EasyChair Preprint

№ 6786

A Language Independent Approach to Multilingual Document Representation Including Arabic

Souhila Boucham and Hassina Aliane

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

October 6, 2021

A language independent approach to multilingual document representation including Arabic

Souhila Boucham
Faculty of Electronics and Computer Science
USTHB University
Algeria
sbouchem@yahoo.fr

Hassina Aliane
Research Center on Scientific and Technical Information
University of Huaguoshan
CERIST,Algeriaa
haliane@mail.cerist.dz

***Abstract* - Arabic language is of increasing interest in the field of Multilingual Information Retrieval (MIR). We deal in this work with the problem of multilingual document representation including Arabic. The proposed approach combines a surface analysis and a Latent Semantic Analysis (LSA) algorithm in a new way to break down the terms of LSA into units which correspond more closely to morphemes. These morphemes are the variable length character N-gram candidates extracted from different fragments separated by borders. The length of the character N-gram candidates is variable because each language has its own properties. This strategy brings an interesting performance for languages such as Arabic in which the words are not explicitly defined and different words are not separated by spaces. The obtained results are encouraging and variability shows that they are perfectible.**

Keywords - multilingual document representation, multilingual information retrieval (MIR), virtual document, principle of border, variable length character N-grams, concept types, pivot language.

I. INTRODUCTION

Several research projects are investigating and exploring various techniques in Information Retrieval (IR) systems for the English, European and Asian languages. However, in Arabic language, there is little ongoing research in Arabic IR or MIR systems including Arabic.

Arabic language is one of the most widely spoken languages. It has a complex morphological structure and is considered as one of the most prolific languages in terms of linguistic articles. Therefore, Arabic IR models need specific techniques to deal with its complex morphological structure [1].

The Core of a MIR system is the indexing process and the retrieval model. In this study, we will focus on models which use an indexing process to store data and to determine how multilingual documents (Arabic, French and English) are represented.

When processing a large corpus with a statistical tool, the first step typically consists of subdividing the text into information units called tokens. These tokens usually correspond to words. This tokenization process may appear to be quite simple, if not to say trivial-tokenization.

However, from an automated processing point of view, the implementation of this process constitutes a challenge. Indeed, how to reliably recognize words? What are the unambiguous formal surface markers that can delineate words, i.e. their boundaries? These questions are relatively easy to answer for languages such as French or English: basically, any string of characters delimited by a beginning space and an ending space is a simple word. But for many other languages, such as Arabic, the answer is much more complicated. In Arabic, subject pronouns and complements are sometimes attached to the verb. In this case, a token like katabtuhu "كتبته"corresponds in fact to a sentence (here, "I wrote it" or "I've written it") ("je l'ai écrit" in French). Obviously, the simple notion of tokens defined as character strings separated by spaces is an oversimplification that is highly inadequate for many situations and languages.

Considering the above, what then could constitute a reasonable atomic unit of information for the segmentation of a text, independently of the specific language it is written in?

The proposed approach avoids the use of tokenizers, stemmers or other language-dependent tools which are complex and may bring noise to representation especially for the high morphologically complex languages including Arabic. The approach is characterized by:

- Language -independent.
- Easy to apply and does not require Natural Language Processing (NLP) tools.
- Construction of the feature terms is based on the N-grams of characters.

The rest of the paper is organized as follows: Section II introduces existing related work. Section III presents the proposed approach to multilingual document representation. Experimental results and analysis are reported in Section IV. Discussion in Section V, and the last section concludes this paper.

II. RELATED WORK

Several approaches have been proposed to solve the document representation problem:

• In [2], a character string delimited with trivial separators such as spaces or punctuation signs, constitutes a basic sense unit,

• [3] and [4] proposed an approach based on the concept of word: inflected or lemmatized form, term, multi-terms.

• An ontology based approach has been used in [5], [6] and [7]. The ontology model preserves the domain knowledge of a term present in a document. However, automatic ontology construction is still a difficult task due to the lack of structured knowledge bases.

