International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 9, Issue.11, November 2021 E-ISSN: 2347-2693

Fish Schooling Algorithm and Hash Based Indexing for Text Document Retrieval

Vinod Sharma

Department of Computer science Application and Engineering, SCE (M.P.), India

DOI: https://doi.org/10.26438/ijcse/v9i11.2428 | Available online at: www.ijcseonline.org

Received: 18/Nov/2021, Accepted: 20/Nov/2021, Published: 30/Nov/2021

Abstract— Publishers are getting content frequently as demand of publication increases day by day. To resolve an issue of identifying the research paper class as per content this work proposed a hybrid model. Features were select by the fish schooling genetic algorithm and indexing was provide by hash structure. In order to maintain the privacy of the user and server data model work on key based searching of relevant document. Each document has set of keywords and each keyword has its own unique key. So user query pass as set of unique keys and searching of cluster document was done by matching keys with hash index. Experiment was done on real dataset having set of document from different field of publication. Result shows that proposed model FSGA has increases the result outcome by fetching more relevant text documents as per user query.

Keywords— Clustering, Genetic Algorithm, Text Mining, Pattern Feature.

I. INTRODUCTION

In today's world, a massive amount of data is collected and stored in data warehouses. There is a distinction to be made between the data that is saved and the knowledge that we gain from it. The shift will not happen by itself, which is why the phrase "data mining" was coined [1]. Although some knowledge of data is required for data analysis, data mining can assist us in gaining a more comprehensive understanding of the data. The primary goal of data mining is to derive insights from acquired data. For short intervals, human effort

is employed for data processing, and for large data, it builds a variety of models that employ techniques such as artificial neural networks, feature reduction, and so on [2].

One of the most fundamental types of information retrieval is document retrieval based on an input query. This application is best exemplified via web searches [3]. For this aim, a number of algorithms have been created that take an input query and match it with stored documents or text samples, then rank the results based on their similarity score to the given query. These algorithms work by comparing indexed documents, which store information about phrase frequencies and locations, to individual query terms [4].

Each document is given a score based on its similarity value. The score of a query phrase in relation to a document is high if it appears frequently in that document. The goal of effective information retrieval should be to obtain only the information that is deemed relevant to a particular query [5].

II. RELATED WORK

The combination of cross-lingual and semantic search described by Zhitao Guan et al in [6] remains an open topic for searchable encryption. No previous research has looked into the problem of cross-lingual ranked search over encrypted cloud data to our knowledge. We propose a cross-lingual multi-keyword rank search (CLRSE) scheme based on the Open Multilingual Wordnet to overcome this problem.

In [7], Jeong, Soyeong, and colleagues introduced an Unsupervised Document Expansion with Generation (UDEG) framework that uses a pre-trained language model to produce a variety of supplemental phrases for the original document without the use of labels on query-document pairings for training. We stochastically tweak their embeddings to generate more diverse sentences for document expansion when producing sentences.

The system architecture, data distribution technique, and retrieval system we constructed are described in [8] by H. Chiranjeevi et al. For effective retrieval and indexing of data for crawling, a convolutional neural network (CNN) is used to classify text documents. An API-based micro-service architecture is used to disseminate and retrieve information depending on the identifying key. The system provides a platform for extracting knowledge and channelling data for use by the company, as well as allowing support centres to provide on-demand services.

In [9], Soyeong Jeong et al. introduced an Unsupervised Document Expansion with Generation (UDEG) framework that uses a pre-trained language model to produce a variety of supplemental phrases for the original document without the use of labels on query-document pairings for training.

We stochastically tweak their embeddings to generate more diverse sentences for document expansion when producing sentences.

In [10], Tuyen Thi-Thanh et al. established a strategy for retrieving related texts to a query using a semantic information retrieval model for Vietnamese. The semantic analysis in the proposed system identifies the semantic dependency graph of phrases, and the retrieving process computes the text document's significance using these semantic dependency graphs. The transformation rules are explored to apply on dependency parse utilising lexicon ontology for Vietnamese in order to determine the semantic dependency graph of a phrase. The Jaccard-Tanimoto distance is applied to the ranking function for rating retrieval results.

