IJCSI International Journal of Computer Science Issues, Volume 14, Issue 1, January 2017
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.20943/01201701.115118

The effect of N-gram indexing on Arabic documents retrieval

Emad Fawzi Al-Shalabi

Department of Information Technology, AL-BALQA Applied University
Al-Huson University College, Irbid, Al-Huson, 50, Jordan

Abstract

This article presents a comparison between 3-gram and 4-gram term indexing in Arabic document retrieval. The calculation of similarity between query and documents is performed using single term and two term query, based on corpora of Arabic language documents collected from Arabic news websites available online.

Keywords: *n-gram, Arabic text indexing, information retrieval, text similarity***.**

1. Introduction

In the past several years, due to the increase in availability of Arabic documents, significant attention has been directed towards Arabic Information Retrieval (IR). The main research focus in this domain has been concerned with formal language retrieval in news documents, OCR-based document retrieval and cross-language retrieval. Other aspects of Arabic IR have also received attention in the past years, including document image retrieval, social media and web search services, and speech search.

However, the currently existing efforts in Arabic language IR techniques, continue to be deficient in many aspects and are still far behind the efforts that have already been made in IR concerned with other languages.

The rest of this article is organized as follows: section 2 describes related works and past attempts in the IR field. Section 3 presents the problem statement. In section 4, the methodology and proposed approach are explained. Experimental work and practical results are shown in section 5. Finally, the proposed approach is concluded in section 6.

2. Related work

One of the most commonly used approaches in English to Arabic translations is N-gam techniques [3], the same approach has also been widely used in continuous speech recognition [2]. An n-gram [6] can be defined as a set of n repeated characters extracted from a single word. Following this approach, similar words will produce a high percentage of n-grams in common. Typical values of n used in the n-gram approach are 2, 3 and 4, that is, 2-gram, 3-gram and 4-gram respectively. This approach was proposed by [5]. As an example, the word (البيانات) following the 4-gram approach will result in the generation of { البي, لبيا, يانا, انات}, while following the 3-gram approach will result in generating gram {الب, لبي، بيا, يان, انا, نات}.

Other techniques include String-similarity [8] approaches to conflation, which involve the system calculating a measure of similarity between the input query term and each of the distinct terms in the document. Conflation [4], is a computational procedure that is designed to bring together words that are semantically related, before reducing them to a single form for retrieval purposes. This process can be divided into two main classes: stemming algorithms [9]; which are language dependent algorithms designed to handle morphological variants, and string-similarity algorithms, which are (usually) language independent algorithms that are designed to handle all types of variants.

For example, the word سيارة (car) may appear in text سيارات (cars). Multi forms of a given word, as a result of addition of different prefixes and suffixes. One way to improve this problem is to use a conflation algorithm, reducing them to a single form for retrieval purposes.

3. Problem Statement

One of the main challenges faced when attempting to conduct the existing IR and clustering approaches on Arabic text documents, is the variation in word forms that are likely to encountered. The most common type of these variations is referred to as different morphological variations where variations of the word do not present enough sharp differences in order to render them distinguishable from one another.

In this article, a comparison is made between two types of n-gram techniques for the purpose of performing IR on Arabic language documents, where the calculation of similarity between input query and documents is conducted using single term and two term query,.

4. Methodology

The proposed methodology consists of using n-gram algorithms for indexing the document. The process of text matching is started as flows:

- Arabic words found in each document are passed through a normalization phase.
- An index file is generated for each document using the 3-gram and 4-gram techniques
- The indexed documents generated previously are stored in a new separate file named documens.txt
- The similarity between the indexed documents and query is then found.
- The use of 3-gram and 4-gram approaches is experimented on Arabic document retrieval, and a comparison is conducted between the two approaches.

Prior to the generation of 3-gram or 4-gram terms from the set of documents, Arabic terms found in the document are normalized using single statement in Perl programming language [5].

$s =~ tr / آأؤإ / ااوا /;

5. Experimental work

5.1 The 4-gram approach

Start extracting terms with 4 letters long from each string in the corpora set, then store the extracted terms in a database file for each document after removing all duplicates,

then perform string-matching between query terms and the documents terms.

The query sentence also processed using 4-gram, and stored in a separate file with calculations of the term frequency for each term in the query.

