IMPROVED MULTI-VERSE OPTIMIZER IN TEXT DOCUMENT CLUSTERING FOR TOPIC EXTRACTION

AMMAR KAMAL MOUSA ABASI

UNIVERSITI SAINS MALAYSIA

2021

IMPROVED MULTI-VERSE OPTIMIZER IN TEXT DOCUMENT CLUSTERING FOR TOPIC EXTRACTION

by

AMMAR KAMAL MOUSA ABASI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

February 2021

ACKNOWLEDGEMENT

First and foremost, I would like to sincerely thank my supervisor Professor Dr.Ahamad Tajudin Khader from the School of Computer Sciences at Universiti Sains Malaysia for his guidance, understanding, patience and most importantly, he has provided positive encouragement and a warm spirit to finish this thesis. It has been a great pleasure and honour to have him as my supervisor. I offer my special thanks to my co-supervisors, Associated Professor Dr.Mohammed Al-Betar and Dr.Nur Syibrah Binti Muhamad Naim for their encouragement, insightful comments, and hard questions.

My deepest gratitude goes to all of my family members. It would not be possible to write this thesis without support from them. I would like to thank my dearest father Kamal Abasi, my mother Nariman, my brothers and sisters.

Last but not least, I would sincerely like to thank all my beloved friends who were with me and support me through thick and thin. Most importantly, I would like to thank Dr.Osama Al Omari, Dr. Zaid Alyasseri, and Dr.Sharif Makhadmeh.

TABLE OF CONTENTS

CHAPTER 9 CONCLUSION AND FUTURE WORK

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ABC Artificial Bee Colony

ADDC Avarage Distance To The Cluster Centroid

AIG The Algorithm Of The Innovative Gunner

ALO Ant Lion Optimizer

AMGM Modified Arithmetic Mean Geometric Mean

ANNs Artificial Neural Networks

BA The Bat Algorithm

BBO Biogeography-based Optimization

BMVO Binary Mvo

CCOA Cultural Coyote Optimization Algorithm

CHI Chi-square

CMO Center Of Mass Optimization

CMVO Chaotic Multi-verse Optimization Algorithm

COA Coyote Optimization Algorithm

CPD Conference Papers Datasets

CS Cuckoo Search

CSI Co-occurrence Statistical Information-based Keyword Extraction

CSKH Cuckoo Search And Krill Herd

CSP Chaotic Swarming Of Particles

CSTR Centre For Speech Technology Research

DA Dragonfly Algorithm

DE Differential Evolution

DE Dolphin Echolocation

DEKH Differential Evolution Kh

DF Document Frequency

DR Dimension Reduction

EA Evolutionary-based Algorithms

EPSO Environment Factor-inspired Pso

ES Evolution Strategies

FFOA Fruit Fly Optimization Algorithm

FS Feature Selection

GA Genetic Algorithms

GF-CLUST Gravity Firefly Clustering

GP Genetic Programming

GSO Glowworm Swarm Optimization

GWO Grey Wolf Optimizer

HDE Hybrid Differential Evolution

H-FSPSOTC Hybrid Pso Algorithm With The Gos

HLBMVO Hybrid Link-based Mvo

HS Harmony Search

IFA Improved Firefly Algorithm

IG Information Gain

kCC K-means Based Co-clustering

KHA Krill Herd Algorithm

KU Krill Updating

LABIC Laboratory Of Computational Intelligence

LBMVO Link-based Mvo

LBPSO Link Based Particle Swarm Optimization

LFMVO Levy-flight Multi-verse Optimization

LGSI Ludo Game-based Swarm Intelligence

LSI Latent Semantic Indexing

MAD Mean Absolute Difference

MFO Moth-flame Optimisation

MI Mutual Information

MM Mean-median

MVO Multi-verse Optimizer

NB Naive Bayes

NFL No Free Lunch

NIPS Neural Information Processing Systems

NLP Natural Language Processing

NoC Network On Chip

NP Noun Phrase

NSS Neighborhood Selection Strategy

PCA Principal Component Analysis

PSO Particle Swarm Optimization

QS Queuing Search

RAKE Rapid Automatic Keyword Extraction Algorithm

ROA Ray Optimization Algorithm

SAD Scientific Articles Datasets

SFO Sailfish Optimizer

SGA The Search Group Algorithm

SHO Spotted Hyena Optimizer

SI Swarm Intelligence

SOM Self-organizing Maps

SSA Salp Swarm Algorithm

SSC Stud Selection And Crossover

SVM Support Vector Machine

TAs Trajectory-based Algorithms

TCD Text Clustering Datasets

TD Text Document

TDC Text Documents Clustering

TDCP Text Documents Clustering Problem

TE	Topic Extraction
TF	The Most Frequent Based Keyword Extraction
Tf-Idf	Term Frequency Inverse Document Frequency
TLBO	Teaching-learning-based Optimization
TR	Textrank
TV	Term Variance
UAVs	Unmanned Aerial Vehicles
UCI	Machine Learning Repository
VSM	Vector Space Model
WEP	Probability Of The Wormhole Existence
WWO	Water Wave Optimization