III. PROPOSED METHODOLOGY

This project is focused on document organization in order to improve document retrieval. The dataset was clustered using numeric term attributes that were unique to each term. To provide anonymity for the term, each of them is assigned a unique number, and each document has a set of keywords that are all the same length. For the grouping of the texts presented in [11], the Fish Schooling Genetic Algorithm FSGA was applied.

Fish Schooling Algorithm

For document retrieval without any structural input to the dataset, this strategy was proposed in [11] (our prior model). Because the model is unsupervised, no prior knowledge of the document type is necessary. The documents were clustered using the Fish Schooling Genetic Algorithm (FSGA). To calculate the distance between documents, the proposed model uses a pattern characteristic from the content.

Document Clustering Module

In this paper, document retrieval was offered as a way to quickly retrieve documents based on a query. Term characteristics were used to do hash-based indexing of the dataset document. To ensure that the phrases are kept private, each one is assigned a unique number, and each document has its own hash index key. In all different evaluation parameters, the proposed work has increased the retrieval efficiency of the work. As a result, hash-based indexing ensures document retrieval privacy while still being efficient.

The entire project is divided into two sections:

1. Training module
2. Testing module

In the training module, a hash index was proposed, however in the testing module, a user pass query was recommended.

ASSIGN TERM ID

Assign a number to each term in the various documents. As a result, a dictionary of words with their numbers is constructed, with each text being assigned a unique number. Words from various documents that are already in the dictionary are not updated in this case. As a result, those terms that aren't in the dictionary are added to it with a unique term.

