SCIENTIFIC ARTICLE CLUSTERING USING R TREES AND K-MEANS ALGORITHM WITH GWO OPTIMIZATION TECHNIQUE

Abstract

Through the employment of technical methods, data resources, and service techniques, the conventional science and technology literature search primarily offers users with trustworthy and in-depth information materials and services. Increases in network, computer, and information technology have led to a proliferation of digital information resources that are having an ever-increasing effect on the conventional approach to providing knowledge services. Under the weight of enormous data volumes, certain tried-and-true technical approaches and service mechanisms fail to deliver the necessary information to users. The method indexes scientific and technical literature using R-trees, and then applies an enhanced k-mean clustering algorithm to develop a clustering model for the data. Literature search efficiency and accuracy were found to be significantly enhanced in experimental settings using datasets culled from academic journals in science and technology.

Keywords: Clustering, GWO, Optimization, K-Means.

Authors

Venkata Nagaraju Thatha
Department of Information Technology
MLR Institute of Technology
Hyderabad, Telangana, India
Nagaraju.thatha@gmail.com

B. Veera Sekharreddy
Department of Information Technology
MLR Institute of Technology
Hyderabad, Telangana, India
Bhargavisekhar68@gmail.com

G. Uday Kiran
Department of Computer Science and Engineering (AI&ML)
B V Raju Institute of Technology
Hyderabad, Telangana, India
udaykiran.goru@bvrit.ac.in

V. Srilakshmi
Department of Computer Science and Engineering (AI&ML)
B V Raju Institute of Technology
Hyderabad, Telangana, India
Srilakshmi.v@bvrit.ac.in

Srinuvasarao Sanapala
Department of Computer Science and Engineering
B V Raju Institute of Technology
Hyderabad, Telangana, India
ssrinuvas@gmail.com

I. INTRODUCTION

The amount of textual resources available online has grown exponentially. Social media platforms like Twitter, Instagram, and Facebook, as well as search engines like Google, have become increasingly popular as a result of the proliferation of mobile devices and Internet access. Social media text posts are massive amounts of unorganised data. Every day, more and more scholarly works are added to the library's archive. The proliferation of the web has also resulted in a rise in the number of online articles. While this has increased the availability of scientific articles, it can also overwhelm the reader with too much data. As a result, it's become important to refine web content searches by zeroing in on certain topics of interest.

It is crucial that a researcher who is investigating a specific issue be able to quickly and easily locate relevant papers from the many that have been written on related topics. All too frequently, the items he discovers have little to do with the topic he was looking for. Young researchers, who often lack experience with research work and good article filtering, often struggle with the task of searching for relevant publications. Another issue for contemporary scholars is that they are often only able to view the abstracts of articles without having to pay. There could be insufficient development of the specific topic in the abstract. After reading the entire article, the researcher will know if the piece is useful and if it contains the data they seek.

This method lengthens the time it takes to conduct research and produce a scholarly publication based on that research. In addition, the writers' self-reported keyword usage might not be sufficient for determining an article's topic. This is a common challenge editors encounter when trying to tailor a piece to the expertise of reviewers. It's possible that the words you declare are too vague for your intended audience, and that a single term can have multiple interpretations [1].

The majority of these computer-generated writings are very brief texts that require further scrutiny in comparison to longer, traditionally written texts [2,3]. There are many uses for short sentences online, such as on social media, in product descriptions, in ad copy, on Q&A websites [4] and in many more. The lack of context in short paragraphs makes it more challenging to extract useful information. Researchers are inspired to find answers by this problem. Short texts appear in many different contexts, such as tweets, search queries, chat messages, online reviews, and product descriptions. The disorganised character of short text, which often includes noise, slang, emoticons, misspellings, abbreviations, and grammatical errors, also makes it difficult to cluster. Tweets are a useful illustration of such difficulties. Also, different aspects of people's daily lives are reflected in different types of short texts. 500 million tweets are posted every day on Twitter as an example. Several uses exist for these brief texts, including but not limited to: trend identification [5, 6], user profiling [6, 7], event investigation [8, 9], system recommendation [8, 10], online user clustering [9], and cluster-based retrieval [2, 10].