PENAMBAHBAIKAN PENGOPTIMUMAN PELBAGAI ALAM DALAM PENGELOMPOKAN DOKUMEN TEKS UNTUK PENGEKSTRAKAN TOPIK

ABSTRAK

Dalam dunia digital, dokumen teks besar membanjiri web setiap hari. Pengekstrakan topik (TE) adalah asas penting untuk penerangan kandungan melalui mis. label atau ringkasan. Akibatnya, memanipulasi dokumen teks ini untuk memilih topik tidak dapat dilaksanakan secara manual. Mengenal pasti topik secara automatik boleh menjadi alternatif yang sangat baik untuk merumuskan topik secara manual. Teks Penggabungan Dokumen (TDC) mewakili, secara umum, langkah pertama TE untuk mengenal pasti dokumen, yang membahas masalah yang berkaitan. Kajian ini bertujuan untuk mencadangkan pendekatan TE yang sesuai, yang memberikan gambaran keseluruhan dokumen teks yang lebih baik. Untuk mencapai tujuan ini: Pertama, Kaedah pemilihan ciri baru untuk TDC, iaitu algoritma pengoptimum berbilang ayat binari (BMVO) dicadangkan untuk menghilangkan ciri-ciri yang tidak relevan, berlebihan dan memperoleh subkumpulan baru yang lebih bermaklumat. Kedua, tiga algoritma pengoptimum berbilang ayat (MVO), iaitu MVO asas, MVO diubah, MVO hibrid dicadangkan untuk menyelesaikan masalah TDC; algoritma ini adalah peningkatan tambahan dari versi sebelumnya. Ketiga, kaedah ensemble novel untuk TE automatik dari koleksi dokumen teks dicadangkan untuk mengekstrak topik dari dokumen berkelompok. Untuk menilai kaedah yang dicadangkan untuk TDC, enam langkah luaran (iaitu ketepatan, ketepatan dan penarikan balik, ukuran-F, kemurnian, dan entropi) digunakan. Selanjutnya, enam belas set data, termasuk enam set data teks standard dan sepuluh set data penerbitan ilmiah digunakan dalam eksperimen tersebut. Hasil yang dihasilkan oleh algoritma

yang dicadangkan untuk TDC dibandingkan dengan kaedah pertimbangan yang baik, termasuk kaedah pengelompokan dan kaedah berasaskan metaheuristik. Mengejutkan. Kaedah yang dicadangkan dapat unggul pada semua kaedah perbandingan dalam semua set data yang digunakan menggunakan hampir semua ukuran luaran. Selanjutnya, untuk menilai kaedah TE ensembel yang dicadangkan, tiga langkah luaran (iaitu ketepatan, penarikan, dan ukuran-F) digunakan. Sekali lagi, sepuluh set data penerbitan ilmiah yang sama juga digunakan dalam eksperimen. Hasil yang dihasilkan oleh kaedah TE ensembel yang dicadangkan dibandingkan dengan yang dihasilkan oleh lima kaedah statistik yang ditetapkan dalam literatur. Eksperimen menunjukkan hasil yang menjanjikan bahawa kaedah TE ensembel yang dicadangkan dapat mencapai rata-rata 49.29 %, 45.22 %, dan 46.90 % masing-masing dengan ketepatan, penarikan, dan ukuran-F Oleh itu, kaedah TE ensembel yang dicadangkan mampu mengungguli semua kaedah perbandingan menggunakan keseluruhan pengukuran luaran untuk hampir semua set data.

IMPROVED MULTI-VERSE OPTIMIZER IN TEXT DOCUMENT CLUSTERING FOR TOPIC EXTRACTION

ABSTRACT

In the digital world, large text documents are inundating the web every day. The topic extraction (TE) is an important basis for description of the contents through e.g. labels or summaries. Consequently, manipulating these text documents to selecting topics is not feasible manually. Automatically identifying topics can then be an excellent alternative to manually formulating topics. Text Document Clustering (TDC) represents, in general, the first step of TE to identify the documents, which address a related subject matter. This study aims to propose a suitable TE approach, which provides a better overview of the text documents. To achieve this aim: First, A new feature selection method for TDC, that is, binary multi-verse optimizer algorithm (BMVO) is proposed to eliminate irrelevantly, redundant features and obtain a new subset of more informative features. Second, three multi-verse optimizer algorithm (MVOs), namely, basic MVO, modified MVO, hybrid MVO is proposed to solve the TDC problem; these algorithms are incremental improvements of the preceding versions. Third, a novel ensemble method for an automatic TE from a collection of text document is proposed to extract the topics from the clustered documents. To evaluate the proposed methods for TDC, six external measures (i.e., accuracy, precision and recall, F-measure, purity, and entropy) are used. Furthermore, sixteen datasets, including six standard text datasets and ten scientific publications datasets are used in the experiments. The results produced by the proposed algorithms for TDC are compared with well-regard methods, including clustering methods and metaheuristic-based methods. Surprisingly. The proposed method can excel at all comparative methods in all datasets used using almost all

external measurements. Furthermore, to evaluate the proposed ensembled TE method, three external measures (i.e., precision, recall, and F-measure) are used. Again, the same ten scientific publications datasets are also used in the experiments. The results produced by the proposed ensembled TE method are compared with those produced by five statistical methods established in the literature. The experiments showed promising results that the proposed ensembled TE method can achieve an average 49.29%, 45.22%, and 46.90% by precision, recall, and F-measure, respectively. Accordingly, the proposed ensembled TE method is able to outperform all comparative methods using the entire external measurements for all almost all datasets.

