Ministry of Higher Education
and Scientific Research
University of Diyala- College of Science
Department of Computer Science

Pattern Discovery for Text Mining Measured by Levenshtien Distance

A Thesis

Submitted to the Department of Computer Science\ College of Science\ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Layla Abd Al.Hak Ismael

Supervised by

Prof. Naji Mutar Suhaib

September 2019

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

﴿هُوَ الَّذِي جَعَلَ الشَّمْسَ ضِيَاءً وَالْقَمَرَ نُورًا وَقَدَّرَهُ مَنَازِلَ لِتَعْلَمُوا عَدَدَ السِّنِينَ وَالْحِسَابَ مَا خَلَقَ اللَّهُ ذَٰلِكَ إِلَّا بِالْحَقِّ يُفَصِّلُ الْآيَاتِ لِقَوْمٍ يَعْلَمُونَ 5﴾

صدق الله العظيم

سورة يونس

الآية (٥)

الإهداء

أهدي جهدي المتواضع هذا

إلى الذي سهر على تعليمي بتضحيات جسام.....إلى مدرستي الأولى في الحياة

أبي الغالي على قلبي أطال الله في عمره

إلى من كانت سندي في الشدائد..... وكانت دعواها لي دائما بالتوفيق

أمي الغالية جزاها الله عني خير الجزاء في الدارين

إلى من كانوا ملاذي وملجئي.....إلى من تذوقت معهم أجمل اللحظات

إلى من هم اقرب إليّ من روحي.....أخواتي

إلى من آنسني في دراستي وشاركني همومي.....زوجي العزيز

إلى الشموع التي ذابت في كبرياء.....لتنير كل خطوة في دربنا

فكانوا رسلاً للعلم والأخلاق .....أساتذتي

ليلى عبد الحق إسماعيل

ACKNOWLEDGMENTS

First and foremost, I would like to thank to Allah SWT for his bless and mercy who has guided me in finishing this thesis. Then I would like to thank my supervisor, ***prof. Naji Mutar Suhaib****, professor of computer science at Diyala University – collage of science, for the extraordinary exertion he applied, I would like to thank him for his profitable direction and backing through his supervision of this work. I am very fortunate to receive such great support from him.*

I would like to thank my father, mother and sisters who have continued to provide encouragement and assistance over the last two years of study.

At long last, there are no words enough to thank my husband for being strong and having faith in me constantly and his support to complete this master project.

Abstract

The textual based information amount is rapidly accumulating Which stored electronically on our computers or Web. Any computer (laptop or desktop) is capable of accommodating enormous data amounts owing to the improvements in the storage devices.

Texts are included in text dataset and this dataset are unstructured. These unstructured data can be handled by text mining. The complexity and the considerable number for these data uncover numerous new capabilities to the analysts. Therefore, this work presents an enhancement of extracting useful patterns from text documents in the field of text mining using Pattern Taxonomy Model (PTM) and Levenshtein Distance Algorithm (LDA).

There are various methods to handle text documents. In this thesis, text mining system was suggested to overcome the problems that have occurred in term-based method and phrase-based method. The proposed system based on the behavior of LDA algorithm and PTM for determining the best accuracy of the extracted patterns with a short time and to prove that pattern based method is the best solution for text mining without any problems in the information extracted from the text.

The strength of the two algorithms (PTM, LDA) are tested using threshold values from 1 to 10 to get 1% to 10% of information in the text. The proposed system used "Openosis opinion dataset" and "Reuters 50_50 dataset" which stored in a file of ".txt" or text document.

The results of this test obtained by comparing among values of four features which are (global probability, local probability, absolute support, relative support) for the text to get higher average accuracy.

The results of proposed system have been compared with other systems. The proposed system get (98.68%) average accuracy for Unigram grammar and (99.65%) average accuracy for Bigram grammar while a system that used the Levenshtein Edit Distance for automatic lemmatization for modern English achieved an accuracy of 96% for English language and the system that used the process of pattern evolving and pattern deploying get 62% of precision and 82% of recall. So, using LDA with PTM achieved a better results compared to other systems.