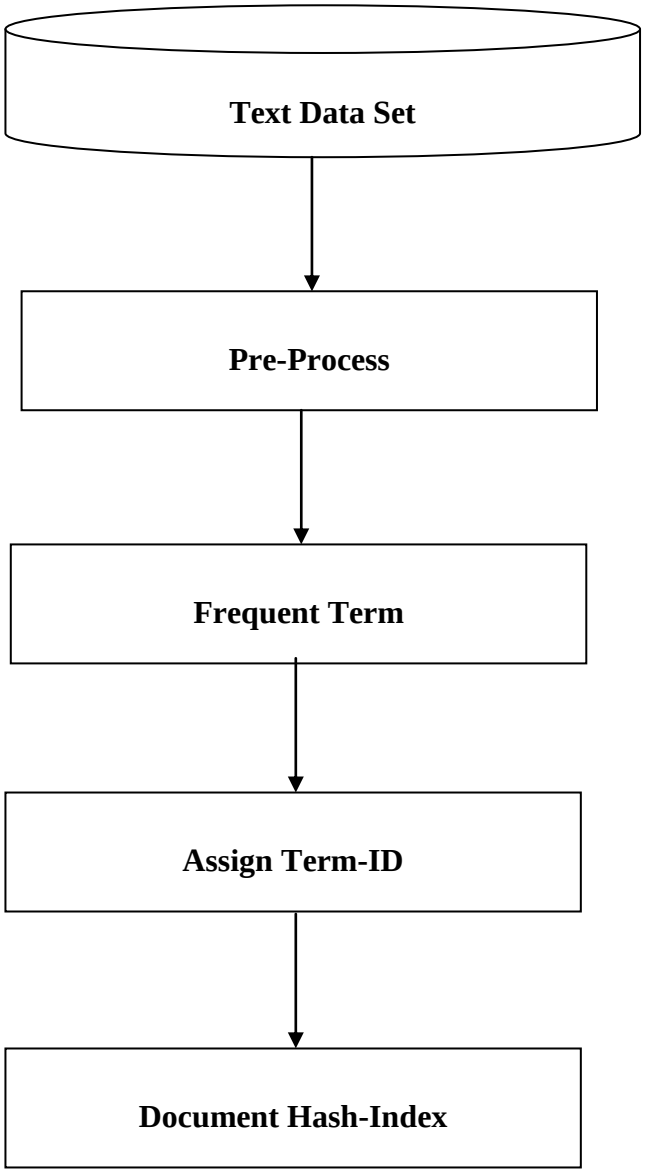

Fig. 1 Block diagram of training module

DOCUMENT HASH-INDEX

This stage determines the document index based on the terms extracted from the document. All of the terms in the paper are arranged in decreasing order based on their frequency values.

TESTING MODULE

A user query is received as input, and these words are then entered into the testing modal, which finds relevant documents based on the query. So, let's see if the user text query Q=State Government of India works. Stop words are deleted from the query after pre-processing.

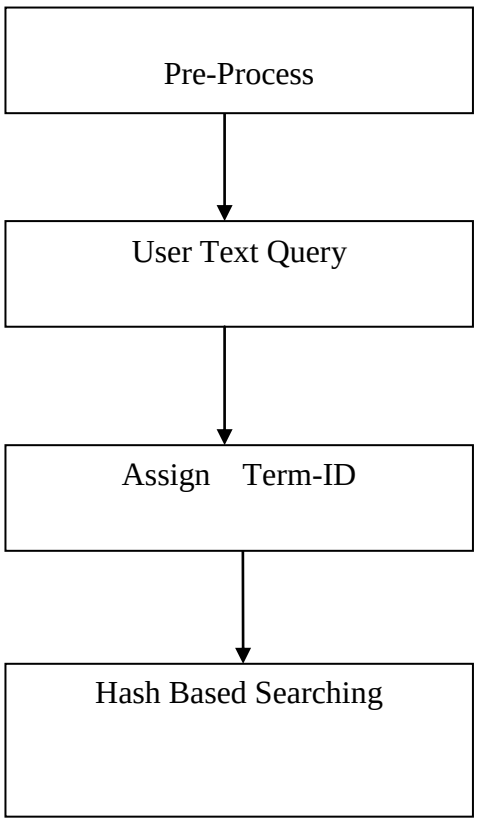

Fig. 2 Block diagram of proposed Searching Model

HASH BASED SEARCHING

The keywords (terms) from the user text query have their own term id in this phase, while the number is the same as in the dataset. The user query's id privacy is increased as a result of this word. Now, all term-ids in the text query are used as keys in the hash function, which retrieves each document set from the matching index. Apply intersection to those two sets now. As a result, the most prevalent elements are fetch and comparable documents.

Let term id be the key, and the modulus function M be used to index the hash base. As a result, the function's output is Y, which is the insert key's index position.

$$M(X) \rightarrow Y \quad \text{------// Hash function}$$
$$M(X) = ||X, C||$$

where C is fix constant use for finding the index position of the key.

The entire step of pre-processing and assigning a term-Id is the same in this searching model as it was in previous steps, though the terms received after pre-processing are not filtered according to their frequency in the query. As a result, hash-based searching and retrieval of linked documents is a novel search paradigm. Keywords are terms that are derived from a user text query and are used to persuade others.

Time complexity of above algorithm is O(nm+u+n) where n is number of documents, m is number of words and u is unique words obtained form n number of documents.

IV. EXPERIMENTS & RESULTS ANALYSIS

Because of the large number of inbuilt functions in MATLAB software, such as text-scan to separate strings into words, reading and writing of text files, word comparison, word collection into structure, and so on, the proposed genetic algorithm based document clustering approach model was implemented.

Dataset

Testing dataset was taken from [12], has various research paper from three field "Electrical, Computer, Electronics". In order to compare this document retrieval method clustered document from each method UFCGA [13], FFDC [14] and FSGA were pass in hash based index module which can produce fetch the documents. So relevancy of document as per user query is compared by NDCG parameter as well.