CHAPTER 1

INTRODUCTION

1.1 Background

In such a vast digital-driven age and owing to gigantic technological advances, the Internet development and advanced online technologies including hugely powerful data servers and voluminous information all constitute an issue that we daily encounter. The International Data Corporation (IDC) has released a report, which anticipates 175 zettabytes of data worldwide by 2025 [1]. Such voluminous data are accumulating in mainframes, servers, and public cloud environments. A significant amount of such enormous data is represented in text format. Various text mining applications were introduced in the existing literature. These applications involve the enhancement of the query results that are returned by search engines, unsupervised text organization systems, knowledge discovery processes, as well as information retrieval services, in addition to text mining processes (Emrouznejad & Yang, 2018). Also, many approaches were proposed with the aim of organizing unsupervised text documents for efficient use (S. F. Hussain & Haris, 2019).

Topic extraction (TE) can be useful for many real-world applications (Y. Zhang et al., 2016). For example, by examining recent publications in computer science domains, areas that are becoming increasingly important can be identified and their trends and popularity can be further predicted in the foreseeable future. In addition, they as a fundamental problem of information retrieval can help the decision mak-

[1] https://www.idc.com/

ers to efficiently detect meaningful topics. Therefore, it has attracted much attention such as public opinion monitoring, decision supporting and emergency management (Mottaghinia, Feizi-Derakhshi, Farzinvash, & Salehpour, 2020).

However, there is a lot of uncertainty regarding how to define these topics and an ongoing debate about how to automatically extract them. In addition, extracting these topics using manual methods is slow, expensive and not accurate (Shaikh, 2018). One of the most commonly used techniques to identify topics is to cluster documents to determine certain groups of papers representing a related subject matter (S. Wang & Koopman, 2017). After that, the most relevant terms from each cluster are extracted and ranked.

Text documents clustering (TDC) is one of the powerful and efficient unsupervised learning technique in text mining. it represents, in general, the first step of TE to identify the documents, which address a related subject matter. TDC aims to divide documents into groups (also called clusters), where similar documents are placed in the same cluster and dissimilar documents in different clusters. This technique helps construct meaningful partitions of massive amounts of heterogeneous digital documents. Partition text documents clustering (PTDC) is defined by Bouras and Tsogkas (2012) as follows, " the process of partitioning a collection of documents into several subcollections based on their similarity of contents ". TDC has been widely studied in the last few years due to two main reasons. First, it is difficult to assign the related documents manually to extract meaningful information. Second, to avoid personal biases in judging documents that belong to any filed or category (Shafiabady et al., 2016). According to (Cagnina, Errecalde, Ingaramo, & Rosso, 2014), even human experts in

automatic clustering do not intervene, which entails that there is no need for any prior knowledge about the texts (i.e., without consulting class label of the documents).

Generally, each document in TDC is represented as a vector using the vector space model (VSM). A widely used approach for document representation is a bag of words (Salton, Wong, & Yang, 1975), where each distinct term that is present in the documents' collection is considered as a feature for the documents' representation. Therefore, a document is represented by a multi-dimensional feature space, where the cell value of each dimension corresponds to a weighted value, e.g., term frequency inverse document frequency (Tf-Idf) (Salton & Buckley, 1988), of the concerned term within the document. Hundreds of thousands of informative and uninformative features (i.e., irrelevant, redundant, and noisy features) originate from the transformation process (K. K. Bharti & Singh, 2015).

High-dimensional feature space of VSM is one of the most important challenges in text clustering because it increases the computational time while decreasing the efficiency of clustering performance (Chandrashekar & Sahin, 2014). Therefore, a dimension reduction (DR) technique is necessary to remove irrelevant, redundant and noisy features without sacrificing the performance of the underlying algorithm (K. K. Bharti & Singh, 2014). Feature selection (FS) techniques are robust DR methods that are used to determine the optimal subset of informative text features (C. Liu, Wang, Zhao, Shen, & Konan, 2017). The filter method uses statistical analysis to evaluate the selected subset of the features from the original large set (K. K. Bharti & Singh, 2015; Guyon & Elisseeff, 2003). Typically, the filter methods are often computationally less expensive than other methods because they are independent of any learning algorithms. They can

perform without any foreknowledge of the class label of the document (C. Liu et al., 2017).

1.2 Research Motivation

Over the last decade, the amount of the available online text information has increased at an unprecedented rate. Moreover, this online text information is diverse. There is no consistency whereas everyone can upload information on the internet in any format they choose. The Publishers use different sorts of strategies to be ranked at the top of search engine results without having any importance to the search. It is a tricky business to get relevant information from these search engines. they will return millions of pages or documents to a query in response. The effect is even greater if the query is vague as search engines try to retrieve documents for all of a query's meaning. Search results clustering is a way to organize this vast number of documents in the form of groups in which members of the group share similar qualities. We took note of some of the search engines that used the clustering results, such as Yippy [2], and Lingo3G [3]. Work on cluster labeling is also going hand in hand as search result clustering is being widely researched. Topic Extraction (TE) is equally essential because the cluster does not describe its content, then there is less possibility that a user would be able to pick it even though it contains the information they choose. TE of clusters should be relevant and describe the clusters correctly to lead the user into the right cluster of documents.

TDC technique aims to cluster a document collection into smaller groups (clus-

[2] http://yippy.com
[3] http://www.carrot2.org

ters), where each group is on a different topic. The same topic group is shared by the text documents that are in the same cluster and different clusters represent different topics (Kushwaha & Pant, 2018). After that, the clusters pass through one of the topic extraction approaches such as the statistical approach to obtain the most important topics contained in the documents in each cluster (Velden et al., 2017). This will generate a partition of document clusters with labels. Finally, a web interface is built allowing users to effectively browse and locate topics of interest in the documents. Figure 1.1 illustrates the workflow of extracting topics that are contained in text documents.

