Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 41 (2005) 819–827
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Character contiguity in *N*-gram-based word matching: the case for Arabic text searching

Suleiman H. Mustafa

Department of Computer Information Systems, Faculty of Information Technology, Yarmouk University, Irbid, Jordan

Received 7 October 2003; accepted 6 February 2004
Available online 25 March 2004

Abstract

This work assesses the performance of two *N*-gram matching techniques for Arabic root-driven string searching: contiguous *N*-grams and hybrid *N*-grams, combining contiguous and non-contiguous. The two techniques were tested using three experiments involving different levels of textual word stemming, a textual corpus containing about 25 thousand words (with a total size of about 160KB), and a set of 100 query textual words. The results of the hybrid approach showed significant performance improvement over the conventional contiguous approach, especially in the cases where stemming was used. The present results and the inconsistent findings of previous studies raise some questions regarding the efficiency of pure conventional *N*-gram matching and the ways in which it should be used in languages other than English.

Keywords: *N*-grams; String matching; Text searching; Stemming; Word conflation

1. Introduction

N-gram matching is one of the main techniques in probabilistic string matching. It has been used for a variety of information retrieval applications, including index term conflation, text searching and retrieval, text categorization, image-based text retrieval, degraded text recognition, and spoken-document retrieval.

Tauritz (2002) has pointed out 13 areas of application in which *N*-gram matching has been used during the last 50 years. The technique is described as being robust, complete, domain independent, efficient, and simple. It has been well established within the information retrieval (IR) research community that, in certain cases (such as generalized multilingual systems involving no customization for particular languages), *N*-gram matching performs as good as, or even better than, word-based techniques. It has also been shown to be effective for many languages.

There are relatively few studies on the retrieval of Arabic documents in the literature. The lack of a realistically large test corpus has been a problem in past studies on Arabic retrieval (Xu, Fraser, & Weischedel, 2002). Furthermore, a claim is sometimes raised that some statistical NLP techniques for IR on

E-mail address: smustafa@yu.edu.jo (S.H. Mustafa).

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.02.003

European languages do not transfer well to Arabic because of the complexities involved in the morphological and orthographic structure of the language (Goweder & De Roeck, 2001).

Over the last few years, a number of researchers have addressed this issue and investigated various statistical techniques and strategies for Arabic text retrieval and for some other applications, including speech recognition (Kirchhoff et al., 2002), OCR-degraded text recognition (Darwish, 2003), and document classification (Sawaf, Zaplos, & Neyt, 2001).

Besides, there has been some effort to initiate an Arabic corpus for experimental purposes such as the TREC 2001 test suit, which included 383,872 newspaper articles (Larkey, AbulJaleel, & Connell, 2002), and the corpus described by Goweder and De Roeck (2001), which included 42,591 newspaper articles. Apart from these cases, most researchers have constructed their own data sets.

Statistical language models offer a possible alternative to traditional morphology-based stemming techniques. Typically, Arabic stemming algorithms operate by "trial and error". Likewise, morpho-syntactic approaches have limited scope for deployment in IR. Even if substantial, their morpho-syntactic coverage remains limited and processing efficiency implications are often unclear (De Roeck & Al-Fares, 2000).

Statistical methods can provide a more language-independent approach to conflation. Related words can be grouped based on various string similarity measures. Such approaches often involve N-grams (Larkey, Ballesteros, & Connell, 2002). This paper examines the performance of N-gram matching as a basis for Arabic term clustering. Given a text T, the paper addresses the problem of finding all occurrences of words sharing the same root in T.

Typically, one slices a string into a set of contiguous N-grams. However, the term N-gram as defined in the literature can include the notion of any co-occurring set of characters in a string such as an N-gram made of the first and third character of a word (Cavnar & Trenkle, 1994). The main objective of the research, reported in this paper, was to test the impact of N-gram character contiguity on the overall performance as measured by clustering recall and precision.

The notion of using non-adjacent characters in N-gram computation is not new. The issue was raised earlier by Sawaf et al. (2001), who pointed out that: "gap-n-grams" should be investigated for Arabic, so that the more abstract level of morphological analysis can be reached. Pirkola, Keskustalo, Leppanen, Kansala, and Jarvelin (2002) have also reported using this strategy in a cross-lingual study. In their technique (which they termed "targeted S-gram"), N-grams were classified into categories on the basis of character contiguity in words. The results indicated that the S-gram technique outperformed the conventional N-gram matching technique.

Arabic morphology involves a complex infix structure. Word infixes may occur in two places: after the first root radical, and before the last root radical. This might lead to an assumption that: adjacent N-grams might fail to capture the similarity between related infixed words. If this were the case, we would expect that non-contiguous N-grams should improve the overall performance of N-gram matching.

2. *N*-gram matching and Arabic

It has been sometimes assumed that some of the statistical techniques that have widely been applied to many languages cannot be expected to perform well on languages like Arabic in which suffixing is not the only inflectional aspect (Larkey et al., 2002). Given this assumption, a number of research studies have been carried out during the last three years or so. An overall look at the experimental results of these studies points out some differences. From an information retrieval perspective, these studies fall into two categories: the first focuses on applying statistical techniques for string searching and term conflation, while the other focuses on document retrieval as a basis for N-gram matching.