Normalized frequency used [4] to calculate term frequency for each term in the documents database.

$$tf = \frac{tf}{\max tf} \quad (1)$$

Then the weights *w* computed using inverse document frequency *idf*.

$$idf = \log\frac{N}{n} \quad (2)$$

$$w = tf \times \log\frac{N}{n} \quad (3)$$

Similarity [4] between documents and query computed as the inner product between query and documents.

$$sim(d,q) = \sum(w_d \bullet w_q) \quad (4)$$

According to 4-gram generated documents we use a query consist of two terms in Arabic language {مباريات رياضية}.

Start divide the query into terms of 4 characters' table (1) and measure the similarity between the query and documents using Eq. (4).

Table 1: query terms using 4-gram

Term	Frequency	*tf*
اريا	1	1.000
اضية	1	1.000
باري	1	1.000
ريات	1	1.000
رياض	1	1.000
مبار	1	1.000
ياضي	1	1.000
max	1	

The results for similarity measure were 5 relevant documents shown in table (2).

Table 2: 4-gram relevant documents arranged descending

Document	Similarity
d1034gram.ngr	0.388674
d84gram.ngr	0.018656
d24gram.ngr	0.018074
d64gram.ngr	0.011238
d44gram.ngr	0.005054

Document {d103} is the relevant document and we can see that it's the first document in the list because it has the highest similarity measure.

5.2 The 3-gram approach

The same Eq. (1,2,3,4) used to compute the similarity for 3-gram, using the same query {مباريات رياضية}, using 3-gram indexing.

- Start by dividing the query into terms of 3 characters' table (3).

Table 3: query terms using 3-gram

Term	Frequencies	*tf*
اري	1	0.5000
اضي	1	0.5000
بار	1	0.5000
ريا	2	1.0000
ضية	1	0.5000
مبا	1	0.5000
يات	1	0.5000
ياض	1	0.5000
max	2	

- The similarity is measured between the query and documents using Eq. (4).

The results for the query using 3-gram indexing were 11 documents retrieved, but only one is relevant.

Table 4: 3-gram relevant documents arranged descending

Document	Similarity
d93gram.ngr	0.006033
d1033gram.ngr	0.004586
d73gram.ngr	0.003674
d23gram.ngr	0.002434
d83gram.ngr	0.001952
d33gram.ngr	0.001870
d43gram.ngr	0.001611
d63gram.ngr	0.000346
d13gram.ngr	0.000185
d103gram.ngr	8.69E-05
d53gram.ngr	7.14E-05

From table (4) we can see that the relevant document {d103} is not on the top of the rank, it's the second one.

5.3 Evaluation

5.3.1 Precision [4] for two term query

The first experiment done using two term query, the results of computing precision shown in table (5) and presented using line chart fig (1).

Table 5: Precision two term query

	Precision	
Query	4-gram	3-gram
نهائيات اسيا	0.200	0.100
ركلات الترجيح	0.111	0.100
قانون الإنترنت	0.286	0.200
النشاط الإشعاعي	0.167	0.091
شركات التأمين	0.200	0.100
محاسبة التكاليف	0.111	0.091

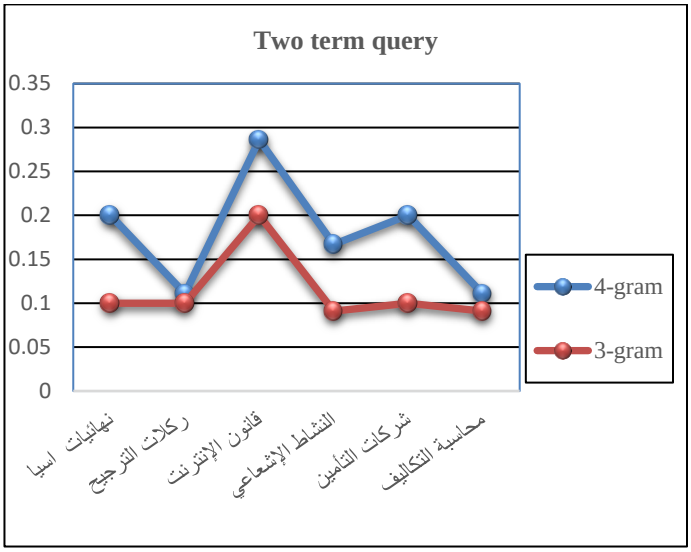

Fig. 1 Precision for two term query

It can be seen from fig (1), that a higher precision has been achieved using 4-gram indexing. The average precision for 4-gram was found to be 0.179 while for 3-gram it was found to be 0.114, using two term query.