Testing Query Set	
Query1	'data mining computer privacy'
Query2	'digital image data fetching'
Query3	'Solar power plant'
Query4	'solar wind power load balance'

Result

Table 1 NDCG Parameter Based Comparison.

User Query	UFCGA[33]	FFDC[14]	ProposedFSGA
Query1	0.4628	0.7718	0.7925
Query2	0.5718	0.6152	0.6317
Query3	0.4455	0.5705	0.9088
Query4	0.7277	0.8795	0.9032

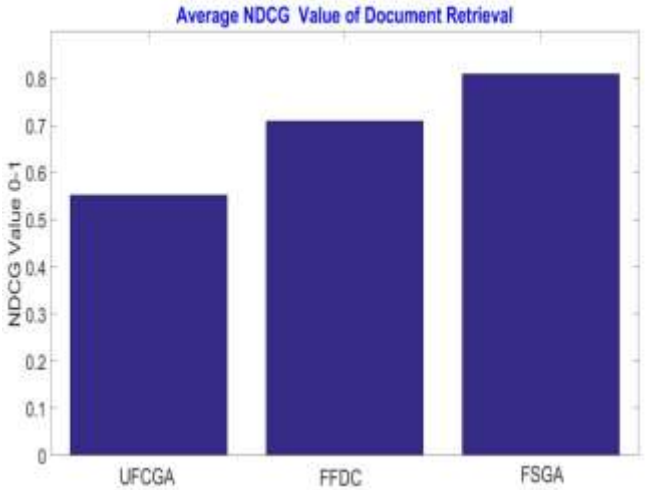

Fig. 3 Average NDCG Parameter Based Comparison.

Above table 1 and fig. 3 shows that proposed work FSGA has improved the NDCG evaluation parameters values as compared to previous work UFCGA [33] and FFDC . Hash based indexing has increase the relevancy of desired document as per input user query. As Fish based genetic algorithm has cluster document in high relevant manner, so NDCG value of this is also high as compared to fire fly in FFDC and normal genetic algorithm in UFCGA.

Table 2 Precision Parameter Based Comparison.

User Query	UFCGA[33]	FFDC[14]	ProposedFSGA
Query1	0.4167	0.7083	0.7083
Query2	0.5417	0.7083	0.7083
Query3	0.4583	0.7083	0.875
Query4	0.625	0.875	0.875

In comparison to prior work UFCGA [33] and FFDC, table 2 demonstrates that the suggested work FSGA has increased the Precision evaluation parameters values. Hash-based indexing has improved the relevancy of the desired document in response to the user's query. Because the Fish-based genetic algorithm clusters documents in a highly relevant manner, its Precision value is also higher than that of the fire fly in FFDC and the regular genetic algorithm in UFCGA.

Table 3 Recall Parameter Based Comparison.

User Query	UFCGA[33]	FFDC[14]	ProposedFSGA
Query1	0.4348	0.7391	0.7391
Query2	0.5652	0.7391	0.7391
Query3	0.4783	0.7391	0.913
Query4	0.6522	0.913	0.913

In comparison to prior work UFCGA [33] and FFDC, table 3 shows that the suggested work FSGA has improved the Recall evaluation parameters values. Hash-based indexing has improved the relevancy of the desired document in response to the user's query. Because the Fish based genetic algorithm clusters documents in a highly relevant manner, it has a higher recall value than the fire fly in FFDC and the regular genetic algorithm in UFCGA.

