Improving Arabic Information Retrieval System using n-gram method

RAMMAL MAHMOUD
Legal Informatics center
Lebanese University
Sami Solh Street- Bp5396/116
LEBANON
mrammal@ul.edu.lb

SANAN MAJED
Faculty of Science
Lebanese University
Hadath compus
LEBANON
Sinane80@hotmail.com

ABSTRACT. This paper presents the application of the indexing method and the Retrieval systems based on N-grams to the Arabic legal language used in official Lebanese government journal documents. In our work we have used N-gram as a representation method, based on words and characters, and then compared the results using the vector space model with three similarity measures: the TF*IDF weighting, Dice's coefficient and the Cosine Coefficient.
The experiments demonstrate the use of trigrams to index Arabic documents is the optimal choice for Arabic information retrieval using N-grams. But using N-grams to indexing and retrieval legal Arabic documents is still insufficient in order to obtain good results and it is indispensable to adopt a linguistic approach that uses a legal thesaurus or ontology for juridical language.

Keywords : Arabic language, Indexing, N-grams, Information Retrieval, Word segmentation

1 Introduction

In this paper, we will discuss the implementation of techniques that allowed us to use n-gram based retrieval methods on an Arabic corpus.

In fact, to allow documents to be accessed by their content it is necessary to first index them. This process will vary depending on the style of access desired, but it should essentially summarize each document into the basic information that the matching routine will require.

It is often not possible for a retrieval engine to take a single query and find every document the user would consider relevant to that query. There are two main reasons for this limitation: that a query is rarely complete enough to specify all relevant documents, and that the retrieval engine is not powerful enough to match all the relevant documents.

In these cases we must calculate the effectiveness of the retrieval method and draw a conclusion about the results obtained.

Our retrieval methods have previously been used on Arabic documents only. The Arabic language presents a number of challenges to information retrieval (IR), which make an exact keyword match inadequate.

Two techniques spelling normalization and stemming are well known for IR. But previous experiments [1,2] show that while these techniques can significantly improve retrieval, they are still inadequate.

The third technique, retrieval based on character N-grams, has been tested in a few studies [5]. In this paper we will focus on this third technique.

2 Information Retrieval Issues

Information retrieval (IR) is a process that informs a user of the existence and whereabouts of information related to a specific request. The retrieval process is influenced by the indexing process as well as by the natural language that is being indexed [3].

N-gram models have been extensively used in text retrieval. A text retrieval process is typically divided into three stages.

The first stage is document indexing, in which content-bearing terms are extracted from the document text.

The second stage is index weighting, which is used to increase the relevance of the documents retrieved for a query.

The final stage ranks documents, relative to the query, according to a similarity measure.

At the indexing stage, documents are normally represented as a bag of words. However, in Arabic, a bag of words is not adequate for segmentation. One method for avoiding word segmentation is to use n-gram character features. In our work, we present a text segmentation and indexing software based on n-grams as a vector space model [17], in this model, documents and queries are represented as vectors in an N-dimensional hyperspace where N is the number of terms in the document or the query in which the documents and query are represented. Three measures of similarity were used: The first measure is the TF*IDF weighting [18], this statistical measurement allows evaluation of the importance of a word in a document extracted from a collection or a corpus. The second measure is Dice's coefficient, this measure has been used on many lexical similarities between word [15] [23]. Now that the documents are represented as vectors, the vector space model considers their similarities to be based on the angle between the two vectors in space. The third measure is the Cosine Coefficient [18], it a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them, often used to compare documents in text mining.

3 Retrieval Effectiveness (Precision and Recall)

The matching results of IR are imprecise and inexact, as in Database, so we need to measure IR effectiveness [7].

The first step taken in developing a methodology to test Arabic's retrieval effectiveness was to map the peculiarities of the Arabic language that might affect retrieval effectiveness. These can be seen simply as three features of Arabic. The Arabic language uses:

- Prefixes

For the definite article, some particles, and some plural forms

- Infixes

For some plural forms

- Suffixes

For some pronouns

The second step was to investigate how the effect of these features could be measured. We are used the twin measures of recall and precision [21], which were used in the information retrieval systems.

4 Challenges of the Arabic Language

The purpose of IR systems is to find relevant documents and provide the user with clear control mechanisms and a rapid response. A number of techniques and algorithms have been implemented within search engines.

