Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Machine Learning Based Automated Semantic Retrieval Method for Document Extraction

[1]Mrs . M. Divya, [2]Dr.S.Sukumaran

[1]Ph.D Research Scholar
Department of Computer Science
Erode Arts and Science College, Erode, Tamilnadu, India
[2] Associate Professor
Department of Computer Science
Erode Arts and Science College, Erode, Tamilnadu, India

Abstract: Retrieving information from the large amount of document is important in many contexts of the researchers who want to handle the retrieve on a research topic patents of the records related to a certain wish to retrieve the research papers as well as the common scenarios are the users in need of identifying relevant textual information in a text manuscript assortment. The files analysis is accomplished through interactive visualization besides the machine learning methods to a domain expert can regularly steer a system in to identifying the relevant files. Text type utilizing several equipment in records retrieval except of the machine getting to know has to be obtained an awful lot attention to each academy of the enterprise evolved researchers. On this paper, from the key phrases categorize the files utilizing automated semantic facts retrieval technique needs to be then finding the relevant documents making use of automated Semantic Retrieval method (ASRM). The proposed approach produce the excessive accuracy, f-measure, precision in addition to keep in mind parameter values helps to locating the relevant record be extracted without problems.
Keywords: Information Retrieval, Document Retrieval, Relevance Ranking, Machine Learning

1. Introduction

Statistics Retrieval (SR) is extensively used method in numerous domain names except the contexts that require searching, filtering, ranking or summarizing the facts from textural or multimedia sources. For times statistics retrieval can be utilized to assemble a search engine that returns the relevant web pages or images based totally on user queries. It is to be developing the recommender device that indicates merchandise, services or content primarily based on consumer alternatives besides the selection process. Information Retrieval may be applied to create a textual content summarization tool that generates concise in addition to informative summaries of long files or articles. This will be implemented to layout a sentiment analysis tool that detects the feelings otherwise evaluations expressed in social media posts or opinions [14]. Now a day's data retrieval plays a critical position in human day by day life thru its implementation in a number of sensible packages together with internet of the question in addition to answering structures, personal assistants, chatbots must be the digital libraries among others.

The number one objective of SR is to discover in addition to retrieve records that are applicable to a user's question. The primary purpose of SR is to become aware of the retrieve statistics that is related to a person's query. As a couple of records can be applicable of the outcomes are often ranked in line with their relevance score are regularly ranked [6]. The time period based totally retrieval systems ought to be several limitations including polysemy in addition to synonymy besides of the lexical gaps between the questions of the files. The advancements in computing electricity needs to be the provision of large labelled datasets have greatly impacted inside the area of herbal Language Processing (NLP) by using enabling researchers to make use of deep learning for a one-of-a-kind sort of innovations [15]. Those form of strategies were utilized to improve the traditional textual content retrieval structures as well as to overcome the limitations of time period-based totally retrieval structures.

The packages of net information retrieval is popular applications with the make use of the gaining knowledge of to rank the machine studying with records structures reduces the drawback of conventional rating systems which extended engagement of studying to rank algorithm in various fields. Stock Portfolio choice is used to Prediction the utility of the getting to know to rank algorithm provides the robustness to the device.

Message to automobile reply is quit to cease technique for producing particular sort of content material messages that is having reaction to selection besides of the response to set generator in addition to triggering models additives. Image textual content in addition to image ranking is employed a photo content description this is comparable pictures can be retrieved.

2. Related Works

Textual content indexing as well as information retrieval systems can be the index of the facts repositories as well as allow users to look the online get admission to to aware about the facts needs to be no want to fear approximately in which the facts must be saved. Customers can queried frequently that data must be decided the index with a unmarried seek. This associated works describes the list of records retrieval documents has to be the algorithms are desired by way of various authors in addition to extraordinary years. All the systems utilize the beyond historic statistics as well as with the help of machine mastering except of the ranking gadget loss mistakes that can be reduced. Evolved a inventory portfolio selection by using combining the functions of ListNet as well as RankNet algorithm with extra dependable predictions.

