postmarketing product surveillance, similarity, Classification, clustering

I. INTRODUCTION

Document similarity is a basic task, and can be utilized in numerous applications, for example, document classification, clustering and ranking. Traditional methodologies use bag-of- words (BOW) as document representation and compute the document similarities utilizing distinctive measures, for example, cosine, Jaccard, and dice. However, the entity phrases rather than simply words in documents can be basic for assessing the relatedness between texts. For instance, "New York" and "New York Times" represent to various meanings. "George Washington" and, "Washington" are comparable in the event that they both allude to individual,however can be somewhat diverse something else. If we can detect their names also, types, they can help us better assess whether two documents are similar. Also, the links between entities or words are also informative.

We are progressively leaving a more noteworthy measure of digital footprints. A large portion of this content is unstructured, basically in the type of content, additionally with a high level of network among various pieces of content. We allude to this data type that can be depicted as a network of entities, where each entity is connected with text content, as an document network. There are cases proliferate of such data type. Familiar with academic publications and the references linking them. For another, frequently encounter Web pages and the hyperlinks among them. In social networks, for example, LinkedIn, Facebook, or Twitter, has client profiles and associations.

Because of their significance and wide appropriateness, document networks have been a serious subject of research, especially in data recovery and link analysis. Moderately less consideration has been paid to genuinely necessary techniques for conducting exploratory investigation on document networks. Analyzing a document network is extremely challenging because of the high dimensional nature of the information. In one sense, an document can be communicated in terms of the occurrences of words (i.e., the dimensionality of content). In another sense, an document can likewise be communicated as far as its availability to alternate documents (i.e., the dimensionality of network).

This paper [1] proposes a two-stage, unsupervised methodology for finding similar documents in a corpus. To start with, we fabricate a weighted network of documents linked on the premise of their cross-referenced data, characterized as the same term(s) reported in them. A gathering of document- level terms can be considered as a structure itself (potentially related to a idea) and, in this manner, the link at the document network level incorporates the content similarity of two documents. Second, we apply certain network analysis also, statistical clustering algorithms that exploit the weighting plan built in the initial step to distinguish the groups of documents in the network topology. Since the text based methodology is completely robotized,

itcould give essential proficiency increases over the code-based strategy.

II. LITERATURE REVIEW

The paper [1] investigate the creation of document networks based on various thresholds of shared data also, diverse clustering algorithms on those systems to recognize document clusters comparative clinical cases. Authors created networks from vaccine adverse event report sets utilizing seven methodologies for connecting reports. They then connected three clustering algorithms [visualization of similarities (VOS), Louvain, k-means] to these systems and assessed their capacity to distinguish known clusters.

This paper [2] presents topical n-grams, a topic model that finds topics and also topical phrases. The probabilistic model produces words in their textual order by, for every word, first sampling a topic, then sampling its status as a unigram or bigram, and afterward sampling the word from a topic-specific unigram or bigram distribution. Along these lines our model can display "white house" as an extraordinary significance phrase in the "politic" topic, however not in the 'real estate' topic. Progressive bigrams frame longer phrases.

In paper [3] authors described the architecture of their intelligence system and illustrate the system's functionalities using several practical case studies. They also anticipate that the given patent intelligence system will be incorporated into the technology planning process to assist experts in the formulation of technology strategies.

In paper [4] authors addressed the issue of embedding a document network's high-dimensional representations in terms of text and network connectivity in a low-dimensional space. They also formulate this as a generative model tying together the various representations of a document (words, links, topics, and coordinates), which are called as PLANE. Through comprehensive experiments on four real-life datasets extracted from the Cora collection, they shown that it outperforms existing baselines in topic modeling, document embedding, and network embedding, especially in terms of the quality of embedding coordinates (as features in classification and scatter plot visualization).

In paper [5] authors given a Chinese text clustering algorithm depending on the complex network theory. The algorithm is based on the given idea. If two communities are joined by only a few intercommunity edges, all paths through the network from vertices in one community to vertices in the other must pass along one of those few edges. Given a suitable set of paths, one can count how many go along each edge in the graph, and this number they then expect to be largest for the intercommunity edges, thus providing a method for identifying them.