Research and development (R&D) of Arabic text still has long way to go. Although academia has made significant achievements, the complex morphological structure of the Arabic language provides challenges; techniques must be developed to make IR efficient for Arabic [8].

Existing Arabic text retrieval systems could be classified into two groups [7]:

- Full form based IR: The mean surface string search (exact entered word) with the option to

extend their results to the inflection of the entered word [24].

- Morphology-based IR: The engines are content based linguistic search that have a strong base of linguistic processors. The efforts that have been made in the academic environment to evaluate more sophisticated systems provide a preview of the next generation of Arabic search engines. Evaluation of systems has been performed using different approaches of incorporating morphology: stem, root based and light stem [8]. The experiments show that the using stemmers improve the recall as well as the precision [9] and the light stemmer performs better than the regular stemmer.

While each of these methods has been proposed as an alternative solution for Arabic text retrieval, none of them are claimed to provide the optimum solution. For example, the word and stem methods are good at providing a more focused output, but may miss relevant texts [22].
The root method, on the other hand, is very efficient at retrieving all relevant text, but may also retrieve a great deal of irrelevant text. This is the quest for a more effective method of Arabic information retrieval

5 Features of the Arabic Language

The Arabic language is an inflectional language and not an analytic language [8]. The derivation in Arabic is based on morphological patterns, and the verb plays a greater inflectional role than in other languages. Furthermore, Arabic words are built up from roots representing lexical and semantic connecting elements.
Arabic offers the possibility of attaching particles and affixed pronouns to words. In other words, Arabic allows a great deal of freedom in the ordering of words within a sentence.
Thus, the syntax of the sentence can vary according to transformational mechanisms such as extraposition, fronting and omission, or according to syntactic replacement, such as the use of an agent noun in place of a verb.

The Arabic language is distinguished by its high context sensitivity in several areas. At the written level, the shape of a letter depends on the letter that precedes it and the one that follows it. At the syntactic level, the different synthetic coherence relations, such as case-ending, matching, connecting, associating and pronominalizing, represent various examples of syntactic sensitivity.
The context sensitivity feature is not only limited to letters, words, and sentences. Arabic sentences are embedded and normally connected by copulatives, exceptive particles and adversative particles. For this reason it is more difficult to identify the end of an Arabic sentence than it is to identify the end of a sentence in other languages.

6 Spelling Normalization and Mapping

Arabic orthography is highly variable. For instance, the changing of the letter YEH (ي) to ALEF MAKSURA (ى) at the end of a word is very common (not surprisingly, as the shape of the two letters is very similar). Since variations of this kind usually result in an “invalid” word, in our experiments we detected such “errors” by using a stemmer (the Buckwalter Stemmer) and restored the correct word ending.
A more problematic type of spelling variation is that certain glyphs combining HAMZA or MADDA with ALEF (e.g. إ ,أ and آ) are sometimes written as a plain ALEF (ا), possibly because of their similarity in appearance. Often, the word intended and what is actually written are both valid words.
This is much like confusing “résumé in French” with “resume” in English. Since both the intended word and the written form are correct words, it is impossible to determine the intended spelling without the use of context.
We explored the use of two techniques to address the problem.

1) With the normalization technique, we replaced all occurrences of the diacritical ALEFs by the plain ALEF.
2) With the mapping technique, we mapped a word with the plain ALEF to a set of words that could potentially be written as that word by the changing of the diacritical ALEFs to the plain ALEF. In this absence of training data, we could assume that all the words in the set were equally probable.
Both techniques have pros and cons. The normalization technique is simple, but it increases ambiguity. The mapping technique, on the other hand, does not introduce additional ambiguity, but is more complex.

7 Advantages of using N-grams of characters

We can define an N-gram of characters as a sequence of N characters. Suleiman [13] study a new idea by comparing the performance of two N-gram methods, the contiguous N-gram and the hybrid N-grams combining contiguous and non-contiguous characters. This method is used in many search engines that allow the search for all the entered words within the same sentence without regardless their order.
However, character level N-gram models have been used successfully in many information retrieval problems and offer the following benefits:

1- Language independence and simplicity: Character level N-gram models are applicable to any language, and even to non-language sequences such as music or gene sequences.
2- Robustness: Character level N-gram models are relatively insensitive to spelling variations and errors, particularly in comparison to word features.
3- Completeness: The vocabulary of character tokens is much smaller than any word vocabulary and is normally known in advance. Therefore, the problem of sparse of data is much less serious in character N-gram models of the same order.

