AIN SHAMS UNIVERSITY
Faculty of Computer &
Information Sciences
Information Systems Department

Enhancing Information Retrieval through Dependency Modeling

Thesis submitted to the Department of Information Systems
Faculty of Computer and Information Sciences
Ain Shams University, Egypt

In fulfillment of the requirements for the Master of Science Degree (MSc) in Computer and Information Sciences

By

Doaa Mabrouk Abd El-Fatah Mabrouk
Teaching assistant at Egyptian Chinese University

Under the Supervision of

Prof. Dr. Mohamed Essam Khalifa
Professor of Mathematics,
Faculty of Computer and Information Sciences,
Ain Shams University
Vice President for Graduate Studies and Research,
Egyptian Chinese University

Prof. Dr. Nagwa Badr
Dean of Faculty of Computer and Information Sciences,
Professor of Information systems,
Ain Shams University

Dr. Sherine Rady
Associate Professor, Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

May 2019

Table of Contents

List of Abbreviations

ACC	Accuracy
ANN	Artificial Neural Network
APM	Admixture Poisson Model
BG	Bi-Gram
BIR	Binary Independence Retrieval
BOW	Bag of Word
BRS	Bibliographic Retrieval Service
CDRM	Context Dependent Relevance Document
CM	Co-occurrence Model
CNN	Convolutional Neural Network
CONF WEIGHT	Confidence Weight
CRTER	Cross Term
DBN	Deep Neural Network
DL	Deep Learning
DM	Dependence Model
FD	Full Dependence
FI	Full Independence
FLC	Fuzzy Logic Controller
GRU	Gated Recurrent Unit
GVSM	Generalized Vector Space Model
IAP	Interpolated Average Precision
IDF	Inverse Document Frequency
IGR	Information Gain Ratio
IR	Information Retrieval
IRS	Information Retrieval System
KNN	K-Nearest Neighbor
LM	Language Model
LM	Link Model

LSI	Latent Semantic Indexing
LSTM	Long Short-Term Memory
MAP	Mean Average Precision
Max-FC	Maximum- Feature Count
Max-No-TDI	Maximum Number Term Dependency identification
ML	Machine Learning
MRF	Markov Random Field
NARX	Non-linear Auto Regressive model with Exogenous
NBC	Naïve Bayes Classifier
NLP	Natural Language Processing
NSLM	Non-Separated Link Model
PLM	Positional Language Model
PMRF	Poisson Markov Random Field
PRF	Pseudo Relevance Feedback
RNN	Recurrent Neural Network
SD	Sequential Dependence
SDM	Sequential Dependence Model
SRBC	Subsumptions Rule Based Classifiers
SSD	Solid State Drive
SVM	Support Vector Machine
SWDM	Semantic Weighted Dependence Model
TC	Text Classification
TDI	Term Dependency Identification
TF	Term Frequency
TF-IDF-AP	Term Frequency-Inverse Document Frequency-Adaptive Position
TRDM	Term Relevance Dependency Model
TREC	Text Retrieval Conference
UG	Unigram
VSM	Vector Space Model
XML	Extensible Markup Language

List of Figures

List of Tables

Abstract

In every field in our life, there are many problems especially in the field of computer. These problems increased due to the rapid spread of the internet. Today, the most important field in our life is information retrieval and the search to convey user's need. With the growth of using the internet and available information on the web, Information Retrieval "IR" became a fact of life for users. The internet is providing the user with vast knowledge and information in different domains. The major research areas include biology, chemistry, commerce, tourism, earth, education, mathematics, physics, economics, agriculture, and information and computer sciences.

In this thesis, the following problems are introduced: Term dependency, especially, that some of the mathematical models assume terms are independent. One of these models is Vector Space Model "VSM", while others, assume that terms are dependent such as Markov Random Field "MRF", Unigram and Bigram models. Term weighting is a core behind mathematical retrieval modeling which is important in document ranking. There are some methods such as Term Frequency Inverse Document Frequency "TF*IDF", Information Gain Ratio "IGR", Confidence weight "Conf.Weight" and weighted clustering.