• To cope with the morphological processing issue, [8] proposed an approach based on character-level N-grams with linear classifiers in the framework of content-based anti-spam filtering. Their results showed that character N-grams are more reliable features than word-tokens despite the fact that they increase the problem of dimensionality. In the following, we present N-grams based approaches to document representation:

In [9], the authors proposed an N-gram based approach to IR for NLP applications, the character N-grams have a constant number of characters, defined a priori. This approach is applied specifically for Cross Language Information Retrieval (CLIR) Systems. However, the results of CLIR are exclusively for European languages written in the Latin alphabet.

In [10], the authors describe an entirely statistical-based, unsupervised approach to MIR, called Latent Morpho-Semantic Analysis (LMSA). This approach has an important theoretical advantage over LSA: it combines well-known techniques in a novel way to break down the terms of LSA into units which correspond more closely to morphemes. The authors have demonstrated that LMSA is a morphologically more sophisticated alternative to LSA.

In [11], the authors propose an hybrid representation of documents for text classification in optical character recognition of documents. The first step consists in Part-of-Speech tag, then the application of the principle of border. The next step is the representation with character N-grams. Hence the document representation is a merger of N-grams of different fragments separated by the border.

In [12], the primary goal of the authors is to investigate the performance of N-grams within the context of Arabic textual retrieval systems. The main contribution of this work is its demonstration of the effectiveness of the N-gram method compared with the keyword matching method. They also showed the importance of a good choice of the N.

Nevertheless, though using the N-gram method is more effective than keyword matching, it is still insufficient due to Arabic language specificities such as the great number of synonymies, directives, ... Therefore, it is preferable to consider adding a linguistic level. Two approaches have been proposed to solve the MIR problem: document based representation approaches and word based representation approaches.

In [13], the authors proposed a novel bigram alphabet approach for features construction and its application in text classification. Term frequency of bi-gram alphabet was used as a weighting scheme to represent document contents. The approach is language independent and does not require NLP tools. Using Vector Space Model - Sequential Minimal Optimization (VSM-SMO) classifier, the proposed approach has proved the ability to classify collections of Arabic and English text documents successfully.

In conclusion, we are aware of only few methods using language independent indexing methods. Moreover works in the field of MIR including Arabic language are very few and are not yet mature and the problem of MIR including Arabic language is still open. By considering several issues discussed in the previous and this section, we propose an approach in order to improve document representation in MIR.

III. A PIVOT LANGUAGE BASED APPROACH TO MULTILINGUAL DOCUMENT REPRESENTATION

We propose to cross the language barrier and complex morphological structure of the Arabic language using a concept based pivot language. This language is used to represent the document and the query. The problem then is the definition of a pivot language for MIR and, conversion and reconversion from natural language to this representation language.

A. Concept types

In concept classification, the objects considered as similar are grouped in a same group. Among the unsupervised approaches, LSA is an automatic statistical method for IR that re-describes the textual data in a new smaller semantic space. The fundamental aim of an LSA model is to achieve a conceptual representation of documents. The LSA technique may be seen as the introduction of an indexing language which is constituted of concept types by changing the expression space of index vectors on new dimensions concretizing the language: the document and the query are represented in a common space independent from a language. This approach is based on the vector model. The whole problem lies in defining the vector space.

B. Vector space

The vector formalization of a document, which reduces it to an unordered list of index terms, sufficient to reveal resemblances, semantic proximities between documents (documents/ queries) in a corpus. The problem posed is to find, regardless of the language or script, the descriptors or basic units of information that are identifiable and extractable well, as most relevant to a document collection written in a given language. [14] suggest that this unit should be defined according to the goal we set ourselves when reading or processing a text. More precisely, from a numerical classification based knowledge extraction viewpoint, the definition of the basic unit of information to be considered depends on the following:

• The unit of information must be a portion of the input text submitted to the numerical analysis processor;

• From an automated processing point of view, it should be easy to recognize these units of information;

• The definition of the unit of information should be independent of the specific language the text is written in;

• The units of information must be statistically meaningful when evaluated or compared between themselves. It should be easy to compute their frequencies in various parts of the input text, as well as to estimate their distribution and the regularity with which some units co-occur in certain portions (segments) of the text.