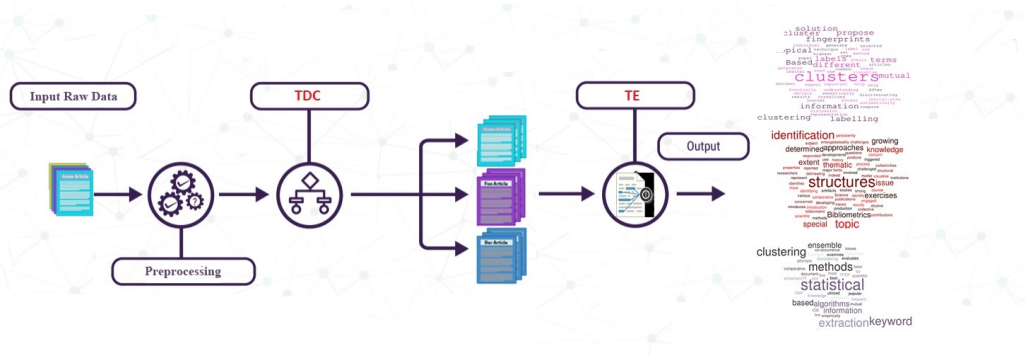

Figure 1.1: Workflow of extracting topics contained in text documents.

1.3 Problem Statement

From the literature on TDC and TE techniques, there are still some challenges that hinder the development of text mining tools to achieve a better representation of huge text documents. The most challenging aspects can be summarized as follows:

First, in the pre-processed step, the text documents must be transformed into a numerical vector of features for each document. Since the dimension of a corpus is usually very large, which normally contains informative and uninformative text fea-

tures (i.e., irrelevant, redundant and noisy features) (K. K. Bharti & Singh, 2014). The existence of irrelevant, redundant and noisy features not only increases the computational complexity of the clustering algorithm, but also decreases the clustering performance and sometimes misleads it (K. K. Bharti & Singh, 2015). Therefore, FS is a commonly used method to reduce the corpus dimension of corpus by selecting a discriminative subset of features from a high dimensional feature space. FS is a type of binary optimization problem. Various types of search techniques have been proposed in the literature such as sequential forward and backward feature selection(SFS, SBS) to overcome FS problems. However, these techniques may have a premature convergence problem or have more computational complexity. To alleviate these problems, evolutionary computation techniques which are population based solvers produce optimum solutions with less computational cost for binary optimization problems. These techniques have been widely used for finding the global optimum and gaining popularity day by day. There are so many meta heuristic algorithms such as, particle swarm optimization (PSO) (Kushwaha & Pant, 2018), artificial bee colony (ABC) (H. Wang et al., 2016), genetic algorithms (GA) (Too & Abdullah, 2020) and ant colony optimization (ACO) (Ahmad, Bakar, & Yaakub, 2019), are used for FS problem. Accordingly, this study aims to propose an efficient and new FS method as one of the pre-processing steps for text clustering to improve the performance of the underlying algorithm by including informative and avoid uninformative text features.

Second, in the clustering step, many related studies used the metaheuristic optimization algorithms like GA (Jiang, Wang, Chu, & Yu, 1997), PSO(Cura, 2012), ABC(K. K. Bharti & Singh, 2016a) trying to solve the text documents clustering problem (TDCP) with same levels of complexity of the documents (i.e., there is no di-

versity in the size of datasets). The obtained results, as well as the analysis of these studies were successful to some degree in improving the clustering performance; they have overcome the problem of local optimum in the local methods such as K-Means (Mishra, Saini, & Bagri, 2015), k-medoids (Park & Jun, 2009). In TDCP, the meta-heuristic algorithms do not strike a good balance between exploration and exploitation in the search space. This means that moving towards the global optimum solution is not guaranteed (Mirjalili, Mirjalili, & Hatamlou, 2016). In addition, they usually start by creating a set of random solutions; these initial solutions are then moved, evolved or combined over the iterations or generations during the execution. Until today, create a random initial population (solutions) has been the standard method. Therefore, the obtained results of the algorithm depend (among other factors) on the quality of the solutions in the initial population (K. K. Bharti & Singh, 2016a).

Third, in the TE step, the term frequency-inverse document frequency (TF-IDF) method (Lee & Kim, 2008), the most frequent based keyword extraction (TF) method (Z. Wang, Hahn, Kim, Song, & Seo, 2018), Latent Dirichlet allocation (LDA)(Bhat, Kundroo, Tarray, & Agarwal, 2020), and the mutual information (MI) method (Koopman & Wang, 2017) are the most widely used statistical TE techniques. These methods suffer from one main limitation. Each statistical method relies on a different metric of various characteristics. For specific document collections, each statistical method produces varied topics from another statistical method on the same cluster. The method, which yields the optimum results of a specific clustering solution, may not perform in the same way with another clustering solution, which means that the results are extremely varied, i.e., there is a weakness in the selected topics.

Accordingly, this study aims to address the above-mentioned problems.

1.4 Research Objectives

This study aims to develop an effective automated TE method for text documents based on optimization clustering, which provides a very compact summary of the clustered text documents. To achieve this aim, the following objectives are addressed:

1. To propose a feature selection technique for finding a new subset of more relevant features to reduce the high dimensional feature space of text collections.
2. To enhance the text documents clustering technique by:
    - proposing a suitable population-based algorithm for text documents clustering problem.
    - improving the balance between exploration and exploitation phases of the population-based algorithm in the search space.
    - enhancing the quality of the initial candidate solutions of the population-based algorithm using the local search strategy.
3. To propose an ensemble topic extraction method by combining several statistical methods to provide more coherent topics.