The proposed algorithm of the power sets a theory to discover all the combinations between words in documents. Moreover, the judgement of the results uses accuracy measurements by Subsumptions Rule-Based Classifiers "SRBC" through two ways (Maximum-Number –Term Dependency Identification "Max-No-TDI" and Maximum-Feature Count "Max-FC").

This thesis introduces a survey of mathematical information retrieval systems' using dependency modeling and term weighting. The enhancement of dependency modeling is through performance, effectiveness and efficiency in addition to term weighting which considers another factor that affects the result. It also contains the power set theory to discover Term Dependency Identification "TDI" between words in Text Classification "TC" and measure accuracy of all generated random experiments. The result is Max-No-TDI which is better than Max-Fc with 96% accuracy level.

List of Publications

[1] Doaa Mabrouk, Sherine Rady, Nagwa Badr and M.E.Khalifa ;"A Survey on Information Retreival Systems' Modeling using Term Dependency and Term Weighting"; *In Intelligent Computing and Information Systems (ICICIS), 2017 IEEE Eighth International Conference* on, pp. 321-328. IEEE; Cairo; Egypt. https://ieeexplore.ieee.org/document/8260073/

[2] Doaa Mabrouk, Sherine Rady, Nagwa Badr and M.E.Khalifa; "Modeling Using Term Dependencies and Term Weighting in Information Retrieval Systems"; *Egyptian Computer Science Journal Vol.42 No.3 May 2018; pp. 57-73 ;* ***ISSN-1110-2586.***

http://ecsjournal.org/Archive/Volume42/Issue3/5.pdf

[3] Doaa Mabrouk, Sherine Rady, Nagwa Badr and M.E.Khalifa; "Enhancing Term Dependency using Power Set in Text Classification"; International Journal of Engineering Research and Applications (IJERA); vol.9 No.5 May 2019; pp. 33-39.
https://www.ijera.com/papers/vol9no5/Series-2/G0905023339.pdf

Acknowledgment

First and foremost, I am grateful to Almighty Allah for His immense blessings and graciously helping me to complete this thesis. This thesis owes its existence to the help, support, and inspiration of many people. In the first place, I owe my deepest gratitude to my supervisors Prof. Dr. Mohamed Essam Khalifa, Prof. Dr. Nagwa Badr and Dr. Sherine Rady for their great knowledge, experience and sharp sense of research direction have provided invaluable feedback to improve the quality of this thesis. This thesis would not have been possible without their sound advice and encouragement. It was my great pleasure and honor to have such professors as my supervisors for their endless support and cooperation during my study and research, in addition to their final revision of the thesis.

Last, but not least with all my appreciation and love that no words can express, I would like to thank all my friends and family members for their endless love and support. I offer my love and heartfelt thanks to my parents for their lifelong support in all my endeavors. I dedicate this thesis to my dear parents, they are behind any success in my life.

Chapter 1.

INTRODUCTION

1.1. Background

Information Retrieval "IR" research is the act of storing, retrieving and searching information according to the users' need. An IR system searches in any collections of data. These data are classified as structured or unstructured or semi-structured. With the increase's of computer technologies, the retrieval become grew also based on speed of processor and storage. Goals of IR are creating the ways that support humans to have better access to information to carry out their tasks. IR task is defined as finding documents characterized by unstructured "Text" that satisfy information need from large collections. So, data retrieval performed as Extensible Markup language "XML". The evaluation criteria of Information Retrieval System "IRS" is identifying measurable properties such as speed, performance, efficiency, complexity and accuracy.

