Computer Science Department
Faculty of Computer & Information Sciences
Ain Shams University

Designing and Building a Retrieval System for Arabic Documents

Thesis submitted as a partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences.

By

Sally Saad Mohamed Ismail

B.Sc. in Computer and Information Sciences,
Demonstrator at Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

Under Supervision of

Prof. Dr. Moustafa Mahmoud Aref
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Shaimaa Arafat
Faculty of Computer and Information Sciences,
Ain Shams University

May 31, 2018

ACKNOWLEDGEMENT

After praising ALLAH for everything HE gave me and for all the inspiration used for this work to show out, I would like to express sincere appreciation to my parents; the late General Saad Mohamed Ismail and the Physiotherapist Azza Mohamed Ezzat; for their love and care in raising me up during my years in Graduate School. I would like also to thank my husband; Eng. Mohamed Ezzat; for his enormous support and encouragement for finishing my degree.

My gratitude is for my supervisors; Prof. Dr. Moustafa Aref; who really made a great effort with me in order to finalize my research and taught me a lot of stuff not only in work but also in life and gave me self-confidence; as well as Dr.Shaimaa Arafat; who was the initiator helper in our published papers. Also, special thanks for the late Prof. Dr. Moustafa Syam who was the originator of this research.

Special Thanks for my brother and sisters who really cared for me.
Finally; my friends and colleagues; specially Bedour; who helped me a lot in collecting the Arabic suffixes, prefixes and stop-words from the language grammar; and Wael who sincerely assisted me in completing the steps to my degree.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF TABLES

LIST OF FIGURES

Publications

1. Shaimaa Arafat, Sally Saad. *An Affix Removal Stemming Algorithm for Arabic Language*. International Journal on Intelligent Computing and Information Systems (IJICIS), Volume 8, No 2, pp 141-153, Egypt, July 2008.

2. Sally Saad, Mostafa Aref. *Applying Vector Model for Arabic Information Retrieval* ,The Fourth International Conference for Intelligent Computing and Information Systems (ICICIS 2009), pp 755-762, Faculty of Computer and Information Sciences, Ain Shams University , Egypt, March 2009.

Chapter 1

Introduction

Chapter One

Introduction

"But do you know that, although I have kept the diary [on a phonograph] for months past, it never once struck me how I was going to find any particular part of it in case I wanted to look it up?" said Dr. Seward, from the story of Dracula, since 1897 by Bram Stoker [1]. This statement reveals the problem a person might have when he has a lot of information and needs to retrieve a part of it. It's the very early need for an Information Retrieval (IR) System that will explore uncovered information, and satisfy the user needs.

Automated IR systems are used to reduce information overload. Many universities and public libraries use IR systems to provide access to books, journals, and other documents. Web search engines such as Google, Yahoo search and Live Search (formerly MSN Search) are the most visible IR applications.

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevance.

An object is an entity which keeps or stores information in a database or a documents corpus. User queries are matched to objects stored in the database. Depending on the application the data objects may be, for example, text documents, images or videos.

Arabic is one of the most important languages in history, not only for the civilization of old Arabs and their leadership in many fields yielding to scientific books that are considered a treasure, but also due to its use in the Quran; the holy book of Islamic Religion; one of the most widespread religions nowadays.

English IR systems have been adopted and worked for since more than fifty years. Whilst Arabic language has been ignored till recently, where there's not much work. So far, most of the researchers in Arabic IR are non-Arabs, so that was a good motivation for us to conduct this research hoping that it helps in the field.

1.1 Overview on Information Retrieval

The problem of information storage and retrieval has attracted increasing attention. It is simply stated: we have vast amounts of information to which accurate and speedy access is becoming ever more difficult. One effect of this is that relevant information gets ignored since it is never uncovered, which in turn leads to much duplication of work and effort.

Information retrieval is the science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational stand-alone databases or hyper textually-networked databases such as the World Wide Web.

There is a common confusion, however, between data retrieval, document retrieval, information retrieval, and text retrieval. Each of these has its own bodies of literature, theory, praxis and technologies. IR is

interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics and physics.

The simplest IR system that can come to one's mind when he's thinking of searching for documents; is Keyword Search. It matches the query keywords with the document keywords. In this kind of search relevancy can be either the query string appears exactly in the document, or a less strict notion; the words in the query appear frequently in the document, in any order (bag of words).

Keyword Search has many problems. It might not retrieve relevant documents that have words of the same root of the query terms; like "talking, talks, talked" from the root "talk". The same is for documents having synonymous words with the query terms; like "restaurant" and "café". Another Problem with Keyword search is that it might retrieve irrelevant documents due to the use of ambiguous words; i.e. words having more than a meaning. For example, searching for the word "apple"; does the user require documents talking about fruits or documents talking about "Apple" company?

An Intelligent IR system can overcome these problems by taking into account one or more of the following; the order of words in the query, the roots of the words in the query (Stemming), the user feedback and adapting to it (Relevance Feedback), the meaning of the words used (Using Thesaurus), and the authority of the source.

Relevance of the retrieved documents in Intelligent IR system may include; being on the proper subject, being timely (recent information),

being authoritative (from a trusted source), and satisfying the goals of the user and his/her intended use of the information (Information Need).

Most IR systems compute a numeric score on how well each object in the database match the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query [35].

IR applications can be classified into mono-lingual or cross-lingual. Mono-lingual are the standard and most common applications; where the user submits his search string in the same language of the retrieved documents; while Cross-language information retrieval (CLIR) is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query. For example, a user may pose his query in English but retrieve relevant documents written in French [36]. Our research is of the 1st type; i.e. the user gives his query in Arabic such as the language of the documents he's searching for.

Arabic IR is considered one of the open areas for research nowadays, due to the great importance of the Arabic language as well as the increase of Arabic documents specially over the internet.

Arabic is currently the sixth most widely spoken language in the world. The estimated number of Arabic speakers is 250 million; of which roughly 195 million are first language speakers and 55 million are second language speakers. Arabic is an official language in more than 22 countries. Since it is also the language of religious instruction in Islam, many more speakers have at least a passive knowledge of the language [2].

As late as the year 2002, the Arabic language wasn't one of the most common languages over the internet. The percentage of public sites was for