Lists of Contents

Lists of Figures

Lists of Tables

List of Algorithms

List of Abbreviations

CCAT	Criteria Cognitive Aptitude Test
CSV	Comma-Separated Values
HTML	Hyper Text Markup Language
IE	Information Extraction
IF	Intermediate Form
IPE	Individual Pattern Evaluation
IR	Information Retrieval
IT	Information Technology
JSON	Java Script Object Notation
KDD	Knowledge Discovery in Databases
LDA	Levenshtein Distance Algorithm
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
NOSQL	Not Only Structured Query Language
PAT-tree	Patricia Tree
PTM	Pattern Taxonomy Model
RCV1	Reuters Corpus Volume1
SQL	Structured Query Language
SUP_a	Absolute Support
SUP_r	Relative Support

SVM	Support Vector Machine
TF	Term Frequency
T_H	Threshold
TM	Text Mining
TREC	Text Retrieval Conference
TXT	Text Document
XML	Extensible Markup Language

Chapter One

General Introduction

Chapter One

General Introduction

1.1 introduction

80% of all the information saved in governments, businesses, industries, and other organizations are storing in the form of text. Texts are the most popular means of formally information exchanging. There is a requirement in our modernistic life to have a tool of business intelligence that is capable of extracting information as quickly as possible from texts. [1].

Data mining term is also called Knowledge mining that is the significant extracting of inherent, formerly unknown and possibly beneficial information from data in the dataset. It has many techniques such as Decision tree classifier, Neural network, Genetic algorithm, Rule extraction [2].

Different applications like business managing and market analyzing can benefit from the utilization of extraction information from a huge data amount. The process of knowledge discovery can be represented as a non-trivial extracting of information from big databases, information which is presented in an implicit manner in the data, formerly unknown and possibly beneficial to users. Data mining is a fundamental stage in the knowledge discovery process in dataset. Based on the approaches of data mining, a considerable number of patterns are created[3].

The processes of finding the effective use and updating the patterns remain an open research problem. The developed model of knowledge discovery satisfies this issue and is capable of applying the patterns in the domain of text mining [4].

The techniques of text mining are very useful to users for finding the desired knowledge from a massive data amount. It is extremely significant to retrieve efficient and relevant information for the users. Term-based approaches were used to provide these requirements. But these approaches have several drawbacks like polysemy and synonymy. Polysemy refers to a word which has several meanings, while synonymy refers to the words which have the same meaning[5].

For overcoming these drawbacks, the phrase-based approaches were presented. But these developed approaches also had several drawbacks such as

1- Lower statistical properties to terms.

2- The occurrence frequency of the phrases is minimal compared with the keywords

3- A considerable number of redundancy and noisy phrases.

In the presence of these drawbacks, pattern-based approach have been proposed to get rid of the problems in the previous approaches [6].

1.2 Related Work

- Ning Zhong et al. (2012) [7], proposed a technique for discovering patterns to handle the issues of low frequency and misinterpretation of text mining. The experiments were conducted on Reuters Corpus Volume1(RCV1) data collection and Text Retrieval Conference (TREC) topics. This technique also used the pattern deploying and evolving processes for finding the obtained patterns in text documents.
- Dipti S. Charjan and Mukesh A. Pund (2013) [8], proposed a technique for patterns discovery that encompasses the pattern deploying and evolving processes for improving the efficiency of utilizing and updating discovered patterns to find interesting and pertinent information. The presented technique used a document of .txt as input and applies different algorithms for getting useful patterns.
- Bharate Laxman, and D. Sujatha (2013) [9], presented a technique to discover patterns and then compute the patterns specificities to evaluate the concept of weights as per their distribution in the obtained patterns. This technique works on updating patterns which show ambiguity that is a characteristic called pattern evolution. Patterns deploying and evolving are also used. The obtained results on the prototype application expose that the obtained result is useful in the text data mining field.
- Charushila Kadu et al. (2013) [10], presented a hybrid system is working on minimizing the dimension dataset and similarity constraints. A feature-based analysis is used to reduce the dimensional of the massive datasets. This system uses the evaluation of feature for reducing the high dimensionality of

text vector, and then identifies the term frequency, after that, these frequencies are weighted via utilizing the inverse document frequency technique. The documents weight is utilized in clustering. The system combines both the semantic and similarity constraints and combined the Individual Pattern Evaluation PTM (IPE).

