On Information Retrieval Using Co-word Analysis and Data Mining Techniques

by

Doaa Aboshady

Mathematics Department, Faculty of science, Ain Shams University

march 2018

Acknowledgments

First and most of all, I am very grateful to my creator almighty ***ALLAH***, the most Merciful and beneficent, To inspire and give me the strength and patience to end this work and compromise in it.

My deep thanks to ***Professor Dr. Faid Ghaleb***, ***Professor Dr. El-Sayed Atlam*** and ***Dr. Dowlat abd elaziz*** my Master. supervisors, for his guidance, his kind support, continuous encouragement and help throughout the accomplishment of these studies. I appreciate his in exhaustive efforts, unending cooperation and advice, his deep insight that always found solutions when problems supervened and very creative criticism.

My deep thanks to ***Dr. Mahmoud Badeea*** Assistant Professor of Computer Science, Mathematics Department, Tanta University, Egypt, for his guidance, his kind support and help. I will never forget his kindness and patience in dealing with me, his kind support and continuous encouragement.

My thanks and appreciation to all my colleagues and friends at the department of mathematics, Faculty of science, Ain Shams University, for help and support.

Also, it is a great honor for me to take this opportunity to express my sincere appreciation and my deep respect to all members of Mathematics Department, Faculty of Science, Tanta University, Egypt.

I am also grateful to my family (my beloved husband ***Hatem Hosny Hegab***, my beloved daughter ***Lili Hatem***) could not complete this work without the unwavering love, patience, prayer, supplication of them. My deep thanks to all of them for their patience and understanding during the many hours of working and preparing this thesis.

I would like to make a deep thank for my Parents (My beloved mother ***Mervat Elgaiar*** and my beloved father ***Hassan Aboshady*** for their support at all and their doa'a, my brothers ***Mohammed*** and ***Ahmed*** and my sister ***Basant***.

I am thankful to all who knows me from near or far and had wished me the success.

Abstract

In this thesis, the co-word analysis that counts and analyzes the co-occurrence of keywords in the publications on a given subject are used and its modification to measure the relations among a selected sample of FA terms in a common field.

This thesis is devoted to study the previous work of the Retrieval Precision (RP) and focuses on how to use the power link as a tool to improve the extracted field association terms from corpus by the proposed algorithm. The power link analysis is used to classify the scientific papers into its proper field, and is used as a quantitative tool to compute the co-citation relation between two words, depending on the co-frequency and distances among instances of the words.

This thesis proposes a modified method to produce an improvement FA terms dictionary, by using the co-word and Power link analysis. The modified method is used to calculate the levels of FA terms by giving different weights to terms according to their position in the document.

The proposed method is compared with the most recent methods to prove the validity and effectiveness. The obtained results show an improvement in precision, recall, and F-measure.

Summary

Information retrieval (IR) is the science of searching for information in documents, searching for document themselves, searching for metadata which describe documents, or searching within database, whether relational standalone or hypertext networked databases such as the internet or World Wide Web or internet, for text, sound, images or data. Field association terms (FA terms) are the terms that indicate each subject matter category in the classification scheme. In this thesis, co-word analysis that counts and analyzes the co-occurrence of keywords in the publications on a given subject will be used to measure the relations among a selected sample of FA terms in a common field.

The thesis objectives are the outline of information retrieval, co-word analysis, and power link. It is devoted to focus on the previous work of the Retrieval Precision (RP) and focuses on how to use the power link as a tool to improve the extracted field association terms from corpus by the proposed algorithm.

The thesis presents a modified method to produce an improvement FA terms dictionary by using the co-word and Power link analysis. The modified method is used to calculate the levels of FA terms by giving different weights to terms according to their position in the document.

The proposed method uses the power link concept as well as modifications of the rules to classify the scientific papers into its proper field. Instead of the whole document, a given document will be divided into three parts, namely the title, abstract, and body. A given term will be given a weight that depends on the location of the term in a specific document. The greatest weight will be given to the title, then the abstract, and then the body respectively. Results of used data show an improvement in precision, recall, and F-measure in perfect FA terms (Level 1), but with different data the proposed method can give an improvement in level 2 and level 3.