The long history of information doesn't begin with the internet. The earliest computer-based searching system were built in the late 1940s. This system is (memex) system "bush". It used to store human knowledge in desktop device. It helps users to commented association among the stored knowledge. One of the disadvantages of this system is that computers operated in batch processing rather than iterative processing. In the 50's, the first computerized systems were built. In this period, the first large scale of information systems built such as punch cards and light. The prototype of these systems was completed in 1950 and demonstrated in 1951. There were commercial library systems such as dialog and Bibliographic Retrieval Services (BRS). BRS focused on biomedical communication but later added in science, business and technology. The early of 60's, definition of recall and precision were appeared to evaluate retrieval systems. Recall and precision are the traditional measurements in IR. These are used to measure relevance and non-relevance documents. These measurements have not advantaged because of number of exact relevant items didn't know well. So, it is impossible to calculate recall also, precision. The relevance of feedback was introduced. This process was used to support iterative search. In this process, the first step is that the user was selected relevant documents and the second, terms from the documents added to the

query. In the fact of 60's, many people attended IR conferences such as Special Interest Group on Information Retrieval (SIGIR).

During 1970's, retrieval began to grown in to real systems. The most important reason was the development of computer word processing which a lot of text was available in machine readable form. Term frequency "TF" is one of the development keys that was appeared in this period. TF is based on occurrence of words within documents. Inverse Document Frequency "IDF" and the combination between them also was appeared. Also, the Vector Space Model "VSM" was produced at this period. The 1970's and 1980's were a period when databases and office research flourished. Variation of TF*IDF were introduced. IRS locates information that is relevant to user's query. IRS searches in collection of unstructured, semi-structured data (web, documents, images, videos, etc.). With the growth of unstructured information, high speed of networks and rapid access to large amount of information, the only solution is to find relevant items from the large text database was search. In 1980's, stable increase in word processing and stable decrease in the price of disk space meant that more information was available in readable machine. From 1980's to mid-1990's, retrieval models are the basis of the core ranking function of IRS. Advances of Vector Space Model "VSM" was developed and latent semantic indexing (LSI). Introduction of probabilistic approach using language modes used to view matching process between documents and queries. The language Model "LM" provides understanding wide range of IR process such as (relevance feedback, clustering and term dependency).

In the middle of 1990's – until now, World Wide Web "WWW" in the late of 1990, number of websites and quality of pages were small until 1993. Web search engines started to appear in late of 1993 to cope with the growth. Most people use some types of modern IR such as google. Research on web IR focused on short queries, which have little linguistic structure. A lot of this work started with the question answering. Then progressed into more detailed questions. IR research has grown due to number of factors. Normal cycle of Google query functions is: (web server sends query to index server, query travels to DOC server which retrieves stored document, search result is eventually returned to the user in a fraction of second). Figure 1.1 shows the cycle of google query.

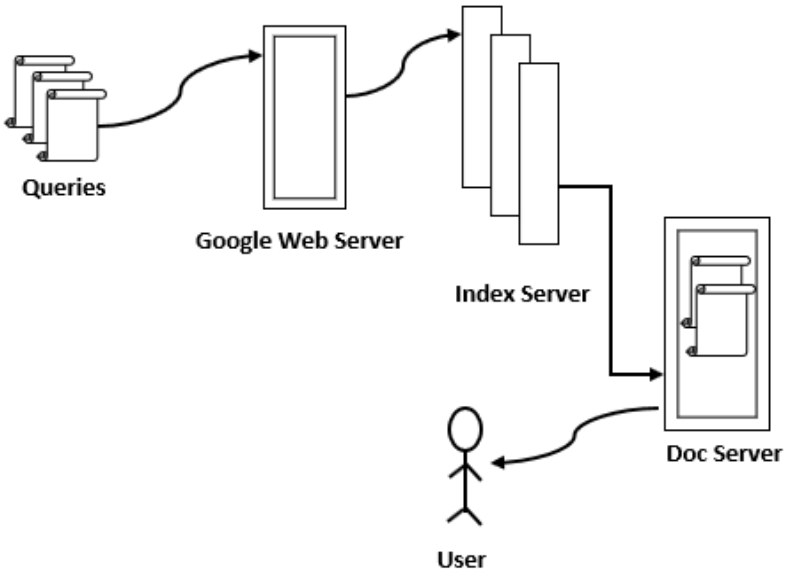

Figure 1.1: Cycle of Google Query.