Traditional literature searches rely on human researchers to compile references and retrieve relevant technical and scientific material. The explosion of e-books has made it impossible for humans to keep up with the volume of literature that needs to be processed by hand. In order to increase both the volume and speed with which literature is processed, when presented with a variety of publications, the effectiveness of various search engines will vary. Chinese scientific and technical writing is distinguished from other types of writing by a number of unique features: the use of standardised words;

Acquisition of information, analysis of information, data processing, and administration of knowledge have all benefited from the information age's numerous fields. Traditional technical ways of information collecting have been continuously influenced by emerging technologies [4]. Unfortunately, there is a dearth of study, practise, and implementation of such technologies in the realm of library knowledge services. Big data provides a context in which studies can help fill in gaps in the literature and address knowledge gaps. It is notable that 90% of the Internet's data volume is generated in the past six years [5], according to data given by the U.S. Internet Data Centre, and that the growth rate of data on the Internet is 50% each year, with data doubling every two years. In this paper, we show that traditional knowledge retrieval models are inadequate to meet the information. Against the backdrop of the age of big data, this study investigates the fundamental building block of the retrieval system, the "database," from the perspective of the query system. By analysing the technical means and realisation methods involved in the data collection, processing, and handling process, and by merging the clustering methods in big data, a knowledge service model is built upon the foundation of big data. In order to address the issues of a database containing a high number of diverse scientific and technical publications.

II. LITERATURE REVIEW

The title, abstract, keywords, body, and references of a scientific or technical paper all contribute to a holistic reflection of the paper's subject matter and serve as unique identifiers [6]. The logical structure of other types of documents, such as legal documents, varies widely. Although completeness and accuracy rates are widely used to evaluate retrieval systems, they are binary (a document is either in the relevant set or the irrelevant set) and therefore limited in their usefulness [9]. According to the search results of Nature and Science, two of the world's most prestigious academic journals.

The primary foci of this study are the information retrieval service model, the knowledge sharing method, and the model based on users' information needs [11]. While progress was made in areas such as software for digital library reference consulting services, knowledge service database building, data mining, and user demand analysis, all of these initiatives were conducted with the conventional information service model in mind. As a conclusion, foreign information institutions and academics have achieved some progress in the practise and research of both big data and scientific and technological documentation services, as well as having some theoretical research bases.

However, international research institutes and researchers are largely autonomous in categorising big data and knowledge service, and they have not yet explored the overlap between the two for practical purposes. Evidently, no foreign countries have yet implemented can be broken down into four categories based on their level of intelligence and semanticization [12]: differs in how intelligently and semantically they process the content of scientific and technical materials.

Retrieval systems that use semantic query expansion take search terms generated by conventional keyword searching and extend them with the use of controlled vocabularies and ontologies. Both MeSH-based query expansion and UMLS-based synonyms for PubMed queries [13] are supported. Using a user-defined weighting system, QuExT [15] ranks search results based on the importance the user places on several categories of conceptsGO2PUB [16] is able to perform semantic expansion of PubMed inquiries by taking use of the gene ontologies.

III. PROPOSED METHODOLOGY

An effective spatial index with a B+ tree-like structure is the R-tree index. R-tree takes the B-tree's concept of spatial segmentation and applies it to the process of decomposing and merging nodes in order to delete and add them without throwing the tree out of whack. Each node in an R-tree of rank M that isn't a leaf has many (P, MBR) data pairings. The smallest rectangle whose boundaries contain the corresponding children is called the Minimal Boundary Rectangle (MBR). In two dimensions, a Minimum Bounding Rectangle (MBR) is simply a rectangle; in three dimensions. To get to the node's offspring, use P as a pointer. In the R-tree, the smallest outer rectangle that contains the associated spatial item makes up the MBR at each leaf node. With the use of the spatial object's identifier, or Oi, information about that object can be retrieved. You can see an example of an R-tree. The following conditions must hold for an R-tree to be constructed successfully: (1) The non-root nodes' leaf nodes contain between m and M record indices (or entries), but the root node's leaf nodes contain no more than m. (3) Each node in the tree apart from the leaves has between m and M children. A balanced tree, like the R-tree, has no more than one level of leaf nodes. R-trees are currently split into two categories, based on their construction, namely, dynamic indexes and static indexes. Indexes for dynamic data are built dynamically, during execution.

Traditional R-tree indexes have poor query speed because of the large rectangular overlap that results from representing the data as it is now. Differentiating the R-tree's dynamic variables is made possible through the introduction of various node breaking and reinsertion strategies.

This work employs an R-tree indexing-based clustering approach to retrieve data from scholarly and technical publications quickly and efficiently. If the distribution law of the data is unknown, then presetting the clustering centres will lead to inaccurate final clustering results, which will have a negative impact on the efficiency of the R-tree clustering model index. In this study, the GWO approach is developed to build the R-tree model, which is used to efficiently estimate the clustering centres.

Algorithm:

- Create a population of n wolf locations by randomly selecting K cluster centres.
- Set the wolf's initials to a, A, and C.
- Find out how effective each search agent is.
- Cluster the data by computing the distance from each document to the cluster's centroid.

- The fitness score must be calculated.
- The optimal search agent is X.
- For the second-best search bot, substitute X.
- Third-best search bot equals X.