Lee H et.Al [8] proposed mastering to rank algorithms with three exclusive strategies allows to improve the rating of the particular files. The popular of the studies needs to be extensive on statistical techniques, might be studied to begin with. One in all its techniques is the vector space technique. The similarity between documents or key phrases is quantified in the area of records retrieval via appointing them into a vector of phrase frequencies in a vector space besides of determining the attitude among the two vectors.

A file provided by using EMC Sara Abdelghafaret.al., [12] imply that the quantity of textual information could be over 40 zettabytes by 2020 that is fifty times of the quantities in 2010. Imagine how meaningful statistics, styles, beneficial insights in addition to the course of information might be hidden inside massive amount of unstructured texts that might be extra gain for an expansion of domains from era, social sciences as well as education.

Mohammadi E, et.Al [11] describes a driving force to broaden techniques in addition to extract the structured significant information must be the knowledge from the huge unstructured raw information. Many vital phrases can be missing from a question, causing a seek engine or records retrieval system to respond badly or ineffectively, leading to outcomes which can be much less applicable to the question.

Yan E et.Al [12] discusses the several kinds of similarity metrics available in terms of semantic in addition to syntactic links that are applied in one-of-a-kind facts retrieval issues. Creator also cautioned a word embedding techniques which can be often utilized. It’s a technique built at the disseminated neural language best word2vec. A non-parametric approach, to retrieve related phrases to a query based at the framework.

Aretha B et.Al [1] speaks some research utilized in LDA to discover the shape of the medical fields. LDA become applied to identify subjects in addition to authors for a fixed of publications listed in the citeseer virtual library. Subject matter evaluation of papers posted inside the court cases of the country wide Academy of Sciences validated the potential of LDA to show the semantic subjects of instructional articles in diverse fields are blanketed. Sentiment analysis at the film evaluate that may be evaluation on the comments of the target market besides of the automatic occasion or crime detection.

3. Proposed methodology

Statistics Retrieval structures have the funds for the collections of thousands, or hundreds of thousands, of files, from which, by means of in case an applicable description, utilizers can improve any person of these statements. The index might also take exclusive bureaucracy, from storing key phrases with hyperlinks to individual documents, to clustering documents under related topics. Tremendous of the paintings in statistics retrieval can be automated.

3.1 Pre-Processing

Semantic extraction mentions to removing or dragging out precise data after the manuscript. Extraction a type includes, Keyword Extraction is the technique helps identify relevant terms has to be the expressions in the text as well as the deep insights when combined with the above classification techniques. The extracted terms for involuntary twitter classification grounded on the word type utilized in the tweets. Feature extraction is

the technique is utilized to identify the extraction of the entities in text, such as names of individuals, organizations, places. This method is typically helpful for customer support teams who intend to extracted relevant information from customer support tickets automatically, including customer name, phone number, query category, shipping details, etc.

Based totally on the key-word searching the report enlargement is a technique utilized in statistics retrieval (SR) structures to improve the overall performance of retrieval through increasing the illustration of every document. The impact late manuscript growth is that by counting additional related terms within the demonstration of every record, the IR gadget will be able to improved contest the question with the applicable documents.

Tokenization or lexical evaluation is the operation of creating phrases from a chain of letters (characters) in a file [1]. Typically, after parsing, lexical evaluation breaks down or tokenizes the document that considered as an input movement into phrases, phrases, or symbols [5]. One of the problems of tokenization is the conversion of acronyms in addition to the abbreviations into a standardized layout. The problem of tokenization varies relying at the language. Because maximum phrases in languages like English and other languages are separated by using white area, they may be mentioned be "space-bound" languages.

Stemming is a word-level amendment is stemming. The stemming factor's (or stemmer's) feature is to institution phrases that proportion a not unusual stem. The aim of this method is to cast off a couple of suffixes, limit the variety of words, and make sure that stems are exactly matched, store memory space in addition to time. For instance, all phrases gift, presentation, offered as well as awarding may be stemmed to the term gift.