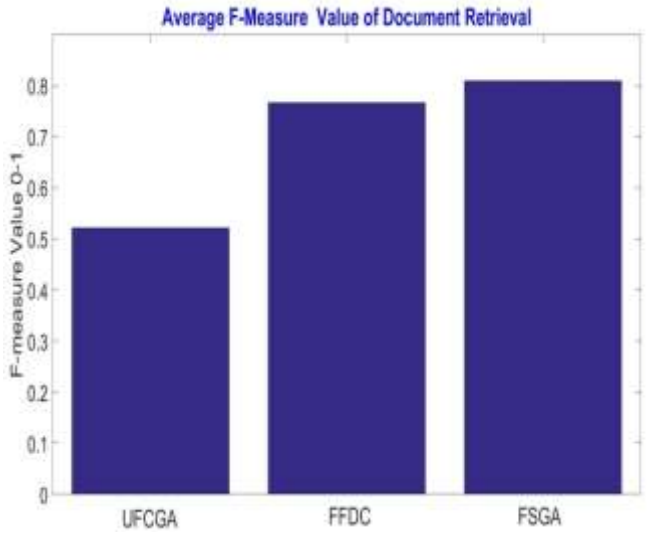

Fig. 4 Average F-Measure Parameter Based Comparison.

Table 4 F-Measure Parameter Based Comparison.

User Query	UFCGA[33]	FFDC[14]	ProposedFSGA
Query1	0.4255	0.7234	0.7234
Query2	0.5532	0.7234	0.7234
Query3	0.4681	0.7234	0.8936
Query4	0.6383	0.8936	0.8936

In comparison to prior work UFCGA [33] and FFDC, the proposed work FSGA has enhanced the F-Measure assessment parameters values (see table 4 and fig. 4). Hash-based indexing has improved the relevancy of the desired document in response to the user's query. Fish-based genetic algorithm has a high F-Measure value when compared to fire fly in FFDC and regular genetic algorithm in UFCGA because it clusters documents in a highly relevant manner.

Table 5 Execution time (Seconds) Parameter Based Comparison.

User Query	UFCGA[33]	FFDC[14]	ProposedFSGA
Query1	0.03	0.0297	0.0218
Query2	0.0327	0.0279	0.0231
Query3	0.0335	0.0288	0.0222
Query4	0.0324	0.031	0.0237

In comparison to prior work UFCGA [33] and FFDC, table 5 demonstrates that the suggested work FSGA has reduced the fetching time evaluation parameters values. Hash-based indexing has improved the relevancy of the desired document in response to the user's query. Because the Fish based genetic algorithm clusters documents in a highly relevant manner, it takes less time than the fire fly in FFDC and the regular genetic algorithm in UFCGA.

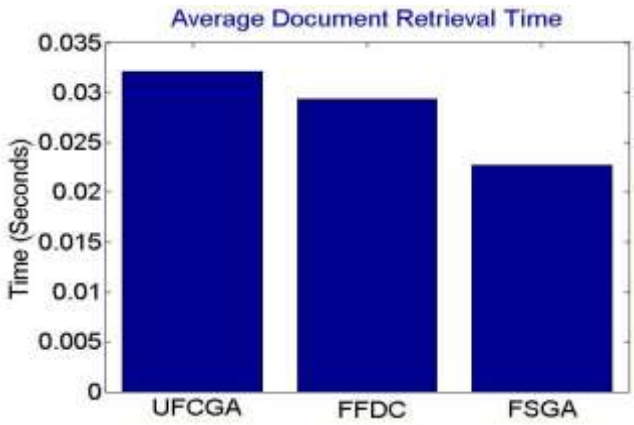

Fig. 5 Average retrieval Time Parameter Based Comparison.

Table 6 Accuracy Parameter Based Comparison.

User Query	UFCGA[33]	FFDC[14]	ProposedFSGA
Query1	46	63.89	63.89
Query2	52.27	63.89	63.89
Query3	47.92	63.89	82.14
Query4	57.50	82.14	82.14

In comparison to prior work UFCGA [33] and FFDC, the proposed work FSGA has enhanced the accuracy evaluation parameters values (see table 6 and fig. 6). Hash-based indexing has improved the relevancy of the desired document in response to the user's query. Because the Fish based genetic algorithm clusters documents in a very relevant manner, it has a higher accuracy value than the fire fly in FFDC and the regular genetic algorithm in UFCGA.

