Chapter II
An Enhanced Text-Classification-Based Arabic Information Retrieval System

Sameh Ghwanmeh
Yarmouk University, Jordan

Ghassan Kanaan
The Arab Academy for Banking and Financial Sciences, Jordan

Riyad Al-Shalabi
The Arab Academy for Banking and Financial Sciences, Jordan

Ahmad Ababneh
The Arab Academy for Banking and Financial Sciences, Jordan

ABSTRACT

This chapter presents enhanced, effective and simple approach to text classification. The approach uses an algorithm to automatically classifying documents. The main idea of the algorithm is to select feature words from each document; those words cover all the ideas in the document. The results of this algorithm are list of the main subjects founded in the document. Also, in this chapter the effects of the Arabic text classification on Information Retrieval have been investigated. The goal was to improve the convenience and effectiveness of information access. The system evaluation was conducted in two cases based on precision/recall criteria: evaluate the system without using Arabic text classification and evaluate the system with Arabic text classification. A chain of experiments were carried out to test the algorithm using 242 Arabic abstracts From the Saudi Arabian National Computer Conference. Additionally, automatic phrase indexing was implemented. Experiments revealed that the system with text classification gives better performance than the system without text classification.

INTRODUCTION

Information-retrieval systems process files of records and requests for information and identify and retrieve from the files certain records in response to the information requests. The retrieval of particular records depends on the similarity between the records and the queries, which in turn is measured by comparing the values of certain attributes of records and information requests. An Information Retrieval System is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text (including numeric and date data), images, audio, video, and other multi-media objects (Jingho and Tianshun, 2002; Salton, 1989).

Classification is the mean whereby we order knowledge. Classification is fundamentally a problem of finding sameness. When we classify, we seek for group things that have a common structure or exhibit a common behavior. Text classification mainly uses information retrieval techniques. Traditional information retrieval mainly retrieves relevant documents by using keyword-based or statistic-based techniques (Jingho and Tianshun, 2004). A standard approach to text categorization makes use of the classical text representation technique that maps a document to a high dimensional feature vector, where each entry of the vector represents the presence or absence of a feature (Lodhi, 2002).

Text Classification is the problem of assign documents to predefine classes or categories. The approaches to topic identification can be summarized in groups: statistical, knowledge-based, and hybrid. The statistical approach infers topics of texts from term frequency, term location, term co-occurrence, etc, without using external knowledge bases such as machine readable dictionaries. The knowledge-based approach relies on a syntactic or semantic parser, knowledge bases such as scripts or machine readable dictionaries, etc., without using any corpus statistics. The hybrid approach combines the statistical and knowledge-based approaches to take advantage of the strengths of both approaches and thereby to improve the overall system performance (Jingho and Tianshun, 2002).

This chapter presents enhanced, effective and simple approach to text classification. The approach uses an algorithm to automatically classifying documents. The main idea of the algorithm is to select feature words from each document; those words cover all the ideas in the document. The results of this algorithm are list of the main subjects founded in the document. Also, in this chapter the effects of the Arabic text classification on Information Retrieval have been investigated. The goal was to improve the convenience and effectiveness of information access. The system evaluation was conducted in two cases based on precision/recall criteria: evaluate the system without using Arabic text classification and to evaluate the system with Arabic text classification. A series of experiments were carried out to test the algorithm using 242 Arabic abstracts From the Saudi Arabian National Computer Conference. Additionally, automatic phrase indexing was implemented. The system was evaluated for the two cases based on precision/recall evaluation as shown before. Experiments reveal that the system with text classification gives better performance than the system without text classification.

GENERAL APPROACHES TO CLASSIFICATION

Classical Categorization

In this approach all entities that have a given property or a collection of properties in common form a category. Classical categorization comes to us from Plato, then from Aristotle through his classification of plants and animals (Joachims, 1999).