The goal of this prevent phrases is to dispose of all of the useless further to insignificant commonplace terms from the tokens streams inclusive of articles, prepositions and so on. Examples of stop words are, "to", "in", "at", "a" and "the". Let's take this sentence as an example to apprehend how the works pass; The most obvious facts retrieval programs are search engines. Put off the classic prevent words, maximum obvious facts retrieval programs engines like google.

3.2 Feature Extraction

Keyword based totally searching allows to extract the characteristic Extraction approach of removing extraneous in addition to superfluous traits from a dataset is referred to as feature extraction. Whilst assigning textual content to one or more companies, accuracy is stepped forward via utilising function extraction strategies [8]. It facilitates the accuracy to get recovered, lessen dimensionality, in addition to has to decrease processing time [11]. The feature extraction algorithm depends on the vector space model, in which a sentence is represented as a dot in an N-dimensional space. The mathematical equation is:

$$Wi, j = t\,fi, j * log(|N|\, d\,fi) \quad \ldots\ldots\ldots (1)$$

The approach of representing files is the second one aspect that's normally referred to as indexing. Basically, it way that the machine creates a record index [2]. As an end result, discover the question representation technique.

In this segment, the consumer writes a question a good way to retrieve applicable information. After that, the gadget searches the index for documents which might be applicable as well as pertinent to the query has to be affords them to the user, this is call ranking. The remaining step is wherein the users can provide the hunt engine with applicable feedback [3]. Challenges the mismatch among how a user conveys the records they're searching for geared toward how the writer of the item expresses the information he is turning in is the primary undertaking in facts retrieval. In different words, the trouble is a mismatch between the user's vocabulary (language) and the writer's vocabulary (language). Except, there are boundaries to specifying the statistics a consumer calls for because of limitations within the person's functionality to explain what records is needed. Uncertainties of the ambiguities in languages are also one of the challenges that a utilizer can face.

The technique of representing files is the second aspect which is typically known as indexing. Basically, it approach that the device creates a document index. As a end result, discover the question illustration manner. In this segment, the user writes a query so that you can retrieve relevant statistics. After that, the system searches the index for files which are relevant in addition to pertinent to the question has to be provides them to the user, this is call ranking. The last step is in which the customers can offer the hunt engine with applicable comments [3]. Challenges the mismatch between how a user conveys the statistics they

may be seeking aimed at how the author of the item expresses the data he is handing over is the principle mission in facts retrieval. In different phrases, the hassle is a mismatch between the user's vocabulary (language) and the writer's vocabulary (language). Besides, there are barriers to specifying the data a consumer calls for due to boundaries inside the consumer's functionality to give an explanation for what records is required. Uncertainties of the ambiguities in languages also are one of the demanding situations.

The main factors of automated Semantic approach to processing the herbal language are:

Hyponyms: its miles a specific lexical entity having a relationship with a more typical verbal entity known as hypernym. Polysemy identifies the word having multiple meaning as well as it's far represented underneath one access. Meronomy is the arrangement of words except the textual content that demote a minor component of phrase file. Synonyms: it's far comparable which means phrases. An antonym is the phrases with contrary meaning. A homonym is the phrases with the identical spelling as well as Pronunciation, and a special that means altogether. The automated semantic method also utilizes semiotics in addition to collocations to comprehend as well as interpret language.

Algorithm of Automated Semantic Retrieval Method (ASRM)

Step 1: *INITIALIZE item array 'items[i]' with individuals of Product class present in GoodRelations Ontology. Remove prefix in URI of each individual*

Step 2: *Newstate ← 0*
for all Log message wj do

Step 3: *state ← 0, i ← 1*
while Forward (state, wji) > 0 do
Forward(state, wji) i ← i + 1
end while
for p = i → Length(wj) do
Newstate ← Newstate + 1

Step 4: *Send values of array elements in the Search box using "send_keys" method of webdriver.*

Step 5: *Invoke check procedure which will validate results of the searched query with queried item.*