An information retrieval system is a system that able to store, retrieve and maintain information, [1]. Information has many types such as text (include numeric and date data), audio, video and other multimedia. Information retrieval modeling is very important to help researchers in designing and implementing an actual efficient information system. Mathematical modeling can be used in several domains such as geographical areas, medical areas…. etc. The model of information retrieval helps the user to predict and explain what a user will find relevant given query. In Figure 1.2, the classification of mathematical models in IR is shown.

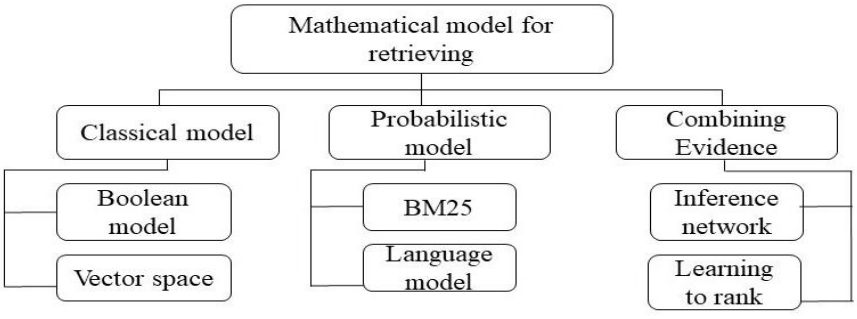

Figure 1.2: Mathematical modeling classification

Mathematical retrieval modeling is classified as classical models (Boolean and VSM), Probabilistic Models (BM-25 "Ranking Function in IR" and LM) and combining evidence models (inference network and language to rank models)

1.1.1. Classical Models

This group contains two models: Boolean and region models. In [2], these models provide exact matching. These models use logic operators (and, or, not). These operators are known as "intersection, union, difference". In Figure 1.3, the different Boolean model operations are shown. Their advantages are given exact match, but no ranking retrieved the document. There is a difference between them which is that region model is designed to search for semi-structured data.

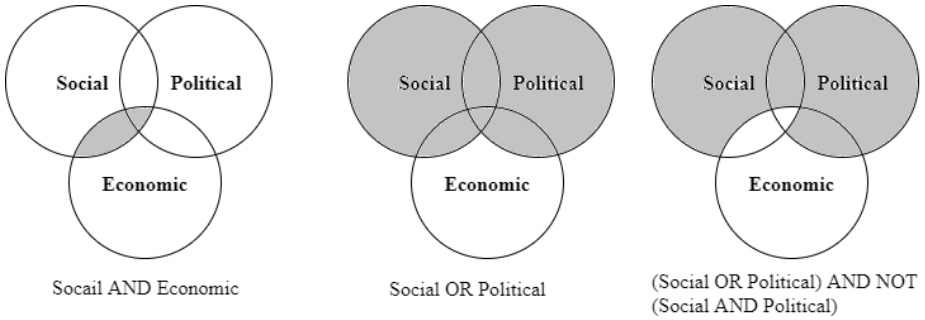

Figure 1.3: Boolean operation using Venn diagram [2].

The vector space model is used to calculate the similarity "distance" between documents and queries through "Euclidean, Manhattan, and cosine similarity. "If cosine angel is zero", then the vector is orthogonal else the degree is zero. The positioning of the query in vector space model is to calculate the centroid of the relevant and irrelevant documents. Moving query towards centroid point of relevant rather than irrelevant meaning it is improving retrieval performance. Both of term weighting's intuition and term's independence are considered the most common disadvantages.

1.1.2. Probabilistic Models

Probabilistic models use conditional probability and require two conditions, [2], relevant document and long queries. The relevant document is obtained through computing the probability that contains document terms. Long queries are used to distinguish between term's presence and term's absence in documents. The good choice is that both relevant and non-relevant documents are available. There is a difference