Although, it makes tokenization (where words are considered as tokens) relatively easy in English or French, it is much more difficult for other languages such as Arabic. Moreover, stemming or lemmatisation, typically used to normalize and reduce the size of the lexicon, constitutes another challenge. The notion of N-grams which, for the last ten years, seems to have produced good results both in language identification and speech analysis, has recently become a privileged research axis in several areas of knowledge acquisition and information extraction.

The concept of N-grams was first discussed in 1951 by Shannon [15]. Since then, the concept of N-grams has been used in many areas, such as spelling-related applications, string searching, prediction and speech recognition.

A character N-gram is a character sequence of length N extracted from a document. To generate the N-gram vector for a document, a window of N characters in length is moved through the text, sliding forward one character at a time. At each position of the window, the sequence of characters in it is recorded. For example, the first 5-grams in " character sequences..." are: " char", "chara", "harac" and "aract". In some schemes, the window may be slided more than one character after each N-gram is recorded.

In the following we explain the advantages of character N-grams encoding:

First, the system can be garble tolerant by using character N-grams as basic terms. If a document is scanned using Optical Character Recognition (OCR), there may be some misread characters. For example, suppose "character" is scanned as "claracter". The word-based system will not be able to match this word because it is misspelled, but a character N-gram based system will still match the other character N-grams such as "aract", "racte"… and take their frequency into account.

Second, by using character N-grams, the system can achieve language independence. In a word-based IR system, there is language dependency. For example, in some Asian languages, different words are not separated by spaces, so a sentence is composed of many consecutive characters. Grammar knowledge is needed to separate those characters into words, which is a very difficult task to perform. Using character N-grams, the system does not need to separate characters into words.

Additionally, character N-gram based systems do not use stop words. This is because the number of unique character N-grams in a document is very big and distribution is very wide. There is few character N-grams that have high frequency. From Ekmekcioglu's research [16], stop words and stemming are superior for word-based system but not significant for a character N-gram based system.

C. Creation of the semantic space as a pivot language

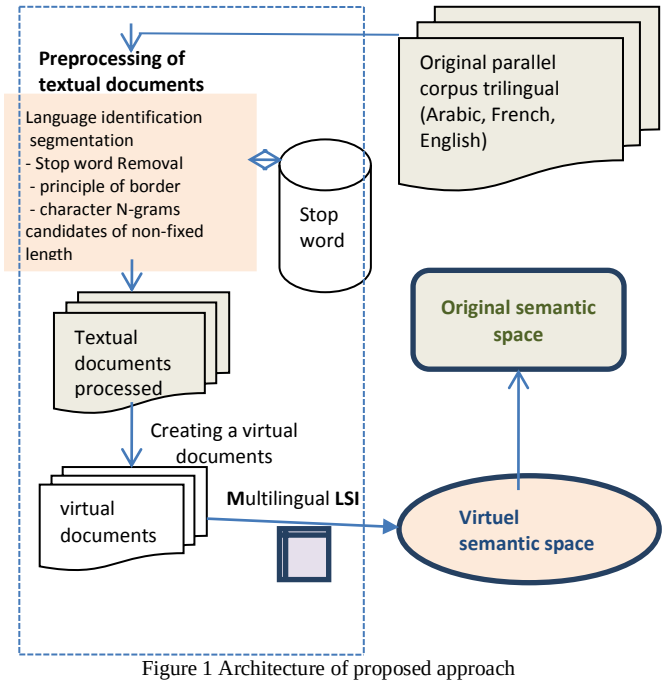

Figure 1 Architecture of proposed approach

To create the semantic space, we proceed in two phases:

• Corpus building: the constitution of our multilingual corpus (Arabic, French and English).

• The second phase comprises the following steps:

1. Preprocessing step and process of feature extraction: These steps are detailed in the following:

Segmentation: it consists in determining the relative positions of each text stream and the paragraph breaks including the end of the sentences which should be kept. Nevertheless, some difficulties may occur in this step namely the removal of any sequences of successive dots or ellipsis and special characters which sometimes indicate the end of the sentence and then need a particular processing.