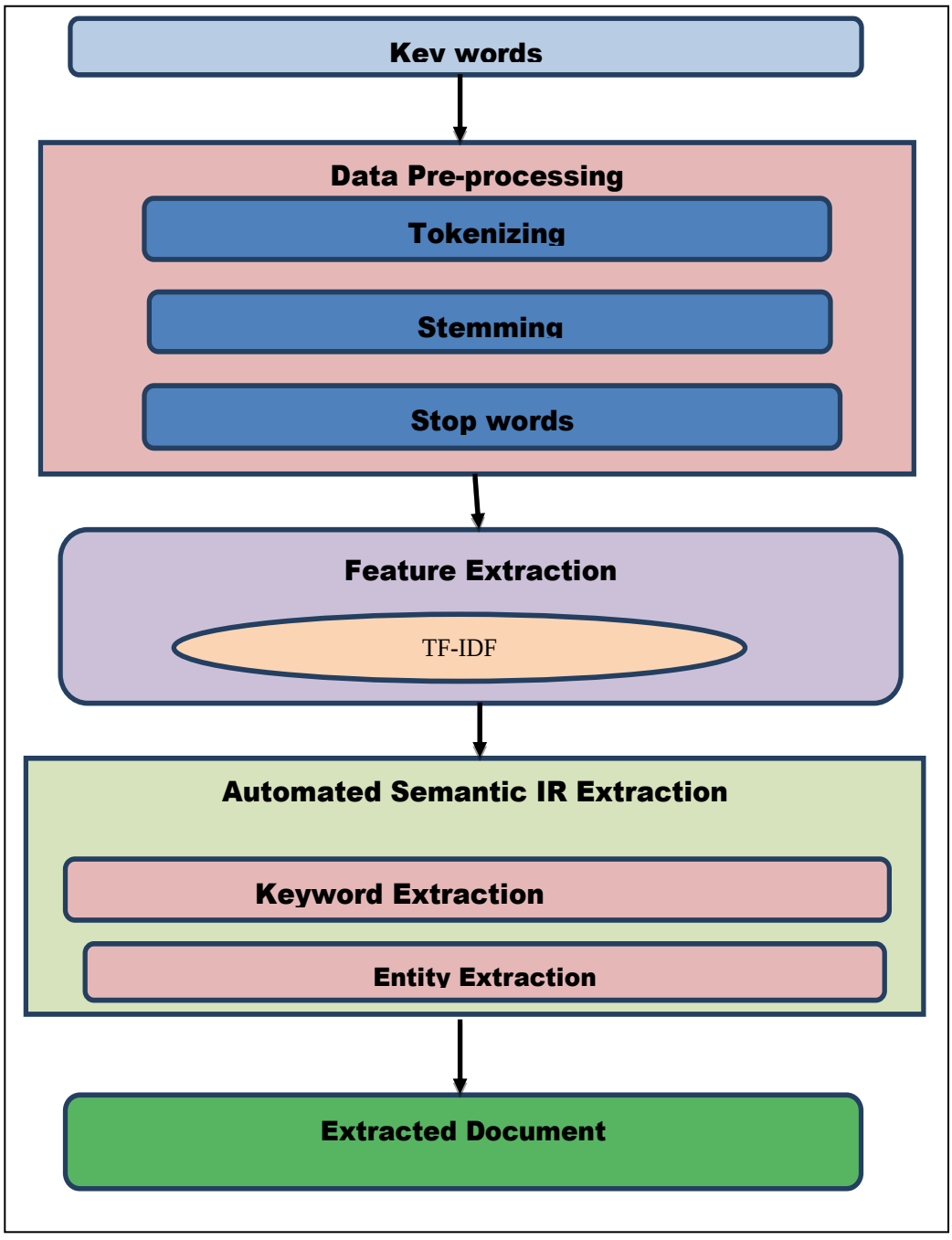
Step 6: *For each instance in the result set, repeat*

Step 7: *If instance matches queried item specifications then the test case is Successful/Passed else it will fail.*

3.3 Automated Semantic Retrieval Method (ASRM)

File parsing is the system of figuring out the content in addition to the shape of text documents which might be written must be supplied in a spread of languages, man or woman units, in addition to bureaucracy.Document parsing entails with identifying needs to be splitting the report structure into wonderful additives with the intention to shape it into unitary documents.