8 Character N-grams in Arabic documents

Broken plurals are very common in Arabic. There is no existing rule-based algorithm to reduce them to their singular forms, and it seems that it would be not be straight-forward to create such an algorithm [16]. As such, broken plurals are not handled by current Arabic stemmers.
One technique to address this problem is to use character N-grams. Although broken plurals are not derived by attaching word affixes, many of the letters in broken plurals are the same as in the singular forms (though sometimes in a different order). If words are divided into character N-grams, some of the N-grams from the singular and plural forms will probably match.
This technique can also handle words that have a stem but cannot be stemmed by a stemmer for various reasons. For example, the Buckwalter stemmer employs a list of valid stems to ensure the validity of the resulting stems. Although the list is quite large, it is still not complete. N-grams in this case provide a fallback where exact word match fails.
In previous work [11], experiments were conducted with N-grams created from stems as well as N-grams created from words. N-grams were created by applying a shifting window of n characters over a word or stem. If the word or stem had fewer than n characters, the whole word or stem was returned. The experiments use two methods of creating N-grams: from words and from stems. The retrieval scores show that stem-based N-grams are more effective than word-based N-grams for retrieval. The probable reason is that some of the word-based N-grams are prefixes or suffixes, which can cause false matches between documents and queries.

9 Experimentation

In our experiment, the corpus will be the 2667 Arabic documents from the Lebanese government's official journals, this documents were extracted from the Lebanese juridical

Database [19]. Using these documents, we will categorize them by trying different parameters for N-gram methods (based on words, stems and characters). We will then try to find the candidate words for each document and compare the results obtained by different method parameters. In our approach we have chosen the N-gram method to represent information in a vector space model. So each document and the query as well, are represented as a vector of terms (descriptors). The term can be a word, 3-gram, 4-gram or 5-gram.

The first level of linguistic processing which typically occurs is the removal of “stop words”--words which have no meaning when taken out of context and the tokens can be normalized by removing their diacritics. Once we have cut our documents (and also our query, if it is a free text query) into tokens, it is simplest if tokens in the query exactly match tokens in the token list of the document. However, there are many cases when things are not quite the same, but you would like a match to occur.

In these cases, the user simply types in a list of keywords of interest, and the search engine retrieves matches, ranking them in descending order of relevance. We will now describe how this works.

- We derive the vector for each document as follows. Each term in the document is given a weight. This weight is typically based on what is termed "IDF * TF”.

- IDF reflects the fact that uncommon words, when specified in a query, are more likely to be useful in narrowing down the selection of documents than very common words.

- TF reflects the fact that if a keyword occurs multiple times in a document, that document is more likely to be relevant than a document where the keyword occurs just once.

- -The query itself can be regarded as an M-dimensional vector, where M is the number of terms in the query. Each dimension is given a weight of 1.

- To compare the relevance of a particular document to the query, we compare the query vector with the document's vector using Dice's coefficient and the cosine coefficient, which is the angle between the two vectors in space. Relevance ranking means sorting the matching documents in descending order of cosine.

10 Experimental Results

To test our approach, we have developed software that segments our corpus using an N-gram with N = 3, 4 and 5 characters, and presented the top 50 results in a table.

The results were obtained by choosing the keywords with the juridical expert.

We have chosen, after conducting some tests on different values of thresholds and showing the results of thresholds in previous experiments [12]:

- 0.6 as a threshold value when using Dice's coefficient.

- 0.7 as a threshold value when using Cosine coefficient.

- 0.01 as a threshold value when using TF*IDF weights.

a) N-gram (3 characters)

Dice's coefficient (threshold =0.6)			Cosine coefficient (threshold =0.7)			TF*IDF weight (threshold=0.01)		
Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
53.3	40.81	46.22	41.34	10.8	21.68	37.34	26.9	31.27
92.85	1.45	2.85	89.28	2.34	4.56	46.42	37.7	41.61
73.075	21.13	24.53	65.31	6.57	13.12	41.88	32.3	36.44

Table 1: result using 3-gram

b) N-gram (4 characters)

Dice's coefficient (threshold =0.6)			Cosine coefficient (threshold =0.7)			TF*IDF weight(threshold=0.01)		
Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
1.33	50	2.59	13.34	58.82	21.74	38.67	36.25	37.42
75	1.17	2.3	67.85	1.06	2.08	60.71	6.67	12.01
38.165	25.585	2.445	63.33	7.2	11.91	49.69	21.46	24.71