At the end of this step, we generate a list of segments. Each segment is marked by a starting offset (the beginning of the segment) and an ending one (the end of the segment). These offsets allow referencing any extracted segment.

Stop word Removal: we removed stop words for French, English and Arabic, using predefined stop word lists. For French, this list contains mainly articles, pronouns, etc.

Arabic is a morphologically rich language with a large set of morphological features such as person, number, gender, voice, aspect, case, and state. Arabic features are realized using both concatenative (affixes and stems) and templatic (root and patterns) morphology with a variety of morphological, phonological and spelling adjustments. In addition, Arabic has a set of very common clitics that are written attached to the word, e.g., the conjunction و+ w+ 'and', the preposition + ب b+2 'with/in', the definite article + ال Al+ 'the' and a range of pronominal clitics that can attach to nouns (as possessives) or verbs and prepositions (as objects).

Therefore, Arabic is an agglutinative language. Articles, prepositions and pronouns stick to adjectives, nouns, verbs and particles to which they relate, which generates ambiguity in morphological analysis of words. All these empty words of Arabic can be concatenated together. For example, for "تلك" we can derive 'فبتلك' = ف+ب+تلك et.وتلك = و+تلك

N-grams of characters: We conducted a search of N-grams regardless of their size. This choice is justified by the fact that it frees from the notion of word, so from any morpho-syntactic analysis. To introduce our approach, we consider the sentence: " le bijoux plaqué or a du charme (the goldplated jewellery has charm)".

The selection and removal of stop words returns the following result: "bijoux plaqué or a charme".

The application of N-grams of characters process gives two possibilities of representation:

• The first representation is based on a bag of selected words. The application of N-grams with N = 5 gives the following result: « _bijo, bijou, ijoux, joux_, oux_p, ux_pl, x_pla, _plaq, plaqu, laqué, aqué_, qué_o, ué_or, é_or_, _or_a, or_a_, **r_a_c, _a_ch, a_cha**, _char, charme, harme, arme ». This application is flawed because it adds noise and unnecessary N-grams. For example: "a_cha'" is a Ngrams which represents noise (N-grams from the fragment "a du charme (has charm)" where the word "du" was deleted). Indeed, the elimination of the stop words of the initial sentence returns irrelevant results.

• A second representation is based on the N-grams of characters for each extracted word separately. As result, we have: " _bijo, bijou, ijoux, joux_, _plaq, plaqu, laqué, aqué_, _char, charme, harme, arme_". This representation corrects the defects caused by the previous method but provides fewer data (in particular with short words). For example, by using the character N-grams with $N \geq 5$, the noun "or" cannot be identified. This deletion causes a loss of information.

Principle of border: The two representations mentioned above have major defects with the introduction of **noise** (first method) and **silence** (second method). Thus, we have introduced a principle of border. In our study, the words giving less information (i.e. stop words list) are replaced by a border. This method corrects the noise added during the first proposed treatment. it takes into account groups of words (e.g. "plaqué or"). The result according to the principle of border is shown below: " X bijoux plaqué or a X charme", "X" represents the border.

Then we can extract the character 5-grams in the two fragments of the text (i.e. " bijoux plaqué or a " and " charme "): " _bijo, bijou, ijoux, joux_, oux_p, ux_pl, x_pla, _plaq, plaqu, laqué, aqué_, qué_o, ué_or, é_or_, _or_a, or_a_, _char, charm, harme, arme_ ".

The proposed algorithm is organized as follows:

Inputs: The set of trilingual documents forming the corpus.

Outputs: Matrix.

For all documents do:

- Segmentation and removal of stop words.
- Application of the principle of border.
- Representation of words extracted with the character N-grams of variable length.
- Assigning weights based on statistical measure (entropy).

End.