- V. Aswini and S. K. Lavanya (2014) [11], presented a technique that utilized the pattern taxonomy model for discovering the patterns from a large data amount and seeking for important patterns. This technique includes the pattern evolving and deploying processes for improving the efficiency of utilizing and updating derived patterns to find important and relevant information. This technique get 62% of precision and 82% of recall as a result.
- Shivani D Gupta and B.P.Vasgi (2015) [12], the designed system that concentrates on the performing of a specific manner to deriving the pattern, in addition, to utilize them for retrieving the relevant text. The pattern deploying and evolving are two processes that are utilized for developing the efficiency of the obtained patterns. Then, the obtained patterns are utilized to search for important and relevant information. to develop the efficiency of utilizing and updating obtained patterns regarding the view of users, the testing on RCV1 data collection fulfills the users needed data from a document.
- Vaishali Pansare (2016) [13], proposed an approach that uses the association rule mining based on the AprioriAll algorithm for discovering frequent patterns in text documents. The input can take different formats of a text file. It uses Hash Tree structure to store candidate itemsets, and find frequent

patterns and itemsets within less time. This proposed approach finds a solution to the misinterpretation and low-frequency issue. The obtained results have demonstrated that the time of execution needed by the algorithm is minimal than the time needed by the other compared algorithms.

- S. R. Lomate (2016) [14], The proposed system introduced a technique for pattern discovery that consists of the pattern deploying and evolving processes, for improving the efficiency of utilizing and updating obtained patterns for finding the important and relevant information. The fundamental prototype of Pattern Taxonomy Model (PTM) using Advanced Apriori Algorithm which focuses on the problem of getting beneficial frequent patterns from the documents. This technique used substantial tests on RCV1 data collection shows that the obtained result provides a hopeful performance. A comparison between the Apriori algorithm and Advanced Apriori Algorithm is done for finding frequent patterns in pattern mining and a show advanced Apriori has better computation time for frequent patterns. Reduce computing time and rule count for frequent pattern generation.
- H. M. Mahedi Hasan et al. (2018) [15], presented a technique for key terms extraction that is depending on semantic relation. This approach is working on extracting a specified number of keywords from documents for identifying the main text content. The data set is collected from various sources like newspapers, books, journals, etcetera. To extract these keywords, several machine learning and statistical techniques have been utilized such as Logistic regression, support vector machine, word co-occurrences, and PAT-tree. This technique presented an approach of modified semantic relation with an accuracy of 77.6% precision and 84.3% recall to chosen data sets.

1.3 The Problem Statement

Lots of researches in the field of text mining have concentrated on improving effective mining algorithms to discover various patterns from larger text documents. Hence, finding interesting and useful patterns remains an open issue. Within this field, the techniques of data mining can be utilized for finding a variety of text patterns, like frequent itemsets, and sequential patterns. This thesis deals with three problems which are:

1- Getting an effective deal with the massive amount of text document.
2- Discovering the useful patterns from digital text documents, and utilizing these mined patterns to develop the performance of the system.
3- How to improve and evaluate the efficiency of the discovered pattern from text documents.

1.4 Aim of the Thesis

The aim of this thesis is to:

1- Extract interesting patterns from text document while these patterns contain accurate information about the document.
2- Reduce the time required for finding the patterns which describe the content of the document.
3- Increase the accuracy of information extracted from the text by using data mining techniques such as:
   a- Summarization.
   b- information extraction.
4- Proof that the discovery of a pattern is the best solution for text mining.

1.5 Outline of the Thesis

The other chapters in this thesis are as follows:

Chapter Two: Theoretical Background

This chapter gives the background and review of text mining and its techniques: (classification, clustering, retrieval information, extracting information, and summarization), the stemming, N-gram and Levenshtein algorithm

Chapter Three: The Proposed System

This chapter describes the proposed pattern discovery system with its design and implementation.

Chapter Four: Experiential Results and Tests

This chapter explains the results and evaluation that have been getting from the proposed system.

Chapter Five: Conclusions and Suggestions for Future Works

This chapter presents the conclusions of this work. Furthermore, it justify provides suggestions for future work.