Table 2: : result using 4-gram

c) N-gram (5 characters)

Dice's coefficient (threshold =0.6)			Cosine coefficient (threshold =0.7)			TF*IDF weight(threshold=0.01)		
Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
1.33	50	2.59	8	42.85	13.48	6.67	45.45	11.65
71.42	7.905	14.23	71.42	8.26	14.81	67.85	9.74	17.03
36.375	28.95	8.41	39.71	25.55	14.14	37.26	27.59	14.34

Table 3 : : result using 5-gram

d) Using the Boolean approach

3 characters			4 characters			5 characters		
Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
89.34	2.71	5.26	84	3.76	7.19	4	21.43	6.74
89.28	0.95	1.88	85.71	1.19	2.34	85.71	1.34	2.63
89.31	1.83	3.57	84.85	2.47	4.76	44.85	11.38	4.68

Table 4: : result using Boolean approach

e) Using word

word		
Recall	Precision	F1
1.34	50	2.61
60.71	7.59	13.49
31.03	28.79	8.05

Table 1 : result using word

11 Discussions

The results of our experiment have demonstrated that using the N-gram method in Arabic information retrieval is more effective than using keyword matching.

But we have obtained a maximum value of F1 = 36.44, which is still an insufficient level of effectiveness for information retrieval.

Thus, though using the N-gram method is more effective than keyword matching, it still insufficient. Therefore, we have to consider adding a linguistic level.

The reason for this insufficiency is the specificity of Arabic language, in particular, the great number of synonymies, directives…

For example, the word "the war" has different synonyms, so all relevant documents that contain these relevant words would not be retrieved because there are no common grams.

Concerning the optimal value of N that gives us the greatest results, we can see that for N=3 characters we have obtained the maximum value of F1.

In fact we have obtained the maximum value of F1 (=36.44) in the case of N=3 grams and using TF*IDF weighting, and also when using Dice's coefficient with a threshold of 0.6 (F1=24.53).

We can explain this by the fact that most Arabic root words are formed from 3 characters, the work in [19] indicates that the use root indexing method will give better result than using word indexing method. The result is confirmed by other studies, for the Arabic text classification in [25], the tri-gram text classification using the Dice measure outperforms classification using the Manhattan measure and for Arabic text categorisation [26], the author's indicate that Dice based TF.IDF and Jaccard based TF.IDF outperforms Cosine based TF.IDF.

We can conclude three things from these results:

1- Using N-grams when searching Arabic documents is more effective than using keyword matching.

2- Using trigrams to index Arabic documents is the optimal choice for Arabic information retrieval using N-grams.

3- Using N-grams by indexing Arabic documents is still insufficient to obtain good results in Arabic information retrieval.

12 Conclusions and Future Work

Arabic is one of the most widely used languages in the world, yet there are relatively few studies on the retrieval of Arabic documents.

N-grams have been widely investigated for a number of text processing and retrieval applications.

The primary goal of this paper is to investigate the performance of N-grams within the context of Arabic textual retrieval systems.

The main contribution of this paper is its demonstration of the effectiveness of the N-gram method relative to the keyword matching method, and the good choice of the number of characters of this method (Value of N).

This work evaluated the use of N-grams for the retrieval of Arabic documents, using Lebanese official journal documents (Arabic corpus) as the test bed.

We can deduce from the "keyword matching" retrieval method and the statistical approach of text mining that if we need to improve the results, we have to think about a hybrid approach somewhere in between the statistical and linguistic approaches--and that might be a concept matching approach using N-grams.

In future work, we can essay, for example, to use stemming or concept matching before applying the N-gram indexing method. In fact, research in the area of Arabic information retrieval has shown that stemming in an improvement of the performance of Arabic information retrieval systems [14, 15, 16].

References:

[1] Riyas Al-Shalabi and Marth Evens, "A computational morphology system for arabic", Workshop on Computational Approaches to Semitic Languages COLING-ACL98, 1998.

[2] Xu, A. Fraser, and R. Weischedel, "Empirical studies in strategies for arabic information retrieval", SIGIR 2002, Tampere, Finland.