After applying these different stages we obtain a representation for each document.
The length N of N-grams is not fixed since every language has its own properties.
As in [10], we propose a character selection method of non-fixed length based on repeated N-grams that are already filtered by the principle of border.
After extractin all candidate tokens for each final N-gram, we filter these candidates to select a single candidate which best represents the final N-gram and maximizes Mutual Information (MI).

$$MI_{max}(\text{N-gram}) = max_{i=1}^{n}\{MI(S_i)\}$$

S_i (i = 1, …, n) denotes the set of all N-grams of length i.

2. Take the documents of the three languages, concatenated to create a set of virtual documents. The virtual document is the concatenation of a source document + its translations in two target languages).

3. Analysis phase: the virtual document is considered as one document regardless of the language. The set of documents is analyzed by LSA algorithm. The terms of the term-document matrix are the morphemes results of the preprocessing step from the corpus. LSA examines the similarity of the contexts in which terms appear, and creates a reduced-dimension feature-space representation where terms that occur in similar contexts are near each other.

4. The result is a reduced semantic space that will serve as a pivot language where related terms are grouped in the same concept: concepts constitute the pivot language.

5. Represent the documents in each language around the space terms.

IV. EXPERIMENTAL EVALUATION

To test our approach, we have developed a software that index our corpus using an character Ngram with N = 3, 4 and 5 characters, and presented the top results.

A. *Building and results of the Quranic Corpus*

The texts are extracted from websites http://www.lexilogos.com/coran.htm; http://www.alargam.com/quran2/quran3/index.htm. Since the Quran comprises 114 Surats, the corpus consists of 114 documents in each language (ie 342 documents) and 100 types of queries (verses), whose relevance is evaluated manually through web search. We will work on the corpus in a raw format.

Example

" الله لا إله إلا هو الحي القيوم لا تأخذه سنة و لا نوم له ما في السموات و ما في الأرض من ذا الذي يشفع عنده إلا بإذنه يعلم ما بين أيديهم و ما خلفهم و لا يحيطون بشيء من علمه إلابما شاء وسع كرسيه السموات و الأرض و لا يؤوده حفظهما و هو العلي العظيم" البقرة 255

Figure 2 Query verse of al-Kursi number 255, Surah the cow (Al-Baqarah)

The top 20 closest documents to this query are given in the following figure:

document title
1. سورة البقرة رقم 2 الله لا إله إلا هو الحي القيوم لا تأخذه سنة و لا نوم له ما في السموات و ما في الأرض من ذا الذي يشفع عنده إلا بإذنه يعلم ما بين أيديهم و ما خلفهم و لا يحيطون بشيء من علمه إلابما شاء وسع كرسيه السموات و الأرض و لا يؤوده حفظهما و هو العلي العظيم(255) "
2. سورة آل عمران رقم 3 الم (1) الله لا اله الا هو الحي القيوم (2)
3. سورة غافر رقم 40 حم * تَنزِيلُ الْكِتَابِ مِنَ اللَّهِ الْعَزِيزِ الْعَلِيمِ * غَافِرِ الذَّنْبِ وَقَابِلِ التَّوْبِ شَدِيدِ الْعِقَابِ ذِي الطَّوْلِ لا إِلَهَ إِلَّا هُوَ إِلَيْهِ الْمَصِيرُ [غافر:1-3]
4. سورة طه رقم 20 (4. و عنت الوجوه للحي القيوم و قد خاب من حمل ظلما (111
5. سورة الحاقة رقم 69 كَذَّبَتْ ثَمُودُ وَعَادٌ بِالْقَارِعَةِ)4(
6. سورة الشورى رقم 42 فَلِذَلِكَ فَادْعُ وَاسْتَقِمْ كَمَا أُمِرْتَ وَلا تَتَّبِعْ أَهْوَاءَهُمْ وَقُلْ آمَنْتُ بِمَا أَنْزَلَ اللَّهُ مِنْ كِتَابٍ وَأُمِرْتُ لأَعْدِلَ بَيْنَكُمُ اللَّهُ رَبُّنَا وَرَبُّكُمْ لَنَا أَعْمَالُنَا وَلَكُمْ أَعْمَالُكُمْ لا حُجَّةَ بَيْنَنَا وَبَيْنَكُمُ اللَّهُ يَجْمَعُ بَيْنَنَا وَإِلَيْهِ الْمَصِيرُ (15](لَهُ مَا فِي السَّمَاوَاتِ وَمَا فِي الأَرْضِ وَهُوَ الْعَلِيُّ الْعَظِيمُ (4
7.سورة الواقعة رقم 56
8. سورة النازعات رقم 79 ءانتم اشد خلقا ام السماء بنيها (27) رفع سمكها فسويها (28) و اغطش ليلها و اخرج (ضحيها (29) و الأرض بعد ذلك دحيها (30
9. سورة لقمان رقم 31 وْ أَنَّمَا فِي الْأَرْضِ مِن شَجَرَةٍ أَقْلَامٌ وَالْبَحْرُ يَمُدُّهُ مِن بَعْدِهِ سَبْعَةُ أَبْحُرٍ مَّا نَفِدَتْ كَلِمَاتُ اللَّهِ إِنَّ اللَّهَ عَزِيزٌ حَكِيمٌ (27 الم - تِلْكَ آيَاتُ الْكِتَابِ الْحَكِيمِ - هُدًى وَرَحْمَةً لِلْمُحْسِنِينَ - الَّذِينَ يُقِيمُونَ الصَّلَاةَ وَيُؤْتُونَ الزَّكَاةَ وَهُمْ بِالْآخِرَةِ هُمْ يُوقِنُونَ
10. سورة الملك رقم 67
11. سورة القدر رقم 97
12. سورة التغبن رقم 64
13 . سورة الاحقاف رقم 46
14. سورة المؤمنون رقم 23
15. سورة الروم رقم 30 في سورة الروم (فَسُبْحَانَ اللَّهِ حِينَ تُمْسُونَ وَحِينَ تُصْبِحُونَ (17) وَلَهُ الْحَمْدُ فِي السَّمَاوَاتِ وَالْأَرْضِ وَعَشِيًّا وَحِينَ تُظْهِرُونَ (18))
16. سورة هود رقم 11 وَهُوَ الَّذِي خَلَقَ السَّمَاوَاتِ وَالْأَرْضَ فِي سِتَّةِ أَيَّامٍ وَكَانَ عَرْشُهُ عَلَى الْمَاءِ لِيَبْلُوَكُمْ أَيُّكُمْ أَحْسَنُ عَمَلًا وَلَئِنْ قُلْتَ إِنَّكُمْ مَبْعُوثُونَ مِنْ بَعْدِ الْمَوْتِ لَيَقُولَنَّ الَّذِينَ كَفَرُوا إِنْ هَذَا إِلَّا سِحْرٌ مُبِينٌ (7
26. سورة مريم رقم 19

Figure 3 20 closest documents

A surat in the quran may be related to other surats. Our experimentation shows that the top results are those surats directly related to the query surat. The same principle is applied to *retrieve documents to a query in French and English* documents. The results obtained are encouraging and variability shows that they are perfectible. Most queries are processed verses and deal with a specific topic.

B. Collection of the Institute of Scientific Information (CISI)

We needed to work on another corpus CISI, because we need to compare the results in terms of relevant documents for a given query. CISI account 1460 documents and 112 queries. After indexing, the corpus account 38821 terms to 38821 for 1460 documents. The CISI corpus has numerous queries that have no or few really deemed relevant documents. For this reason, we limited our study to

all queries having at least 10 relevant documents i.e. 67 requests.

1) Evaluation criteria: We have chosen to compare our measure to the cosine measure and the SimRank measure [17]. The approach presented in [17] consists in comparing document and request on the basis of their relation system. This approach exploits the information conveyed by the relations between terms and documents, those between terms and those between documents.

In order to evaluate our model, we have used the Mean Average Precision (MAP) measure, which is a widely accepted measure in the evaluation of the performance of IR systems. The average precision provides a global view of the performance of an IR model through a set of queries. However, the average set of queries can hide many details. It is not so easy to determine what leads to increase or decrease the average precision. To obtain a better explanation and understanding of the difference between the different models, we carried out a query-by-query analysis. In order to verify the significance of the results obtained, we carried out a test on the query number 62 for which our method majors the methods SimRank and cosine.

