Acknowledgements

First, I thank Allah the Almighty for giving me the strength and the ability to complete this thesis. Many thanks to my parents who are patiently encourage and support me to finish this work. I am also grateful to my husband for his support, understanding and constant encouragements.

My Sincere thanks and real indebtedness are dedicated to my supervisors **Prof. Dr. Fayed Ghaleb**, for his constructive guidance throughout the development of the work. The opportunity I had to share in his boundless vision, enthusiasm and advice is something for which I shall always be profoundly grateful, **Prof. Dr. El-Sayed Atlam** for giving me feedback and for his support as research on this thesis; and **Dr. Azza Taha** for her support and patience. Without their guidance and advice this thesis would not exist.

It is pleasure to express my great thanks to Mathematics department, Faculty of Science, Ain Shams University for give me this opportunity.

List of published papers

- Atlam E-S, Ghaleb F, Taha A, Ismail A. A new retrieval method based on time series variation using field association terms. Math Meth Appl Sci. 2017;1-12. https://doi.org/10.1002/mma.4713

Contents

List of Figures

List of Tables

List of Abbreviations

IR	Information Retrieval
FA	Field Association
SB	Stabilization
DT	Decision Tree
TF-IDF	Term Frequency-Inverse Document Frequency
TF	Term Frequency
IDF	Inverse Document Frequency
LSI	Latent Semantic Indexing
SVD	Singular Value Decomposition
TREC	Text REtrieval Conference
WWW	World Wide Web
NLP	Nature Language Processing
EBP	Error-Based Pruning
CF	Confidence Interval
DF	Descriptive Features
Di_name	Disease name
O_name	Organization name
D_name	Doctor name
T_name	Treatment name
m_name	Medical name

Inc.	Increment
Std.	Steady
Dec.	Decrement
SVM	Support Vector Machine

Abstract

With the advent of the World Wide Web, the significance of Information Retrieval (IR) has grown. IR is the process of searching for the relevant information in the subjects that interest the user and then retrieving it. Recent years have seen enormous increases in the amount of texts that are available electronically on the Internet. These texts are analyzed into useful information widely and then it can be utilized for searching, clustering, classifying, summarizing and retrieving information, etc. Typically, the system of IR searches in collections of data that are either unstructured or semi-structured. The user can get the needed information from a collection of documents by reading the whole documents, then maintaining the relevant documents and neglecting the others. The result of this retrieval may be good but not perfect because the popularity of words in a given period of time is not taken into account in the searching process. The time also is wasted in reading the whole document. In order to treat these drawbacks and to obtain highly effective retrieval results, a retrieval method based on time series variation using Field Association (FA) terms is suggested. Persons can determine the document field when they find particular words or conceptual units which are named FA terms without the need to read the whole document. In this thesis, we study the effects of the time change on the frequencies of FA terms in a given period of time. Furthermore, a method for automatic evaluation of the Stabilization (SB) classes of FA terms is suggested to improve the precision of Decision Tree (DT). The SB classes point out the popularity of list of FA terms depending on time change. The method is evaluated through conducting experiments (using Python programming language) by simulating the result of 1,245 files which are equivalent to 4.15 MB. The F-measure for Increment, Fairly Steady and Decrement classes achieves %90.4, %99.3 and %38.6, sequentially. Moreover, the problem of the scattering of data among classes is handled using two methods to improve the performance of DT. The two methods are random sampling method and data replication method. From the experimental results, the F-measure is %90.8 for Increment class, %99.5 for Steady class and %68.1 for Decrement class using random sampling method. While the F-measure is %93.6 for Increment class, %99.8 for Steady class and %75.7 for Decrement class using data replication method.

Chapter 1

Introduction

Recently, there have been enormous increases in the amount of texts of all types of data that are available electronically in the Internet, digital libraries and company intranets, etc. The abilities to store, process and access these texts from far away locations through a network have improved with the technological progress in computers. The texts are analyzed into important information, features and knowledge to be utilized for searching, clustering, classifying, summarization and retrieving information, etc. There have been many systems of document retrieval sophisticated along with progress in areas like automatic document clustering, summarization and classification and retrieval of similar files. In organizations, more than 70%-80% of all data are considered as unstructured information; Aarthi et al. [1]. Identifying non-trivial pieces of information, interesting knowledge or important words from unstructured information has assumed a great significance in Information Retrieval (IR). Machine learning is a technique used to analysis data for building an analytical model. Without the machine learning, the unstructured information is practically useless. Renewed attention in machine learning is due to factors like increasing amounts and diversities of available data, inexpensive and more powerful computational processing and accessible data storage. The learning is needed when the problem solutions changes in time or relied on the specific environment as well as when persons do not possess any experience or when persons cannot describe their experience. Search engines, medical diagnosis, email spam and credit scoring are some of the applications of machine learning. In machine learning, a performance criterion is optimized utilizing past expertise or example data; Alpaydin [3]. About 70% of machine learning is supervised learning. Supervised algorithms are trained using labeled instances. A learning algorithm takes all inputs with the correct outputs and the algorithm learns by the comparison between its actual outputs and correct outputs to determine errors. Supervised machine learning algorithms include Decision Tree (DT), k-nearest neighbors, support vector machines, Naive Bayes, neural networks, etc. Machine learning is shown to be an efficient method for not only automatic production of classification models, but also for making the previous results better.

There have been many electronic texts which may contain an enormous amount of important knowledge. It is virtually impossible for users to find the relevant information from these collections of texts by reading all of them due to the huge number of texts. Therefore, there is an increasing need for effectively organize, cluster, classify and determine relevant information using automatic methods.