5.3.2 Precision for Single word query

The second experiment done by using single term query as shown in table (6) And presented using line graph fig (2).

Table 6: Precision single term query

	Precision	
Query	4-gram	3-gram
استثمار	0.400	0.250
بروتوكول	0.950	0.500
الوزارة	0.500	0.200
نويدات	0.500	0.091
مونتريال	0.500	0.091
ديناميكي	0.250	0.091

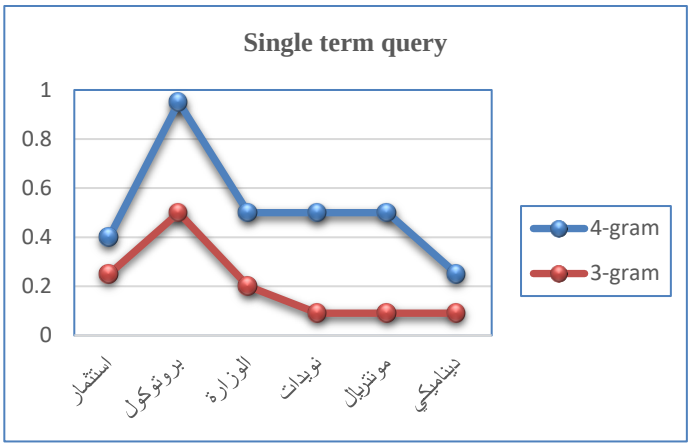

Fig. 2 Precision for single term query

From fig (2) we can see that single term query using 4-gram get the higher precision.

6. Conclusion

In this article, a comparison between two types of text indexing 3-gram and 4-gram has been conducted, the two approaches have been experimented on Arabic documents. As a result, the process of document retrieval has been evaluated using the proposed approaches.

The evaluation results have been predicted for two types of query, single word and two-word query. It has been concluded that following the 4-gram approach for Arabic indexing is far superior to the 3-gram approach, that is, for terms with at least 4 characters long.

Through the experimental results, it has also been shown that 4-gram is more efficient using single term query, by using 4-gram indexing Arabic stop words are removed, without using a special algorithm for stop word removal, this is because the length of most Arabic stop words is less than 4 character, i.e. { من, الى, عن, على, في, حيث }.

References:

[1] El-Halees, Alaa M. "Arabic text classification using maximum entropy." IUG Journal of Natural Studies 15.1 (2015).
[2] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013.
[3] Zaghouani, Wajdi, et al. "Building an arabic machine translation post-edited corpus: Guidelines and annotation." International Conference on Language Resources and Evaluation (LREC 2016). 2016.
[4] Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. Modern information retrieval. Vol. 463. New York: ACM press, 1999.
[5] Orwant, Jon, Jarkko Hietaniemi, and John Macdonald. Mastering algorithms with Perl. " O'Reilly Media, Inc.", 1999.
[6] Cavnar, William B., and John M. Trenkle. "N-gram-based text categorization." Ann Arbor MI 48113.2 (1994): 161-175.
[7] Ababneh, Jafar, et al. "Vector space models to classify Arabic text." International Journal of Computer Trends and Technology (IJCTT) 7.4 (2014): 219-223.
[8] Gomaa, Wael H., and Aly A. Fahmy. "A survey of text similarity approaches." International Journal of Computer Applications 68.13 (2013).
[9] Al-Shalabi, Emad Fawzi. "An Automated System for Essay Scoring of Online Exams in Arabic based on Stemming Techniques and Levenshtein Edit Operations." arXiv preprint arXiv:1611.02815 (2016).

Author Emad Fawzi Al-shalabi received the B.S. degree in Information Technology, Management Information Systems from Philadelphia University , Jordan in 2004, and the MSc in Computer Information Systems from Yarmouk University (YU), Jordan in 2009.