3.3.1 Stemming



Stemming is the development of conflating the opportunity varieties of a phrase into a ordinary symbol, the stem. The phrases: “awarding”, “awarded”, “awarding” may be decreased to a commonplace illustration “award”. That is a broadly applied method in text processing for information retrieval (IR) depends on the supposition that have an effect on a question thru the term providing shows attentiveness in files containing the words presentation and provided.

3.3.2 Stop Words

Stop words are fundamentally a position of frequently utilized vocabulary within every language, not just English language. Forestall phrases are fundamentally a role of regularly applied vocabulary inside every language, not simply English language. Stop terms are essentially a function of frequently utilized vocabulary inner every language, no longer simply English language. The inducement stop phrases is great to several

packages is with the purpose of, abolish the phrases in an effort to are extremely and typically applied in a recognized language, be capable of attention on the great phrases alternatively. Stop words are usually consideration to be a “single set of phrases”. It definitely is in a position to signify diverse matters to several applications. Determiners are susceptible to mark nouns anywhere a determiner frequently is probably observed with the aid of manner of a noun: the, a, an, every different. Coordinating conjunctions connect words, phrases, and clauses: for, an, nor, however, or, but, so. Prepositions specific temporal.

3.3.3 Feature Extraction-Term Frequency-IDF

Inverse record Frequency (TF-IDF) is a mathematical statistic which discloses that a phrase is in what manner big to a manuscript in a set. The TF- IDF is often exploited as a weighting factor in statistics restoration as well as text mining. The price of tf-idf develops proportionally in the direction of the numeral of length a phrase seems within the file, other than is counter appearing thru the prevalence of the phrase inside the corpus. This can facilitate to organize the facts with a view to some phrases are commonly applied files. TF-IDF can be efficiently applied for forestall-words filtering in a selection of issue fields counting text summarization as well as classification. TF-IDF is the object for consumption of two facts that are defined as a term primarily based frequency as well as inverse document frequency.

3.3.4 Automated Semantic IR Extraction Keyword

Extraction: Key-word extraction is the recovery of keywords or key terms after textual content documents. They are decided on amongst phrases inside the textual content document and characterise the record’s topic summarise the maximum commonly utilized methods that automatically extracted key phrases. Strategies that robotically extracted key phrases from the files utilize heuristics to select the maximum utilized as well as sizeable words or terms from the textual content report. The term keyword as an essential term to denote to language expressions that sales on one or extra of the following roles:

- **Terminology:** words or phrases that are applied in a precise area to indicate a specific technical idea.
- **Topics:** terms as well as labels that are portion of a set of concepts systematically accumulated below a specific classification policy.
- **Index terms:** terms representative major concepts, ideas, events, as well as people, mentioned to in a document or book.
- **Summary terms:** words or phrases that are expected to appear as a rapid explanation of the contented.
- ***Entity Extraction:*** Method entity extraction can be seemed as a sub-task of named entity recognition. Named-entity recognition (NER) (additionally called (named) entity identity, entity chunking, of the entity extraction) is a subtask of information extraction that factors to identify and classify participants of inflexible designators from facts proper to diverse types of named entities consisting of corporations, persons, locations. In the system of technique entity extraction, elements are described: corpus series as well as entity extraction. Consequently, to start with, the prevailing works obtained or constructed the statistics corpus for the technique entity reputation responsibilities, must be introduce a few generally utilized facts sources. By expending automated semantic retrieval approach (ASRM) examine the tools has to be concerned commercial enterprise stakeholders can enhance selection-making

4. Results And Discussions

The ability of an IR system to return pertinent documents, as well as the accuracy and precision of these retrieved documents, is commonly measured.

$$Precision = \frac{|relevant\ documents \cap retieved\ documents|}{|retrieved\ documents|} \quad \text{----- (2)}$$