Fukumoto et al. [26] proposed a statistical method for clustering a group of articles based on online dictionary definitions and then used the dictionary definitions to classify these articles, each of which belongs to the restricted topic field. Kaoru and Katsumasa [33] presented a method for determination of useful terms that represent the contents of articles of English contracts using Term Frequency-Inverse Document Frequency (TF-IDF) method. Liman [42] proposed a method for classifying cue phrases as discourse or sentential using machine learning techniques and then compared the automatic linguistically classification models generated from machine learning with manually derived models to determine the accuracy. Although all of these methods were useful in clustering, classifying, searching and analyzing documents but they were unable to define accurately the significance of words in a specific period of time.

Identifying important words has assumed a great significance in effective systems of Information Retrieval (IR). Some words occur frequently in texts and usually connected strongly with the text field. Generally, the frequencies of words in texts change over time. These words are frequently linked with specific period of time. For instance; "mumps" and "skin rashes" are more spread in summer , "common cold", "asthma attacks" and "sore throat" are more spread in winter and "blizzard" is more spread when there is a severe snowstorm with high winds and low visibility. This shows that some words are affected by time change. Politicians', actors' and athletes' personalities always change over time. Therefore, knowing the popularity of words according to time change is important to obtain efficient retrieval results.

Atlam et al. [7] presented a method to study the spread of words according to time change. This method was based on keywords which represent the documents but keywords are not the best representative of the texts.

With the rapid growth of data, there is still a challenge to retrieve and process this huge amount of digital data into useful pieces of information and knowledge. Persons generally can decide the text field when they found clearly defined words or conceptual units without the necessity to read the all text. These words or conceptual units are the Field Association (FA) terms. FA terms are a limited set of discriminating words that match a particular document best. FA terms are linked to scheme named field tree and FA knowledge base is constructed from the field tree along with FA terms under different fields. Some examples of FA terms are the term "Alzheimer" point to the sub-field <Diseases> under a super-field <MEDICINE> and the term "pill" point to the sub-field <Pharmacy> under a super-field <MEDICINE>. As opposed to the traditional methods that depend on probabilistic methods and classical vector space models, a new technique based on FA terms Fuketa et al. [25] and Tsuji et al. [87] has been proved to be useful and very effective in many areas such as retrieval of similar files Atlam et al. [5], classification of documents Fuketa et al. [25] and retrieval of passages Lee et al. [38]. Also, this technique carries greatly promise for application in areas such as machine translation Nguyen-Phan [56], cross-language retrieval Lu et al. [44] and summarization of texts Zhan et al. [92], etc.

The thesis is motivated by the need for automatic methods to determine the popularity of FA terms according to time change. Therefore, a retrieval method based on time series variation using Field Association (FA) terms is suggested. This method is considered as an important step to achieve highly effective retrieval results.

Decision Tree (DT) is a highly effective tool of machine learning and data mining. It is a powerful and performs well with large amount of data in short time. DT is used in several distinct disciplines such as diagnosis, cognitive science, engineering, artificial intelligence and data mining. The model of DT aims to understand the predictive structure of the problem and produce a precise classifier; Petri [60]. The Learning data of DT may contain varied amounts of data in various classes and there are some data in these various classes contain duplicated features. The classes that contain less number of data are considered as noise data compared with the class that contains more data amounts. This fewer number of data is negatively impact on precision of DT; Honda et al. [29]. The thesis is also motivated by the need to manipulate the problem of scattering data numbers among classes to improve the DT precision.

1.1 Contributions and Objectives

The thesis aims to study the effects of the time change on the frequencies of FA terms in specific period of time. It presents a method for automatic evaluation of the stabilization (SB) classes that result from applying the Decision Tree (DT) on the FA terms. This method is assumed to denote the popularity of list of FA terms over the time change and to improve the precision of DT. Moreover, the thesis solves the problem of the scattering (disparity) of data numbers among classes to improve the achievement of DT precision. A series of experiments are made to achieve these aims.

1.2 Thesis Outline

This thesis is organized as follows. Chapter 2 introduces an overview of the fundamental definitions and concepts related to the thesis in IR, text analysis, FA terms, decision tree and time series. Chapter 3 introduces some related works to this thesis. Chapter 4 introduces the retrieval method followed in this thesis for determining the popularity of list of FA terms over the time change and the experimental evaluation (data and results) for this method. Chapter 5 introduces two methods followed to manipulate problems connected with varying amounts of data among classes and to improve the achievement of DT precision, the results of these two methods and comparisons with the traditional method of Atlam et al. [7]. Finally, chapter 6 introduces conclusion and possible future work.

Chapter 2

Basic Concepts

This chapter introduces an overview for concepts and definitions related to IR, FA terms and decision trees. The chapter also discussed concepts associated to time series analysis and text analysis. The work in this thesis depends on these concepts and definitions.

2.1 Information Retrieval

Information retrieval (IR) is the process of finding information that satisfies users' information needs from large information collections. IR is interested with all activities linked to the processing of, organization of, and access to, all formats of information; Chowdhury [15]. The term ”Information Retrieval” was invented by Calvin Mooers in 1950 and acquired popularity in the community of research from 1961. At that period the organizing function of IR was seen as a main progress in libraries which were no longer just stores of books, but became places to catalog and index the information they hold. Over the last years, IR system has matured greatly. Nowadays, several systems of IR are used in every aspect of our daily lifetimes. For example, to find e-mail sent or received on a particular date, to find messages sent by or to a specific person, to find something on the web, and so on.

2.1.1 History of IR

Over the centuries, there was an increasing need to store and retrieve information particularly with the existence of papers and press. After the advent of computers, people used them to store and retrieve huge amounts of information which is represented by text, image, video and sound recording. Finding useful information became important in IR. Bush [13] was the first person who introduced the idea of automatic access to huge amounts of stored information. Researchers used his idea and created specific descriptions for the steps used to search text archives automat-