[3] Norbert Fuhr,2001. in Lectures on information retrieval. Springer-Verlag New York, Inc.,2001, Pages: 21 - 50

[4] Victor Lavrenko. Center for intelligent Information Retrieval University of Massachusetts Amherst. Hopkings IR workshop, 2005

[5] Darwish et al, 2001; Mayfield et al, 2001; Kwok et al, 2001.

[6] Haidar Moukdad and Andrew Large ,"Information Retrieval from full-text Arabic Databases: Can Search Engines Designed for English Do the Job? ", Libri,,vol.51, pages.63-74 – 2001.

[7] Ahmed Abdelali ,"Improving Arabic Information Retrieval Using Local variations in Modern Standard Arabic," New Mexico Institute of Mining and Technology, 2004.

[8] Ahmed Abdelali, Jim Cowie and Hamdy S.Soliman,"Arabic Information Retrieval Perspectives", JEP-TALN 2004,Arabic Language Processing, Fez, 19-22 April 2004.

[9] Larkey L, Lisa Ballesteros, and Margaret Connell, "Improving stemming for arabic information retrieval:Light stemming and co-occurrence analysis", In SIGIR 2002, pages 269 - 274, 2002.

[10] P. McNamee, "Knowledge -light Asian Language Text Retrieval at the NTCIR-3 Workshop," Working Notes of the 3rd NTCIR Workshop, 2002

[11] James Mayfield, Paul McNamee, Cash Costello, Christine Piatko, and Amit Banerjee," Experiments in Filtering and in Arabic, Video, and Web Retrieval", In E. Voorhees and D. Harman (eds.), Proceedings of the Tenth TextREtrieval Conference (TREC 2001), Gaithersburg, Maryland, July 2002.

[12] Suleiman H. Mustafa and Qasem A. Al-Radaideh, "Using N-Grams for Arabic Text Searching", Journal of the American Society for information science and technology, 55(11):1002-1007,2004.

[13] Suleiman H. Mustafa, "Character contiguity in N-gram-based word matching: the case for Arabic text", Searching, in Information Processing and Management (41), 819-827, 2005

[14] Larkey, L.S. and Connell, M.E, "Arabic information retrieval at UMass in TREC-10.", TREC 2001. Gaithersburg: NIST, 2001.

[15] Dice L R. "Measures of the Amount of Ecologic Association between Species" in Ecology, Vol. 26, No. 3, pp. 297-302, 1945

[16] Goweder A, Poesio M, DeRoeck A, Reynolds J, "Identifying Broken Plurals In Unvowelised Arabic Text", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, Sheffield, United Kingdom .

[17] Salton G , A. Wong , C. S. Yang," A vector space model for automatic indexing", Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975

[18] Salton G., Buckley C, "Term-weighting approaches in automatic text retrieval", Information Processing and Management, 24(5), 513-523, 1988.

[19] Ibrahim M.M. El Emary and Jaafa Atwan, "Designing and Building an Automatic Information Retrieval System for Handling the Arabic Data", American Journal of Applied Sciences 2 (11) pp 1520-1525, 2005.

[20] RAMMAL M, Mona Al ACHKAR, Phlippe NABHAN, "Computer Assisted research system on legal Lebanese document", Workshop on Arabic Software, Lebanese American University, 2001

[21] Clarke, S., Willett P, "Estimating the recall performance of search engines" ASLIB Proceedings, 49 (7), 184-189.

[22] Abu-Salem H, Al-Omari M, Martha W. Even, "Stemming methodologies over individual query words for an Arabic Information Retrieval System", Journal of the American Society for Information Science, Volume 50 Issue 6, Pages 524 –529 ,1999.

[23] Adamson G W, Boreham J, " The use of an association measure based on character structure to identify semantically related pairs of words and document titles" in Information Storage and Retrieval, Vol. 10, No 7-8, pp 253-260, 1974.

[24] The search engines are full form retrieval systems used such as Google,Yahoo, Ayna and alltheweb, but we note that the Sakhr has investigate on the Arabic processing technology and produce IDRISI the most sophisticate Arabic search engine (www.sakhr.com).

[25] Khreisat L, "A machine learning approach for Arabic text classification using N-gram frequency statistics" in Journal of Informetrics No 3, pp 72–77, 2009

[26] Thabtah F, Wa'el Musa Hadi W. M, Al-shammare G, "VSMs with K-Nearest Neighbour to Categorise Arabic Text Data" , in Proceedings of the WCECS 2008, 2008, San Francisco, USA 2008