Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 1538–1545

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

An auto-indexing method for Arabic text

Nashat Mansour [*], Ramzi A. Haraty, Walid Daher, Manal Houri

Division of Computer Science and Mathematics, Lebanese American University, P.O. Box 13-5053, Chouran, Beirut 1102 3801, Lebanon

Received 4 July 2007; received in revised form 18 December 2007; accepted 29 December 2007
Available online 8 February 2008

Abstract

This work addresses the information retrieval problem of auto-indexing Arabic documents. Auto-indexing a text document refers to automatically extracting words that are suitable for building an index for the document. In this paper, we propose an auto-indexing method for Arabic text documents. This method is mainly based on morphological analysis and on a technique for assigning weights to words. The morphological analysis uses a number of grammatical rules to extract stem words that become candidate index words. The weight assignment technique computes weights for these words relative to the container document. The weight is based on how spread is the word in a document and not only on its rate of occurrence. The candidate index words are then sorted in descending order by weight so that information retrievers can select the more important index words. We empirically verify the usefulness of our method using several examples. For these examples, we obtained an average recall of 46% and an average precision of 64%.

Keywords: Arabic text; Document auto-indexing; Information retrieval; Stem words; Word spread

1. Introduction

Indexing text documents refers to selecting some words that represent the content of a document. The selected words are referred to as index words. Manual indexing of text documents is considered to be a cumbersome task in information retrieval. The people who perform indexing are usually well trained and have reasonable linguistic background. Manual indexing requires intensive human effort, since it requires people to read the whole document before selecting the candidate index words for that document.

There are two types of indexing: thesaurus based indexing and full-text based indexing. In thesaurus based indexing, the index words selected to represent a document might not exist in the document; but, their synonyms must exist. In this case, the index words are selected based on prior knowledge of what words might be searched for by users. In contrast, full-text based indexing is based on words found within the document itself. However, for both types of indexing, the output is a set of index (key) words from which an index for the

[*] Corresponding author. Tel.: +961 3 379647; fax: +961 1 867098.
E-mail addresses: nmansour@lau.edu.lb (N. Mansour), rharaty@lau.edu.lb (R.A. Haraty), daherwalid@hotmail.com (W. Daher), manoula78@hotmail.com (M. Houri).

0306-4573/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.12.007

document can be constructed. These key words can also be used to define subject headings. Subject headings are phrases composed of more than one keyword. A single document may have many subject headings. The more accurate subject headings are, the more likely it will be for a user to hit that document upon searching for a topic in an information retrieval system.

Auto-indexing refers to automatic selection of key words in a text document. This problem varies in difficulty depending on the language used. Every language is characterized by its syntax, logical structure, and its domain (Harter, 1986). In particular, languages with sophisticated grammatical rules, such as Arabic, require sophisticated indexing methods.

A number of methods have been reported in the literature on related subjects. Classification algorithms for Arabic text have used the *N*-gram frequency statistic technique (Khreisat, 2006). An Arabic part-of-speech tagger that uses statistical and rule-based techniques has been proposed (Khoja, 2001). These parts-of-speech are further divided into nouns, verbs, and particles. The author uses a stemmer to remove all of a word's affixes to produce the stem. The stemmer faced problems when certain letters that appear to be affixes are part of the word, and when certain letters change to other letters when an affix is added. Gawrysiak, Gancarz, and Okoniewski (2002) describe the unigram and *N*-gram document representation techniques that are used frequently in text mining and discuss their shortcomings. They also present a technique that considers the word's position within a document. In Larkey and Connell (2001), the authors implement some standard approaches to handle co-occurrences. They also present several monolingual and cross-language runs and show that stemming improves their system. Term co-occurrence data are also used in Billhardt, Borrajo, and Maojo (2000) for document indexing. Rachidi et al. (2003) describe a multilingual Internet search engine that includes Arabic word stemming. Stemming is performed by removing postfix, infix, and prefix parts of words and by reducing singular, plural, and adjective forms into a canonical form. Larkey, Ballesteros, and Connell (2002) have developed light stemmers based on heuristics and a statistical stemmer based on co-occurrence for Arabic retrieval. Thus, very limited work has been reported on extracting indices for Arabic text.

In this paper, we present a full-text based auto-indexing method for Arabic text documents. Our auto-indexing method consists of three phases. The first phase consists of the processing steps: (a) apply rhyming step to classify all words; (b) determine and exclude stop-list words and phrases; (c) identify verbs and nouns. The second phase is concerned with extracting stem words from different types of verbs and nouns, which leads to determining candidate index words. The third phase is concerned with computing weights for the candidate index words relative to their document, which leads to selecting appropriate index words for the document. The usefulness of our method is demonstrated using empirical work involving several Arabic texts. For brevity and to serve general readers, we omit detailed language analysis.

The rest of the paper is organized as follows. Section 2 describes the preprocessing steps. Section 3 presents the word stemming algorithm. Section 4 describes how the weight of a word is calculated and discusses index word selection. Section 5 presents our experimental results. Section 6 contains the conclusions.

2. Phase 1 – preprocessing steps

Before extracting stem words, the first phase is a preparatory phase that involves preprocessing steps, which are described in the following subsections.

2.1. The rhyming step

The rhyming step preprocesses words by classifying them before stemming. Every word in the document is examined to decide whether it is a noun or a verb, whether a noun is in its singular form or plural form, whether a verb is in its past, present or future tense, and whether pronouns are attached to a word. In this step, every word is compared to an appropriate set of predefined rhythms. There are different sets of rhythms in Arabic. For example, the set of rhythms for used to decide whether a noun in singular or plural is different from the set for determining the attached pronoun. However, all words are rhymed with the derivations of the word 'Fa'ala' (i.e., did) and are marked with the relevant rhythm. The rhythms of the verb 'Fa'ala' are standard in the Arabic grammar.