The second measure is Recall. It is the proportion of documents that are related to the query and have been found.

$$Recall = \frac{|relevant\ documents \cap retrieved\ documents|}{|relevant\ document|} \quad \text{-------(3)}$$

These binary measures benefit to computeadditional information retrieval metrics which is F-measure

$$F - measure = \frac{2*precision*recall}{Precision+recall} \quad \text{------(4)}$$

Accuracy is used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition is

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{-----(5)}$$

where TP = True positive; FP = False positive; TN = True negative; FN = False negative

Table 4.1 gives the performance evaluation of ASRM are analysed by means of existing method like K-means, Support Vector Machine, Naïve Bayes, Roccchio are compared with proposed automated semantic information retrieval (ASRM).

Table 4.1: Performance Evaluation of ASRM

Methods	Precision (%)	Recall (%)	F-Measure (%)	Accuracy (%)
SVM	89.05	88.56	88.64	88.34
NBA	88.23	88.34	88.12	89.50
RA	89.29	89.67	88.45	89.76
ASRM	90.02	90.23	91.04	91.97

Searching academic publications for an evolving, undefined research area across disciplines presents a significant challenge for data collection [9].

From fig 4.1 explains the proposes the performance evaluation of automated semantic information retrieval method are analysed expending existing algorithm like K-nearest neighbour, support vector machine, naïve bayes algorithm,roccchio algorithms are compared with proposed algorithm as automated semantic retrieval method.

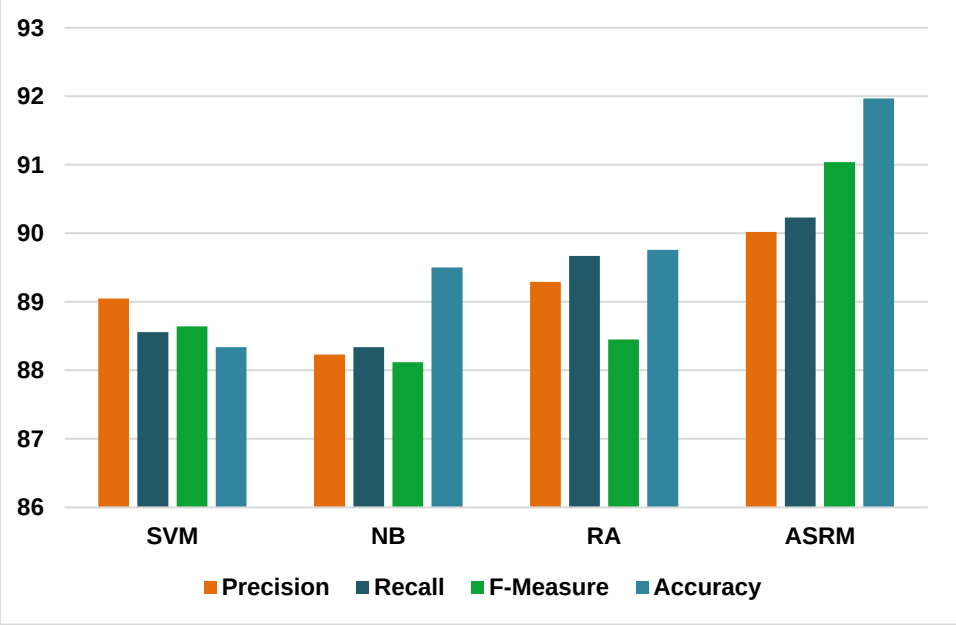


Fig 4.1: Performance Comparison Chart

The compared parameters are precision, recollect, f-measure values. Examine to the existing retrieval algorithms the proposed algorithm produce the excessive accuracy, f-measure, precision as well as recall values.

5. Conclusion

Learning to rank techniques is used for prediction responses with low errors price need to be locating the file to extract the green results. The work is compared with a number of the existed approaches which include SVM, Naïve Bayes, in addition to Rocchio Algorithms. Locating the relevant files is without problems utilising of keyword looking. The proposed work performed well for all the metrics considered. Hence ASRM is more suitable for effective document extraction.