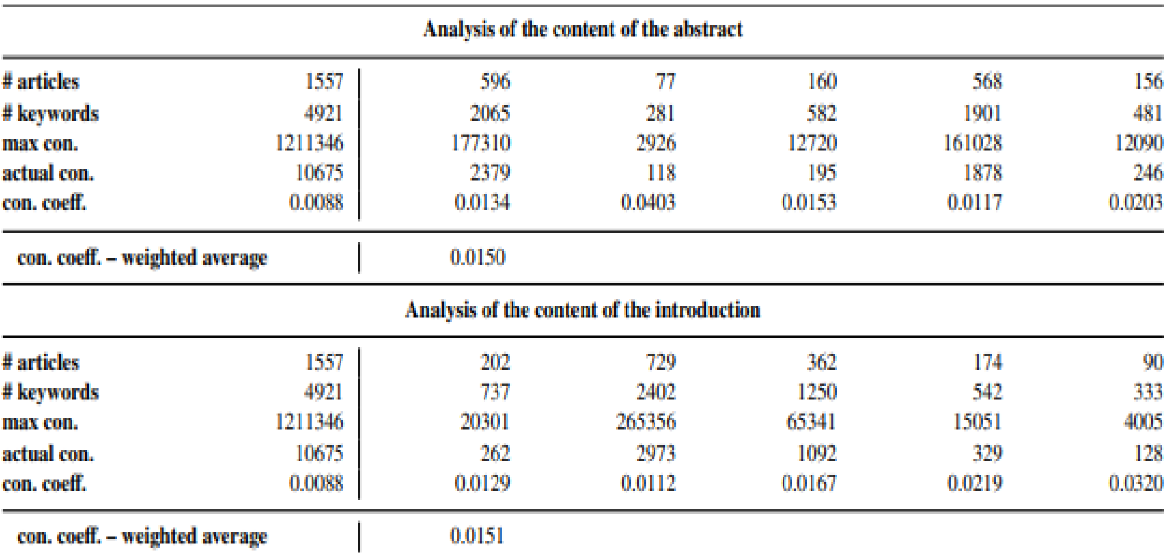
While Hold off stopping until the conditions are met. The current search agent's location should be updated (Equation (22))

While for each wolf encountered.
End
Refresh a, A, and C.
Find out how effective each search agent is.
Cluster the data by computing the distance from each document to the cluster's centroid.
The fitness score must be calculated.
Please refresh X, X, and X.
End return position

IV. RESULTS

We used a dataset comprised of 1557 genuine scientific publications published between 2017 and 2022 in an open-access journal with an impact factor of 2.0 or above for our studies. Since the publications were only available in PDF format, preprocessing was required to convert the PDFs into text and then extract the pertinent passages. The hardest step is always finding the right number of clusters, and that's especially true when working with an unstructured dataset [17]. We intentionally selected to increment the number of clusters in our study by two for each set of results we present. However, in-depth study is required to determine the optimum number of clusters. The K-means technique is used repeatedly as an unsupervised learning algorithm for a fixed number of groups [18], among other applications.

Table1: Experimental Results of proposed method for division in to five clusters.



Analysis of the content of the abstract						
# articles	1557	596	77	160	568	156
# keywords	4921	2065	281	582	1901	481
max con.	1211346	177310	2926	12720	161028	12090
actual con.	10675	2379	118	195	1878	246
con. coeff.	0.0088	0.0134	0.0403	0.0153	0.0117	0.0203
con. coeff. – weighted average		0.0150				
Analysis of the content of the introduction						
# articles	1557	202	729	362	174	90
# keywords	4921	737	2402	1250	542	333
max con.	1211346	20301	265356	65341	15051	4005
actual con.	10675	262	2973	1092	329	128
con. coeff.	0.0088	0.0129	0.0112	0.0167	0.0219	0.0320
con. coeff. – weighted average		0.0151				

The number of clusters, as well as TF-IDF, TF, and binary metrics, should be used to evaluate experimental results. Already at this point, it is clear that the TF-IDF metric provides the most insightful outcomes. This is because in the case of the binary measure, all we know is whether or not a given word was in the document's content, whereas in the case of TF, we know the number of times a given word appeared (albeit this leads to enormous discrepancies). When looking at the article's beginning, many more discrepancies become apparent. When there are more words to examine, each cluster contains fewer documents, and one cluster becomes the focal point for the rest. As can be observed, such a condition holds true for the TF measure even before the abstract is analysed. With two documents, a cluster is produced, and with all of the documents, a third. Although the binary measure is already described by comparable behavior to TF in the introduction, it appears to produce better results.