1.5 Contribution of the Study

The contribution of this study can be summarized as follows:

1. The first adaptation of a multi-verse optimizer (MVO) as new metaheuristic algorithm to solve feature selection problem.

2. Introduced a new optimization technique for solving the text documents clustering problem.

    - The first adaptation of a MVO as new metaheuristic algorithm to solve the text documents clustering problem.
    - The modification of a MVO to improve the balance between exploration and exploitation in the search space for the text documents clustering problem.
    - The hybridization of a MVO with K-Means clustering algorithm to enhance the quality of initial candidate solutions for the text documents clustering problem.

3. Ensemble five popular statistical TE methods: most frequent based keyword extraction (TF), term frequency-inverse document frequency (TF-IDF), co-occurrence statistical information-based keyword extraction (CSI), TextRank (TR), and mutual information (MI) to extract coherent topics from a collection of documents for each cluster extraction, instead of only one TE method, as in related work.

1.6 Scope of the Study

This study focuses on introducing a TE approach for an enormous digital collection of text documents. These text documents (any text) have certain characteristics. They are written in English, they have an unstructured format, in various sizes (medium and large), and they have high-dimensional informative and uninformative text features.

1.7 Methodology Overview

This section briefly outlines the four phases of the adopted methodology to achieve the research objectives of this study. As shown in Figure 1.2, the commonly used standard benchmark datasets in the literature will be implemented to evaluate the results that are obtained from different methods. In the first phase, after the pre-processing step (i.e., tokenization, stop words removal and stemming, transfer the data into numerical form). In the second phase, the FS method will find the highest relevant features' subset with a high accuracy rate using MVO algorithm. After that in the third phase, the MVO algorithm will be adapted directly into discriminative features subset to partition documents into several predefined groups (i.e., the number of clusters as an input parameter K) based on their similarity of contents. In addition, the MVO will be involved properly in the strategy and hybridization of the standard MVO algorithm to improve the exploration-exploitation balance and initial solutions, respectively. Finally in the fourth phase, an ensemble TE method will be applied to extract coherent topics from the obtained clusters and provide good human-understandable labels. A detailed description of the implemented methodology is provided in Chapter 3 of this study.

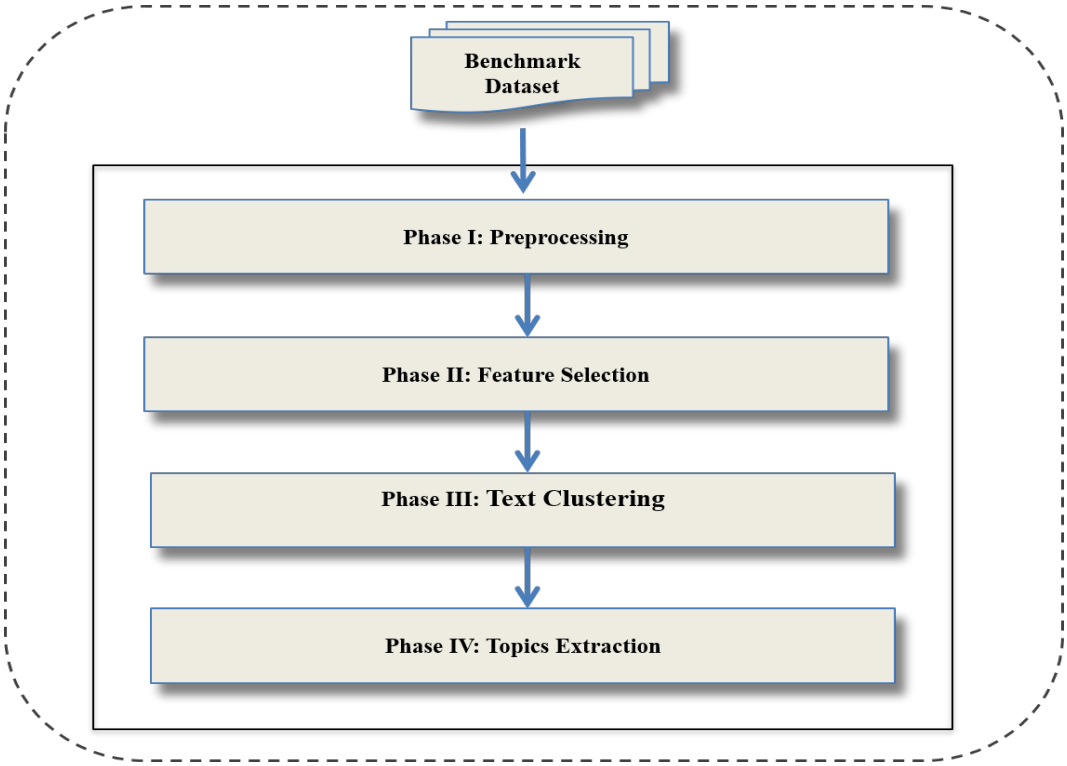

Figure 1.2: Research methodology overview

1.8 Organization of the Study

This thesis is organized into nine chapters, as follows:

Chapter 1: This chapter introduces the topic of the study and discusses the statement of the problem. It also presents the research objectives, scope, methodology, and contribution of the study.

Chapter 2: This chapter reviews the methods that are used to solve the FS problem, TDCP, and TE with some of the state-of-the-art, as well as the inspiration and mathematical model of the MVO algorithm.

Chapter 3: This chapter presents the research methodology. Four main phases in

this research to achieve the research objectives, namely i) Text Preprocessing ii) Text feature selection iii) text document clustering iv) topic extraction.