In paper [6] authors applied network analysis methods to the domain of linguistics and introduced a new method of detecting subjective document communities with commonly used subjective words. Test result shows us the possibility to classify subjective documents without much knowledge about linguistics. They also used an eigenvector-based community analysis method and classified test documents into positive and negative ones.

In paper [7] a document is mapped to a network, in which words are vertices and relations between words are edges. Then eccentricity centrality and degree centrality are adopted to calculate the weight of each word. Finally, the first K words with smaller weight are extracted as keywords. Compared to classical TFIDF algorithm, the algorithm proposed in paper [6] just needs to take the single document into account when extracting keywords. This given algorithm is simple and adaptive. In the meantime, improvements in precision, recall and F-score are achieved compared to TFIDF. The average recall radio is over 80% in approximate match. All of these prove that the algorithm of this paper is effective, stable and adaptive.

In paper [8], authors used semantic parsing and semantic filtering modules to specify the world knowledge to domains, and then model the specified world knowledge in the form of heterogeneous information network, which enables to represent the link information for the documents. By defining a novel document similarity measure, KnowSim (document similarity with world knowledge), the similarity between documents can be measured based on the automatically generated meta-paths in the HIN constructed from the documents.

Table 1. Survey Table

Sr. No.	Title	Paper Details	Method Used	Advantages	Disadvantages
1.	Identifying Similar Cases in Document Networks Using Cross-Reference Structures	Explore the creation of document networks based on different thresholds of shared information and different clustering algorithms on those networks to identify document clusters describing similar clinical cases.	visualization of similarities (VOS), Louvain, *k*-means	This approach supported identification of similar nodes in a document network.	Need to improve accuracy
2.	Topical n-grams: Phrase and topic discovery, with an application to information retrieval	Presents topical n-grams, a topic model that discovers topics as well as topical phrases	N-gram based Topic Models	Improved information retrieval performance on a TREC collection.	The document content is not fully utilized with these approaches since they mainly focus on the identification of patterns in the corpus rather than the initial evaluation of the information structure within each document.
3.	A patent intelligence system for strategic technology planning	Proposes a Subject–Action–Object (SAO)-based patent intelligence system and illustrates the system's functionalities using case studies.	Subject–Action–Object (SAO)-based patent intelligence system	Provides specific functionalities including identification of technology trends and significant patents, detection of novel technologies, and identification of	They defined unusual technological terms used in several technology fields analyzed in case studies, but it is also not enough to cover various technology domains.
4.	Probabilistic Latent Document Network Embedding	Study the problem of embedding, or finding a low-dimensional representation of a document network that "preserves" the data as much as possible	topic-based embedding method PLANE	it outperforms existing baselines in topic modeling, document embedding, and network embedding, especially in terms of the quality of embedding	Need to extensions such as generalizing to directed graph, and pursuing computational optimizations such as hyper-threading or parallel processing
5.	Community structure of the Chinese document network based on content similarity	A fast and efficient method for detecting community structure is proposed.	agglomerative community detection method	This method is efficient to detect community structure in complex networks	Low accuracy
6.	Subjective Document Classification Using Network Analysis	Combination of both supervised machine learning and rule-based approaches are proposed for mining feasible feature-opinion pairs from subjective review sentences.	Subjectivity Classification and Feature-Opinion Pair Mining Method	This method is very effective and have high accuracy	insufficient concept representations
7.	Keywords Extraction from Chinese Document Based on Complex Network Theory	Presents a survey of methods and approaches for keyword extraction task.	keyword extraction, graph-based methods, selectivity-based keyword extraction	Provides guidelines for future research and development of new graph-based approaches for keyword extraction	Only work on keywords extraction.

8.	KnowSim: A Document Similarity Measure on Structured Heterogeneous	Propose a method to represent a document as a typed heterogeneous information network (HIN), where the entities and relations are annotated with	Heterogeneous Information Network (HIN)	KnowSim generates impressive high-quality document clustering.	Time consuming process.