The thesis is organized into four chapters:

Chapter 1: Presents a review of definitions and concepts related to information retrieval, FA terms, co-word analysis, and presents the relation between these fields. Also, this chapter discusses the methods of IR system evaluation.

Chapter 2: Presents a review of the power link analysis, real word spell checker based on power links, and the main steps of this method and its applications in various fields. Also, it presents the traditional algorithm for calculating the levels of FA terms based on power link analysis and the methods to solve spelling errors by using the concept of power link. This survey reflects that the relation between these areas did not studied before.

Chapter 3: Presents the modified algorithm for calculating the perfect FA terms, and presents the Continuity and Transition theme to detect the different parts of every document. Also, it presents Python language that used to write a program for the code of the modified system. Finally, it presents the experiments applied to a set of documents (scientific researches) and the comparison between the traditional and proposed methods that presented in this chapter, which helps in evaluating the system.

Chapter 4: Concludes the thesis and lists important future work.

Contents

List of Tables

List of Figures

List of Publication

El-Sayed Atlam, Fayed Ghaleb, Dawlat A. El A.Mohamed, Doaa Abo-Shady, **"An Improvement of FA Terms Dictionary using Power Link and Co-Word Analysis"**, International Journal of Advanced Computer Science and Applications(IJACSA),Volume 9, Issue 2, 2018.

Chapter 1

Introduction

1.1 Information Retrieval

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). The term "unstructured data" refers to data which does not have clear, semantically overt, easy-for-a-computer structure. It is the opposite of structured data.

The IR system consists of a corpus, one or more indexes of its content, a query interface, a search system, and a results interface. IR did not begin with the Web. In response to various challenges of providing information access, the field of IR evolved to give principled approaches to searching various forms of content. IR is also used to facilitate "semistructured" search such as finding a document where the title contains Java and the body contains threading. IR covers other kinds of data and information problems. The field of IR covers supporting users in other tasks like: browsing, filtering, translation, summarization and further processing a set of retrieved documents.

1.1.1 IR Models

The documents are typically converted into a convenient representation to effectively retrieve relevant documents by IR strategies. Since each retrieval strategy incorporates a specific model for its document representation purposes, the models are classified according to two dimensions: the mathematical basis, and the properties of the model.

- The mathematical basis contains three kinds of models:

    1. Set-theoretic model: that represents documents as sets of terms (the words or phrase). The similarities are usually derived from set-theoretic operations on those sets. Common models are: standard Boolean model, Extended Boolean model and Fuzzy retrieval.
    2. Algebraic model: that represents documents and queries usually as vectors, matrices, or tuples. The similarity of the query vector and document vector as a scalar value. Common models are: Vector space model, Extended Boolean model, Generalized vector space model, and (Enhanced) Topic-based Vector Space Model.
    3. Probabilistic model: that treats the process of document retrieval as a probabilistic inference. The Similarities are computed as probabilities that a document is relevant for a given query. Probabilistic theorems like the Bayes' theorem are often used in these models. Common models are: binary Independence Model, uncertain inference, Language models, Divergence-from-randomness model and Latentnt Dirichlet allocation.

- Properties of models:

  Properties of models are represented without term-interdependencies, but treat different terms/words as independent. This fact is usually represented in vector space models by the orthogonality assumption of term vectors or in probabilistic models by an independency assumption for term variables. These Models allow a representation of interdependencies terms between each other. However the degree of the interdependency between two terms is defined by the model itself. It is usually directly or indirectly derived (e.g. By dimensional reduction) from the co-occurrence of those terms in the whole set of documents. In addition, these models do not allege how the interdependency between two terms is defined. They rely an external source for the degree of interdependency between two terms (e.g., a human or sophisticated algorithms).