V. CONCLUSION

STC is a challenging issue because of the proliferation of increasingly brief texts generated by internet users and social media apps. Common issues in STC include data noise, sparsity, excessive dimensionality, and a lack of information. The search for and development of effective clustering algorithms have become pressing problems. Current STC algorithms can be made more effective with further research into the state of the art in text representation methods (Appl. Sci. 2023, 13, 342). In this paper, we provide a comprehensive review of the research done on STC. The summary discusses how STC can be used. In this article, we introduce STC and outline its steps in detail. We discuss the strategies employed in the depiction of short texts, together with their benefits and drawbacks, and the results of employing various approaches to short texts. Furthermore, we detail the most important deep learning techniques for text. Methods like TF-IDF vectors and BOW, which generate sparse and high-dimensional feature vectors that are less distinctive for calculating distance, perform well in some experiments but badly in others. Poor clustering accuracy can be avoided by addressing associated difficulties in text representation that are short on words.

REFERENCES

[1] Yang, S.; Huang, G.; Ofoghi, B.; Yearwood, J. Short text similarity measurement using context-aware weighted biterms. Concurr. Comput. Pract. Exp. 2020, 34, e5765.

[2] Zhang, W.; Dong, C.; Yin, J.; Wang, J. Attentive representation learning with adversarial training for short text clustering. IEEE Trans. Knowl. Data Eng. 2021, 34, 5196–5210.

[3] Yu, Z.; Wang, H.; Lin, X.; Wang, M. Understanding short texts through semantic enrichment and hashing. IEEE Trans. Knowl Data Eng. 2015, 28, 566–579.

[4] Lopez-Gazpio, I.; Maritxalar, M.; Gonzalez-Agirre, A.; Rigau, G.; Uria, L.; Agirre, E. Interpretable semantic textual similarity: Finding and explaining differences between sentences. Knowl. Based Syst. 2017, 119, 186–199.

[5] Ramachandran, D.; Parvathi, R. Analysis of twitter specific preprocessing technique for tweets. Procedia Comput. Sci. 2019, 165, 245–251.

[6] Vo, D.-V.; Karnjana, J.; Huynh, V.-N. An integrated framework of learning and evidential reasoning for user profiling using short texts. Inf. Fusion 2021, 70, 27–42.

[7] Feng, W.; Zhang, C.; Zhang, W.; Han, J.; Wang, J.; Aggarwal, C.; Huang, J. STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream. In Proceedings of the 2015 IEEE 31st International Conference on Data `Engineering, Seoul, Korea, 13–17 April 2015; pp. 1561–1572.

[8] Ailem, M.; Role, F.; Nadif, M. Sparse poisson latent block model for document clustering. IEEE Trans. Knowl. Data Eng. 2017, 29, 1563–1576.

[9] Liang, S.; Yilmaz, E.; Kanoulas, E. Collaboratively tracking interests for user clustering in streams of short texts. IEEE Trans. Knowl. Data Eng. 2018, 31, 257–272.

[10] Carpineto, C.; Romano, G. Consensus clustering based on a new probabilistic rand index with application to subtopic retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 2012, 34, 2315–2326.

[11] Wang, T.; Brede, M.; Ianni, A.; Mentzakis, E. Detecting and characterizing eating-disorder communities on social media. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, UK, 6–10 February 2017; pp. 91–100.

[12] Song, G.; Ye, Y.; Du, X.; Huang, X.; Bie, S. Short text classification: A survey. J. Multimed. 2014, 9, 635.

[13] Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. Science 2014, 344, 1492–1496.

[14] Zhang, C.; Lei, D.; Yuan, Q.; Zhuang, H.; Kaplan, L.; Wang, S.; Han, J. GeoBurst+ Effective and Real-Time Local Event Detection in Geo-Tagged Tweet Streams. ACM Trans. Intell. Syst. Technol. (TIST) 2018, 9, 1–24.

[15] Yang, S.; Huang, G.; Xiang, Y.; Zhou, X.; Chi, C.-H. Modeling user preferences on spatiotemporal topics for point-of-interest recommendation. In Proceedings of the 2017 IEEE International Conference on Services Computing (SCC), Honolulu, HI, USA, 25–30 June 2017; pp. 204–211.

[16] Alsaffar, D.; Alfahhad, A.; Alqhtani, B.; Alamri, L.; Alansari, S.; Alqahtani, N.; Alboaneen, D.A. Machine and deep learning algorithms for Twitter spam detection. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Cairo, Egypt, 26–28 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 483–491.

[17] Shanmugam, S.; Padmanaban, I. A multi-criteria decision-making approach for selection of brand ambassadors using machine learning algorithm. In Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Uttar Pradesh, India, 28–29 January 2021; pp. 848–853.

[18] Jin, J.; Zhao, H.; Ji, P. Topic attention encoder: A self-supervised approach for short text clustering;SAGE, United Kingdom. J. Inf. Sci. 2022, 48, 701–717.