V. CONCLUSIONS

Getting an relevant information from any search depends on structure of data store. Use of genetic algorithm for feature reduction and clustering increase the relevancy chances in the work. Hash based unique term indexing has increases the privacy of model where user not disclose the query and other document information at any level of search algorithm. Work to be done In comparison to prior work by UFCGA [33] and FFDC, FSGA has enhanced the accuracy evaluation parameter values. Hash based indexing has increase the relevancy of desired document as per input user query. As Fish based genetic algorithm has cluster document in high relevant manner, so accuracy value of this is also high as compared to fire fly in FFDC and normal genetic algorithm in UFCGA.

REFERENCES

[1]. Abhishek Jain, Aman Jain, Nihal Chauhan, Vikrant Singh and Narina Thakur. "Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model". International Journal of Computer Applications **164(6):28-30, April 2017.**

[2]. Dr.M.Suresh Babu, Mr. A.Althaf Ali, Mr. A.Subramaneswara Rao, "A Study on Information Retrieval Methods in Text Mining", International Journal Of Engineering Research & Technology (Ijert) Ncdma, **Volume 2 – Issue 15, 2014**

[3]. P, Mrs. (2020). A Prognostic Rainfall using Machine Learning Technique. International Journal for Research in Applied Science and Engineering Technology. **8: 1 2020.**

[4]. Wu Chuhan, et al. A hybrid unsupervised method for aspect term and opinion target extraction Knowledge-Based Systems, **148, 2018.**

[5]. Giannakopoulos, Athanasios "Unsupervised aspect term extraction with b-lstm & crf using automatically labelled datasets." Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. **2017.**

[6]. Zhitao Guan, Xueyan Liu, Longfei Wu, Jun Wu, Ruzhi Xu, Jinhu Zhang, Yuanzhang Li, Cross-lingual multi-keyword rank search with semantic extension over encrypted data, Information Sciences, **Volume 514, 2020.**

[7]. Jeong, Soyeong and Baek, Jinheon and Park, ChaeHun and Park, Jong. "Unsupervised Document Expansion for Information Retrieval with Stochastic Text Generation". Proceedings of the Second Workshop on Scholarly Document Processing, **2021.**

[8]. H. Chiranjeevi and K. S. Manjula, "An Text Document Retrieval System for University Support Service on a High Performance Distributed Information System," 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), **2019.**

[9]. Soyeong Jeong, Jinheon Baek, ChaeHun Park, Jong C. Park. "Unsupervised Document Expansion for Information Retrieval with Stochastic Text Generation". Proceedings of the Second Workshop on Scholarly Document Processing, **pages 7–17 June 10, 2021.**

[10]. Tuyen Thi-Thanh Do; Dang Tuan Nguyen. "A computational semantic information retrieval model for Vietnamese texts" International Journal of Computational Science and Engineering **Vol.24 No.3., 2021**

[11]. https://ijsret.com/2017/12/14/computer-science/

[12]. Alan Díaz-Manríquez , Ana Bertha Ríos-Alvarado, José Hugo Barrón-Zambrano, Tania Yukary Guerrero-Melendez, And Juan Carlos Elizondo-Leal. "An Automatic Document Classifier System Based on Genetic Algorithm and Taxonomy". accepted March 9, 2018, date of publication March 15, 2018, date of current version **May 9, 2018.**

[13]. Vinod Sharm, "Document Class Identification Using Fire-Fly Genetic Algorithm and Normalized Text Features" Volume 6 Issue 1, ijsret.com.

[14]. Vinod Sharma, Dr. Shiv Shakti Shrivastava and Dr. Sanjeev Kumar Gupta."Fish Schooling Genetic Algorithm for Text document Clustering Using Pattern Features". International Journal of Grid and Distributed Computing (I.J.G.D.C.) **Vol. 13, No. 1, 2020.**