References

[1] Aretha B Alencar, Maria Cristina F de Oliveira, and Fernando V Paulovich. 2012. Seeing beyond reading: a survey on visual text analytics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2, 6(2012), Pp: 476-492, Published Year 2012.

[2] Dr.S.Brindha, Dr.S.Sukumaran, Enhanced Pattern Classification Methods for Text Categorization Mining, International Journal of All Research Education and Scientific Methods (IJARESM), ISSN: 2455-6211 Volume 10, Issue 10, October-2022, Impact Factor: 7.429, Available online at: www.ijaresm.com.

[3] Chung, N. & Koo, C. (2015) 'The use of social media in travel information search.' Telematics and Informatics [Online]. 32 (2): 215–229, Published Year 2015.

[4] F. Beck, S. Koch, and D. Weiskopf. 2016. Visual Analysis and Dissemination of Scientific Literature Collections with SurVis. IEEE Transactions on Visualization and Computer Graphics 22, 1 (Jan 2016), Pp:180-189. https://doi.org/10.1109/TVCG.2015.2467757, Published Year 2016.

[5] Juhee Bae, ToveHelldin, Maria Riveiro, SławomirNowaczyk, Mohamed-RafikBouguelia, and GöranFalkman. 2020. Interactive Clustering: A Comprehensive Review. ACM Comput. Surv. 53, 1, Article 1 (Feb. 2020), 39 pages, Published Year 2020.

[6] Karami A, White CN, Ford K, et al. Unwanted advances in higher education:Uncovering sexual harassment experiences in academia with text mining. Inf Process Manag 2020; 57: 102167, Published Year 2020.

[7] Kowsari, K., JafariMeimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D.(2019) "Text classification algorithms: A survey." Information 10 (4): 150, Published Year 2019.

[8] Lee H and Kang P. Identifying core topics in technology and innovation management studies: a topic model approach. J TechnolTransf 2018; 43: Pp:1291–1317, Published Year 2018.

[9] Langlois, P., and Titah, R.(2020) "Utilisation et impact des outilsd'intelligenceartificielledans des contextes de cyberjustice." Doctoral dissertation, HEC Montréal, Published Year 2020.

[10] Liu, Y., Teichert, T., Rossi, M., Li, H. & Hu, F. (2017) 'Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews'. Tourism Management [Online]. 55: 554-563, Published Year 2017.

[11] Mohammadi E, Gregory KB, Thelwall M, et al. Which health and biomedical topics generate the most Facebook interest and the strongest citation relationships? Inf Process Manag 2020; 57: 102230, Published Year 2020.

[12] Sara Abdelghafar, Ashraf Darwish, and Aboul Ella Hassanien. 2020. Intelligent health monitoring systems for space missions based on data mining techniques. In Machine learning and data mining in aerospace technology.springer, Hershey, PA, USA, Pp:65-78, Published Year 2020.

[13] Yan E, Ding Y, Milojevic´ S, et al. Topics in dynamic research communities: an exploratory study for the field of information retrieval.J Informetr 2012; 6: 140–153, Published Year 2012.

[14] Yan E. Research dynamics: measuring the continuity and popularity of research topics. J Informetr 2014; 8: 98–110, Published Year 2014.

[15] Yan E. Research dynamics, impact, and dissemination: a topic-level analysis. J AssocInfSciTechnol 2015; 66: Pp:2357–2372, Published Year 2015.

Author Profile

Dr. S.Sukumaran, working as Associate Professor, Department of Computer science (Aided) in Erode Arts and Science College, Erode, Tamilnadu, India. He is a member of Board of studies in various Autonomous colleges and universities. In his 33 years of teaching experience, he has supervised more than 55 M.Phil research works, guided 22 Ph.D research works and still continuing. He has presented, published around 82 research papers in National, International Conferences and Peer Reviewed Journals. His area of research interest includes Digital Image Processing, Networking, and Data mining.