To evaluate this query, we computed exact precision measures p@5, p@10, p@30, p@100 and p@200 representing respectively, the precision values at the top 5, 10, 30, 100 and 200 documents returned and R-precision. R-precision: R-precision is defined as precision at cut-off R, where R is the number of relevant documents for the query.

	COS	SimRank	**Our method**
p@5	0	0,08	**0,25**
p@10	0	0,16	**0,5**
p@30	0	0,25	**0,58**
p@100	0,33	0,41	**0,66**
p@200	0,41	0,58	**0,75**
R-Prec	0	0,25	**0,5**

TABLE 1
Improvement in average precision at top n documents returned and R-precision

The following graphic allows to visualize the position of relevant documents retrieved and thus the evolution of this position from one method to another.

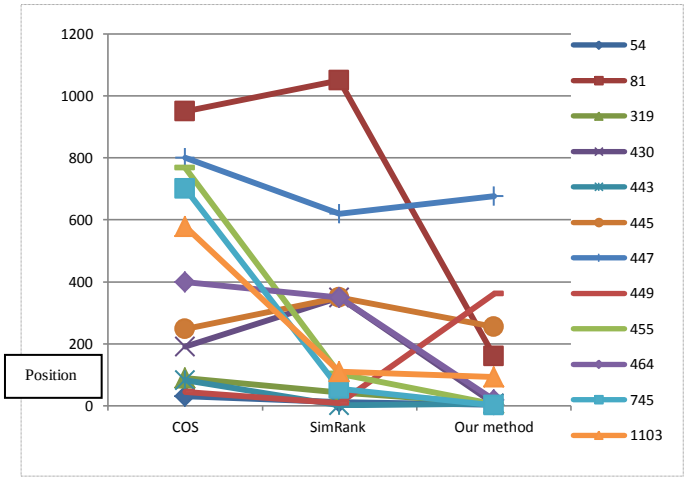

Figure 4 Comparison of three methods of the ranks of relevant documents to the query number 62.

On studying the top relevant documents' of the list: the documents number 1103 and 54, are retrieved in 579th and 31st position by the cosine method, 110th and 12th position by the SimRank method progressing a few ranks through our method (92nd and 3rd).

Similarly, the following two relevant documents number 319 and 443 are retrieved themselves in 90th and 82nd position and are retrieved improved by SimRank which classifies them into 43rd and the first position and such documents are retrieved topping the list by applying our method (8th and 6th position). This is due to the fact that these documents have a strong direct relationship with the query and an inter significant resemblance increasing their similarity to the query. The following two relevant documents, 447, and 445, respectively positioned in 800th, 247th by the cosine method, are around 620th and 350th position by the SimRank and respectively in the 675th and 255th by our method, this indicates an indirect resemblance to documents not resembling the query, hence the lightweight distance from the topping list. The most outstanding gain is obtained by the documents number 430, 81.464, 455 and 745 positioned in191st, 950th, 400th, 768th and 700th by the cosine who find themselves powered the 350th, 1050th, 350th, 103rd and 550th by SimRank and 7th, 160TH, 18th, 5th and 1st by our method.
We suppose this is due to low cosine of these documents with the query and an indirect strong relationship with it.

On 12 relevant documents, 1 document regresses, 2 documents practically keep the same position and 9 documents are progressing in rank relative to SimRank method. However, only one document regresses, 1 document practically keeps the same position and 10 documents are progressing in rank relative to the cosine method.

The documents with a high cosine progresses slightly with SimRank, indeed SimRank measure indicates both direct and indirect resemblance.

(Concerning documents with very strong progression (430, 81.464, 455et 745, 1103, 443, 319), we are pleasantly surprised to see that indirect resemblances can significantly improve documents rank.