Chapters 4, 5, 6, 7: illustrate Binary MVO (BMVO), MVO, Link-based MVO (LB-MVO), Hybrid link-based MVO (HLBMVO), consecutively. Each chapter explains a comprehensive description of the proposed method and the sequence of the procedures conducted. Besides, it discusses the experiments and results of all proposed methods and presents comparisons of each method with the others.

Chapter 8: This chapter illustrate the proposed method to extract the topics from the clustered documents after the HLBMVO is introduced. Firstly, the chapter explains the existing keyword extraction methods — secondly, the proposed method. Finally, the chapter discusses the experiments and results.

Chapter 9: This chapter provides the research conclusion and possible future works.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents a detailed and comprehensive background of FS, TDC, and TE phases. Each of these phases is described, in terms of basic concepts, methods, and related studies. Also, an overview of the MVO algorithm is provided in the chapter. This chapter is organized as follows: Section 2.2 provides an overview on FS methods' categories, which are available in the literature. Sections 2.3 and 2.4 review the approaches of TDC, as well as the TE. Section 2.5 discusses the inspiration and mathematical model of the MVO algorithm. The conclusion of this chapter is provided in Section 2.7.

2.2 Text Feature Selection

Often, a large number of features are involved in Text Documents. Nevertheless, not all the features can have significance because many features can be redundant or irrelevant. This, in turn, decreases the algorithm performance (i.e., the clustering algorithm). The aim of FS involves tackling this issue via the selection of a specific small subset of the relevant features only from the larger set (i.e., the original set of features). Therefore, when the redundant and irrelevant features are removed, the dimensionality of data can be reduced by FS, thereby enhancing performance (Kushwaha & Pant, 2018; C. Liu et al., 2017).

The FS methods are generally classified into three categories according to the varied strategies of the search, including the filter method, the wrapper method, and the hybrid method. Based on the filter method, relevant features can be identified without utilizing any of the machine learning algorithms. Based on the wrapper method, the learning method is used for selecting the informative features. The hybrid method, on the other hand, works on integrating the two methods of feature selection into the learning model. By doing so, high accuracy levels or good performance can be achieved with a moderately acceptable computational cost (e.g. the support vector machines and the least square regression).

The methods of filter feature selection have been widely utilized for the dimension reduction, particularly in text clustering due to their simplicity and scalability; they are less complex computationally, and they are considerably accurate (Ferreira & Figueiredo, 2012). Examples of filter feature selection include information gain (IG) (Salton & Buckley, 1988) and MI (Battiti, 1994), as well as chi-square (CHI) (Li, Luo, & Chung, 2008), in addition to mean-median (MM), and last but not least the modified arithmetic mean geometric mean (AMGM).

FS can be defined as one optimization problem type (K. K. Bharti & Singh, 2016b). The metaheuristic algorithms were applied to tackle the problem of FS successfully by locating the best feature subsets. They used the filter methods to evaluate varied features' subsets until the value of the maximum objective function can be obtained. Examples of these methods are PSO (Kushwaha & Pant, 2018) and ABC (H. Wang et al., 2016), as well as ACO (Aghdam, Ghasem-Aghaee, & Basiri, 2009).

Based on K. K. Bharti and Singh (2015), a hybrid method of feature selection is proposed. This proposed method works on integrating one method's advantages and reducing the drawbacks of another (i.e., Term variance (TV), as well as Document frequency (DF)). This unification or integration method can be traditionally used for merging feature sub-lists. Regarding the union approach, it combines all the present features in the selected feature lists, thereby increasing the total features' number. The intersection approach, however, works on selecting the common features in the entire selected feature lists. A relatively smaller number of features are selected by this approach, but sometimes highly ranked features are ignored. A mid-approach called 'modified union' is, therefore, introduced. This approach works on selecting all the features that are highly ranked, in addition to the common features in the selected feature lists. This can be done without the need to further increase dimensions in the feature space. Principal component analysis (PCA) is applied at a later stage for further refinement of the selected feature subspace. Therefore, dimensions in the feature space are reduced without the loss of a considerable amount of information. The experimental analysis involved three varied benchmark text datasets, including Reuters-21,578 and Classic4, as well as WebKB. The proposed method of dimension reduction was compared with traditional methods of single dimension reduction, as well as traditional strategies of feature sub-lists merging.

According to Abualigah and Khader (2017), a novel feature selection method, which depends on the hybrid PSO algorithm with the GOs (H-FSPSOTC) has been proposed. The k-means clustering can be utilized to assess how effective the achieved features subsets are. The experiments involved eight common text datasets of different characteristics, and the results showed that the H-FSPSOTC algorithm has improved

the clustering algorithm's performance by producing a novel subset of further informative features.

Kushwaha and Pant (2018) introduced a novel method of feature selection for the unsupervised text clustering called link-based particle swarm optimization (LBPSO). This algorithm introduced a novel updating strategy, which learns from the neighbor optimal position rather than the global best. The original text dataset is taken by LBPSO as an input to produce a novel subset of distinguishing features. The k-means clustering algorithm works on taking the features as an input to assess the method of feature selection. This introduced algorithm outperformed other recognized algorithms on the eight subsets of the three benchmark text datasets regarding NMI and RI, as well as purity, in addition to accuracy measures. LBPSO has been verified by the experimental results that showed its effectiveness compared with the binary PSO-based FS method of text clustering. This introduced feature selection algorithm enhanced the results of the text clustering algorithm by multiplying the number of similar groups.