III. PROPOSE SYSTEM

This system takes VAERS dataset as an input for constructing the network and then applies the process. The system identifies the document a cluster which describes similar clinical cases, identification in those documents belongs to same clusters. Initially build the network of documents based on threshold of shared information and different clustering algorithms. Initially we build the document network by linking the reports in VAERS dataset. Then we will apply the clustering algorithm on this network to find the similar kind of documents. The report set is divided into training and testing subsets. Training subset will be used in parameter calculation for clustering algorithms and testing subset will be used for cluster evaluation or result evaluation. For evaluation of clustering algorithm, we will use three performance parameters such as recall, precision and f-measure. For this, initially we read the documents of similar cluster. For accurate and fast clustering process we will use bisect K-means clustering to divide the number of documents into similar clinical cases. We also use topic detection technique. This technique is applied on number of clusters to identify the overall topic of that cluster. Finally we get the document network of similar document. That means each cluster network contain same topic/ likely similar contain in document as an output.

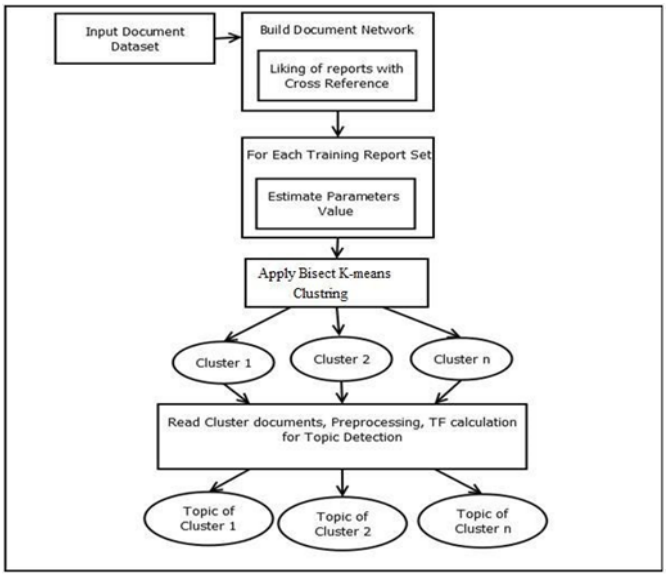

Fig 1. Propose System

IV. CONCLUSION

In this survey, we have discussed on the new problem of document clustering and classification with its features. The need of document clustering, topic identification and detectionis arises because of huge amount of paper documents in electronic format on internet. It is necessary to find the relevant document from huge document dataset. In this paper, we have presented various clustering approaches used for document clustering along with their advantages and disadvantages. From this survey we have identified some major issues of document clustering such as huge amount of data and sometimes their difficult meaning to understand. It is necessary to improve the accuracy and speed of document identification to make such systems better.

REFERENCES

[1] T. Botsis, J. Scott, E. J. Woo and R. Ball, "Identifying Similar Cases in Document Networks Using Cross-Reference Structures," in IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 6, pp. 1906-1917, Nov. 2015.

[2] X. Wang, A. McCallum, and X. Wei, “Topical n- grams: Phrase and topic discovery, with an application to information retrieval,” Proc. 7th IEEE Int. Conf. Data Mining, 2007, pp. 697–702.

[3] H. Park, K. Kim, S. Choi, and J. Yoon, “A patent intelligence system for strategic technology planning,” Expert Syst. Appl., vol. 40, no. 7, pp. 2373–2390, 2012.

[4] T. M. V. Le and H. W. Lauw, "Probabilistic Latent Document Network Embedding," 2014 IEEE International Conference on Data Mining, Shenzhen, 2014, pp. 270-279.

[5] X. Pan, J. G. Liu and G. Deng, "Community structure of the Chinese document network based on content similarity," Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on, Yantai, Shandong, 2010, pp. 1515-1519.

[6] M. Kim, B. T. Zhang and J. S. Lee, "Subjective Document Classification Using Network Analysis," Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on, Odense, 2010, pp. 365-369.

[7] J. Nan, B. Xiao, Z. Lin and Q. Xu, "Keywords Extraction from Chinese Document Based on Complex Network Theory," Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium on, Hangzhou, 2014, pp. 383-386.

[8] C. Wang, Y. Song, H. Li, M. Zhang and J. Han, "KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks," Data Mining (ICDM), 2015 IEEE International Conference on, Atlantic City, NJ, 2015, pp. 1015-1020.