1.1.2 Evaluation Measures

To evaluate the effectiveness of an IR system (i.e., the quality of the system), there are many different measures for evaluating the performance of infor-

mation retrieval systems. In general, measurement depends on a collection of documents to be searched and a search query. Every document is known to be either relevant or non-relevant to a particular query (relevancy) and in practice queries may be ill-posed and there may be different shades of relevancy.

- **Precision:** is the fraction of retrieved documents that are relevant to the information need.

$$Precision = \frac{no.\ of\ relevant\ documents\ retrieved}{no.\ of\ retrieved\ documents} \tag{1.1}$$

  Also, precision is called analogous to positive predictive value In binary classification. Instead of Precision takes all retrieved documents into account, It can be evaluated at a given cut-off rank, considering only the topmost documents returned by the system. This measure is called precision at n or Pan.

- **Recall:** is the fraction of relevant documents that are retrieved to the query that are successfully retrieved.

$$Recall = \frac{no.\ of\ relevant\ documents\ retrieved}{no.\ of\ relevant\ documents} \tag{1.2}$$

  Also, recall is called sensitivity in binary classification. So it can be defined as the probability that a relevant document is retrieved by the query.

  This evaluation measure alone is not enough but one needs to consider the number of non-relevant documents also, for example by computing the precision. Fig.1.1 presents the deferent between retrieved and relevant document.

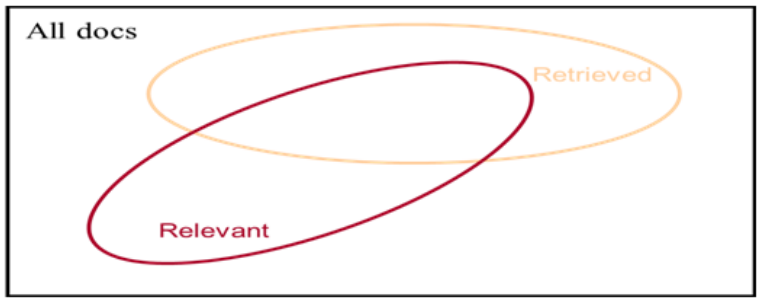

Figure 1.1: Retrieved and Relevant Document.

- **Fall-out:** is the proportion of non-relevant documents that are retrieved, out of all non-relevant documents available.

$$Fall - out = \frac{no.\ of\ non\ relevant\ documents\ retrieved}{no.\ of\ non\ relevant\ documents} \tag{1.3}$$

Also, fall-out is closely related to specificity in binary classification. More precisely: fall-out = $(1 - specificity)$. It can be defined as the probability that a non-relevant document is retrieved by the query.

- **F-measure:** is a single measure that trades off precision versus recall is the F measure, which is the weighted harmonic mean of precision and recall:

$$F = \frac{2 * P * R}{P + R} \tag{1.4}$$

Here, recall and precision are evenly weighted so This formula, also known as the F_1 Measure.

$$F_1 = \frac{(1 + \beta^2) * (P * R)}{\beta^2 * (P + R)} \tag{1.5}$$

This is The general formula for non-negative real β that depended on The F-measure so that F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision" [46]. It is based on van [46] effectiveness measure $E = 1 - (1/\alpha/p + (1 - \alpha)/R)$. Their relationship is $F_\beta = 1 - E$ where $\alpha = 1/(\beta^2 + 1)$. [26]

- **Average Precision of Precision and Recall:** is the precision and recall are based on the whole list of documents returned by the system. Average precision is an average of precisions computed after truncating the list after each of the relevant documents in turn. It is confirmed returning more relevant documents earlier.

$$AvgP = \frac{\sum_{r=1}^{N}(P(r) \times rel(r))}{number\ of\ relevant\ documents} \tag{1.6}$$

Where $P(r)$ precision at a given cut-off rank, r is the rank, $rel(r)$ a binary function on the relevance of a given rank and N the number retrieved.