2.3 Text Document Clustering Approaches

This section reviews the relevant TDC works, which implemented traditional, as well as optimization algorithms.

2.3.1 Classical Partitional Clustering Algorithms

The classical partition clustering algorithms' basic idea involves considering the center of the text documents as the center of the corresponding cluster. K-means (MacQueen et al., 1967), K-medoids (Park & Jun, 2009), as well as fuzzy c-means clustering

(Bezdek, 2013) represent conventionally traditional partitional clustering algorithm. All typical partitional clustering algorithms can be scalable to larger datasets, but these algorithms cannot target a global convergence due to their dependence on the initial position of the cluster centers. They can converge to the solution of the nearest local optimum in the search space from the starting position of the search. Also, the algorithm's multiple runs cannot solve the problem to achieve the global optimum solution. Relatively low time complexity and high computing efficiency are the advantages of this method (D. Xu & Tian, 2015). The following sub-sections explain the k-means and K-medoids for TDCP.

The following subsections explain the k-means and K-medoids for TDCP.

- K-means Algorithm for Text Document Clustering

  K-means represents the most popular clustering algorithm. This algorithm is a prominent technique of partitional clustering. It was introduced over 50 years ago (Figueiredo et al., 2019). The K-means algorithm has been commonly utilized with the aim of dealing with huge databases due to its simplicity. Also, it is easily implemented, and it enjoys low computational complexity, as well as fast convergence (Nanda & Panda, 2014). The process involves two main iterative steps. In this process, the entire dataset is classified into clusters that are heterogeneous. Over the years, many visions of improvement have been developed with the aim of enhancing its performance like the Kernel K-means (Nanda & Panda, 2014), K-harmonic-means (Y. Kumar & Sahoo, 2015), and K-Medoids (Nanda & Panda, 2014; Subhadra, Shashi, & Das, 2015).

  When utilizing K-means, the data can be split into K groups, which are charac-

terized by the centroids (i.e., typically a specific cluster centroid represents the mean of points in a cluster), which is arbitrarily generated artificial data to signify the whole group (Abualigah, Sawaie, et al., 2017). The algorithms' steps are started by calculating the distances between samples and centroids. Every data sample can be allocated to the nearest centroid; each collection of the points is allocated to a specific centroid, which forms a specific cluster; each cluster's centroid can be updated based on the points allocated to a cluster. This process is carried out repeatedly until no further point that can change clusters. Figure 2.1 exemplifies the way this method functions, in which the circles represent data and crosses represent the centroids (Abualigah, Sawaie, et al., 2017). The K-means algorithm steps are provided in Algorithm 1.

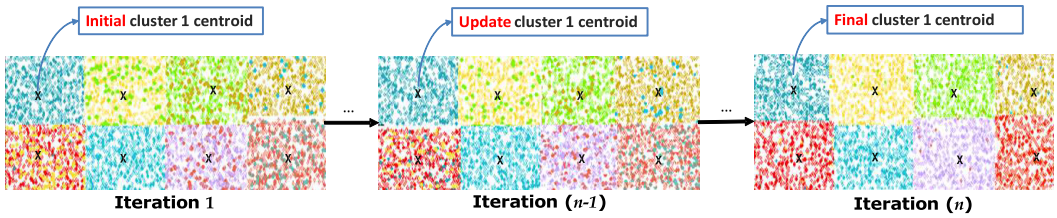

Figure 2.1: Illustrative example of K-means operations.

Algorithm 1 K-means clustering algorithm

Input: A collection of objects D, and number of clusters K.

Output: distribute objects D to clusters K.

Steps:

Step1: select an initial objects as centroids clusters K .

Step2: **while** *the end criterion is not achieved* **do**

Compute new cluster centroid for each k cluster.

Assigning the objects to its closest cluster (less distance between the object and clusters centroids).

end

C.-H. Chen (2017) used K-means with other clustering algorithms after proposing a novel scheme for the distance-based term weighting to encode the term weights via considering distances among the new terms and whether these terms have occurred. The proposed work provided the potential to improve clustering performance.

(S. F. Hussain & Haris, 2019) embed exploiting the statistical information of the data into the k-means algorithm instead of utilizing this information as a distance measure, which is external, thereby presenting a specific framework, which is integrated, namely the k-means based co-clustering (kCC) Algorithm. Also, the initialization step can be modified to involve many points with the aim of representing every cluster center like points, which are close altogether in a specific cluster. They are, however, far from the points, which represent different

clusters. Furthermore, statistics of the neighborhood walk are recommended as a semantic similarity method for cluster assignment, as well as center re-estimation in the iterative process. Evaluating this combined method was carried out on some standard datasets. The proposed approach (i.e., kCC) outperformed the k-means, k-means++, and ICC k-means, as well as Hierarchical Ensemble Clustering, in addition to SSID k-means traditional algorithms.

- k-Medoids Algorithm for Text Document Clustering

  The K-medoids text clustering algorithm has a similar function compared with the K-means text clustering. It begins by selecting k documents randomly as initial medoids with the aim of representing k clusters. Other documents, which are close to the medoid, are involved in the cluster. Subsequently, a novel medoid is chosen, which well represents the cluster. The documents are assigned to the clusters, which have the closest medoid. The medoids modify their location in each iteration. This method works on minimizing the number of dissimilarities among documents and their corresponding medoid. The cycle is repetitive until no medoid modifies its placement. The process ends here and the final clusters along with the medoids are defined. The formed K clusters are centroid on the medoids. All the members of the documents are put in the most appropriate cluster depending on the nearest medoid (Vishwakarma, Nair, & Rao, 2017).

