شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

بسم الله الرحمن الرحيم

HANAA ALY

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

HANAA ALY

شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات

يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

HANAA ALY

AIN SHAMS UNIVERSITY
FACULITY OF ENGINEERING
Computer and Systems Engineering Department

Clustering and Relating Research Papers using Self-Organizing Maps

A Thesis submitted in partial fulfilment of the requirements of the degree of

Master of Science in Electrical Engineering

(Computer and Systems Engineering)

by

Reham Fathy Mahmoud Ahmed

Bachelor of Science in Electrical Engineering

(Computer and Systems Engineering)

Faculty of Engineering, Ain Shams University, 2014

Supervised By

Prof. Dr. Hani M. K. Mahdi
Professor of Computer Systems
Computer and Systems Engineering Department
Faculty of Engineering, Ain Shams University

Dr. Cherif Ramzi Salama
Doctor of Computer Systems
Computer and Systems Engineering Department
Faculty of Engineering, Ain Shams University

Cairo – 2021

AIN SHAMS UNIVERSITY
FACULITY OF ENGINEERING
Computer and Systems Engineering Department

Clustering and Relating Research Papers using Self-Organizing Maps

by

Reham Fathy Mahmoud Ahmed

Bachelor of Science in Electrical Engineering

(Computer and Systems Engineering)

Faculty of Engineering, Ain Shams University, 2014

Examiners' Committee

Name and Affiliation	**Signature**
Prof. Dr. Passent Mohamed ElKafrawy Faculty of Science, Menofia University	.....
Prof. Dr. Hoda Korashy Mohamed Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University	.....
Prof. Dr. Hani M. K. Mahdi Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University	.....
Dr. Cherif Ramzi Salama Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University	.....

Date: 3-4-2021

Statement

This thesis is submitted as a partial fulfilment of Master of Science in Electrical Engineering (Computer and Systems), Faculty of Engineering, Ain shams University.

The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Student Name:

Reham Fathy Mahmoud Ahmed

Signature:

Date: 3-4-2021

Researcher Data

Name	: Reham Fathy Mahmoud Ahmed
Date of birth	: 19.11.1991
Place of birth	: Cairo
Last academic degree	: B.Sc. in Electrical Engineering
Field of specialization	: Computer and Systems Engineering
University issued the degree	: Ain Shams University
Date of issued degree	: 07.2014
Current job	: Senior Business Intelligence and Data Analysis Engineer

Abstract

Text data increases every day with a huge amount. We usually deal with vast quantities of text data through the Internet or on our computer systems. It will be useful to have a method to organize this huge amount of text data. With this huge increase, clustering text papers becomes an important research topic.

For many years and till now many researchers are trying to find the best algorithm to make text clustering. Many algorithms were used to perform text clustering such as Naïve Bayes, Support Vector Machines (SVMs), and Self-Organizing Maps (SOMs).

Research papers are a special type of text documents as they have specific expressions and scientific keywords. This was our motivation to develop an algorithm which can cluster research papers. This thesis proposes a method to cluster research papers based on SOMs.

A SOM is an unsupervised machine learning method. It has some parameters which need to be optimized in order to produce the best possible solution. These parameters are either set manually or using trial and error methods. In this work, we propose to use the well-known genetic algorithm to search the parameter space in an effort to find the best values automatically. Accordingly, in this thesis we decided to use SOM algorithm optimized by genetic algorithm.

First, we built our algorithm and test it on clustering gray colors as a simple case study in order to test our algorithm and measure its efficiency. Then we applied our algorithm on three different research papers data sets to cluster them. To achieve better results, we also integrated our suggested algorithm

with a pretrained Word2Vec model to be able to match different words having similar meaning. Finally, we compared our results with previous research on clustering research papers showing that our work outperform their results which was already compared it to many other earlier methods.

Keywords: Document clustering, Word2vec, cluster validity indices, Self-Organizing Maps, Genetic Algorithm.

Thesis Summary

A Self-Organizing Map (SOM) is a powerful tool for data analysis, clustering, and dimensionality reduction. It is an unsupervised artificial neural network that maps a set of n-dimensional vectors to a two-dimensional topographic map. Being unsupervised, SOMs need little input to be successfully deployed. The only inputs needed by a SOM are its own parameters such as its size, number of iterations, and its initial learning rate. The quality and accuracy of the solution offered by a SOM depend on choosing the right values for such parameters. Different attempts have been made to use the genetic algorithm to optimize these parameters for random inputs or for specific applications such as the traveling salesman problem. To the best knowledge of the authors, no roadmaps for selecting these parameters were presented in the literature. In this thesis, we present the first results of a proposed roadmap for optimizing these parameters using the genetic algorithm and we show its effectiveness by applying it on the classical color clustering problem as a case study.

With the huge amount of published research papers, retrieving relevant information is a difficult task for any researcher. Effective clustering algorithms can help improve and simplify the retrieval process. After testing our proposed approach on the case study, we applied our proposed approach on automatic clustering of text documents. The proposed method is applied to cluster 3 scientific papers datasets using their keywords. Similar research papers were mapped closer to each other.

This thesis is divided into 7 Chapters as follows: chapter 1 is an introduction to the research in this thesis. Chapter 2 discusses document clustering. It defines document clustering highlighting the difference

between clustering and classification. The chapter then elaborates on the text documents clustering problem and its details, the text document pre-processing steps, word embedding, clustering algorithms, and clustering techniques. Chapter 3 introduces Self Organizing Maps, their properties, topologies, steps and applications. Chapter 4 briefly explains the genetic algorithms describing its steps, crossover and mutation operators, selection methods, its advantages and disadvantages, and genetic algorithm applications. Chapter 5 describes the proposed method and how it can applied on any clustering problem such as the colors case study or on clustering research papers. Chapter 6 lists the obtained results for both clustering problem showing that we outperform previous clustering techniques. Chapter 7 concludes the thesis' work and discusses potential directions for future work.