  Balabantaray, Sarma, and Jha (2015) compared the K-means clustering with the K-medoids clustering. K-means is performed utilizing the Euclidean and Manhattan distance on WEKA tool, and K-medoids is performed by using Java programming. The results showed that K-means produced better results compared with K-medoids. However, the use of the k-medoid clustering algorithm suffers

from a couple of disadvantages. First, it needs many repetitions so that convergence is reached in addition to the slow implementation. It is because each of the iterations needs similarity computation or distance measures. Second, the k-medoid clustering algorithms cannot be compatible with the sparse text collection. Also, in the large division, as well as the documents' non-uniform distribution, a text does not include several words in common; the similarity value is quite small among these document pairs (C. C. Aggarwal & Zhai, 2012).

2.3.2 Metaheuristic Algorithms for Text Document Clustering

Over two decades, substantial efforts were exerted by metaheuristic algorithms to solve TDC because the existing deterministic approaches exhibited powerlessness in finding globally optimal solutions to such problems. Most of the algorithms were originated from the evolutionary-based algorithms (EA) survival of the fittest theory, trajectory-based algorithms (TAs), and swarm intelligence (SI) (Alyasseri, Khader, Al-Betar, & Awadallah, 2018). In general, all the metaheuristics are nature-inspired (i.e., inspired by ethology, biology or physics). Their components exhibit stochastic behaviors (involving random variables) and they set many parameters that need to be adapted to the problem (K. Hussain, Salleh, Cheng, & Shi, 2018; Makhadmeh, Khader, Al-Betar, & Naim, 2018).

Evolutionary algorithms are the algorithms that mimic evolutionary processes in their nature. These algorithms are grounded on the survival of the fittest candidate for a specific environment. They begin with a specific population (i.e., a set of solutions) that endeavors to survive in a specific environment (and are defined with fitness evalu-

ation). The parent population works on sharing its adaptation properties in an environment with children having various evolution mechanisms like genetic crossover and mutation. This process remains for several generations (i.e., iterative process) until the most suitable solutions are obtained for an environment. Some of these evolutionary algorithms include GA (Goldberg & Holland, 1988), Evolution Strategies (ES) (Back, Hoffmeister, & Schwefel, 1991), Genetic Programming (GP) (Koza, 1994), Differential Evolution (DE) (Storn & Price, 1997), and Biogeography-Based Optimization (BBO) (Gong, Cai, & Ling, 2010).

TA is an algorithm, which is initiated by a single solution. Iteration by iteration, this solution undergoes improvements using neighboring-moves operators until the most optimal local solution in a similar search space region of the initial solution can be obtained. Although TAs can deeply search the search space region of the initial solution and reach the local optima, they fail to navigate some search space regions simultaneously. The main TAs, which are utilized for TDC, include K-means and K-medoids, whereas other TAs, which are used for TDC, include β-hill-climbing (Abualigah, Sawaie, et al., 2017) and self-organizing maps (SOM) (Bernard, Buoy, Fois, & Girau, 2018; C.-H. Chen, 2017).

SI represents the natural metaheuristics group, which is inspired by 'swarms collective intelligence' (Aljarah, Mafarja, Heidari, Faris, & Mirjalili, 2020; Makhadmeh, Khader, Al-Betar, Naim, Abasi, & Alyasseri, 2019; P. Xu, Luo, Lin, Qiao, & Zhu, 2019). This collective intelligence is built via a homogeneous agents' population; these agents can interact with one another, and they can interact with the environment as well. Good examples of this intelligence can be found in the ants' colonies and

flocks of birds, as well as schools of fish, etc. The PSO (Cura, 2012) is developed following the birds' swarm behavior. The firefly algorithm (FA) (X.-S. Yang, 2010) is formulated according to the fireflies' flashing behavior, whereas the Bat Algorithm (BA) (X.-S. Yang & Hossein Gandomi, 2012) depends on the bats' echolocation behavior.

New metaheuristic algorithms have been recently enhanced, and they were applied with the aim of solving varied types of problems. Some of these algorithms include dragonfly algorithm (DA) (Mirjalili, 2016), coyote optimization algorithm (COA) (Pierezan & Coelho, 2018), hybrid binary ant lion optimizer (Mafarja & Mirjalili, 2019), the algorithm of the innovative gunner (AIG)(Pijarski & Kacejko, 2019), cultural coyote optimization algorithm (CCOA) (Pierezan, Maidl, Yamao, dos Santos Coelho, & Mariani, 2019), farmland fertility (Shayanfar & Gharehchopogh, 2018), center of mass optimization (CMO) (Gholizadeh & Ebadijalal, 2018), artificial neural networks (ANNs) (Sadollah, Sayyaadi, & Yadav, 2018), circular structures of puffer fish (Catalbas & Gulten, 2018), ludo game-based swarm intelligence (LGSI) (P. R. Singh, Elaziz, & Xiong, 2019), sailfish optimizer (SFO) (Shadravan, Naji, & Bardsiri, 2019), queuing search (QS) (J. Zhang, Xiao, Gao, & Pan, 2018), water wave optimization (WWO) (J. Zhang, Zhou, & Luo, 2019). The next subsections outline the recently conducted works that investigated the partitional clustering.

- Genetic Algorithm

  GA (Pal & Wang, 2017) is a very popular evolutionary algorithm, which has been first pioneered by John Holland in the 1970s. The basic idea of GAs is designed to make artificial systems software that retains the robustness of the