• *Average Precision of n documents returned:* The evolution of the average Precision of n returned documents (0 <n <200) for Simrank is 0.15 around 0.025 for the Cosine and 0.4108 for our method. Our method interestingly maximizes both the SimRank and cosine measures.

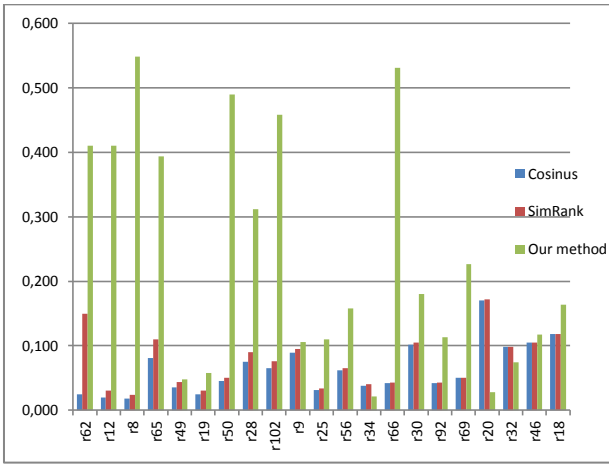

Figure 5 Comparison Cosine / Simrank / our method in terms of the average precision n per query.

For the majority of queries, our method maximizes the cosine and SimRank.
This graph shows 22 requests (of 67 treated) for which our method maximizes the cosine and SimRank. This confirms that our proposed approach appears to have a positive effect in the majority of cases.

	COS	SimRank	**Our method**
MAP	0,064	0,075	**0,2360952**

Table 2 Mean Average Precision computed over all topics

• *The 11-point precision-recall curve:* The 11-point precision-recall curve is a graph plotting the interpolated precision of an IR system at 11 standard recall levels, that is, {0.0,0.1,0.2,...,1.0}. The method for interpolation is detailed below. The graph is widely used to evaluate IR systems that return ranked documents, which are common in modern search systems. the *interpolated precision Pinterp* at a certain recall level *r* is defined as the highest precision found for any recall level $r' \geq r$:

$$p_{interp}(r) = \max_{r' \geq r} p(r')$$

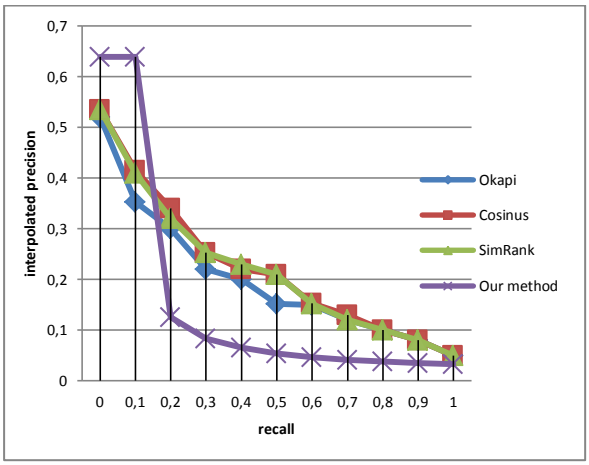

Figure 6 11-point precision-recall curve for Cosine, SimRank, Okapi and our method on Cisi

Figure 6 show that the three measures (Cosine, SimRank, Okapi) obtains similar results. Our method obtains the best scores when the recall is less than 20%. The SimRank achieve the best results when the return rate is between 20% and 30%. The results obtained by the three measures become almost identical when the number of documents returned increases.
Indeed, in our method, when the precision increases, the recall decreases and vice versa. This curve demonstrates that it is be possible to obtain high precision at the cost of low recall or a high recall of low precision prices. The advantage of this interpolation is that it permits to know the precision to standardized values.

V. DISCUSSION

The results obtained with the Quran corpus and test collection (CISI) are encouraging and variability shows that they are perfectible. The good performance demonstrates the merits of document Vector in capturing the semantics of documents and descriptors. Unlike [14], we don't use grammatical labels, we use an independent language surface analysis.

In contrast to [9][10], we use a variable length character N-gram extraction from the relevant word groups located between the borders as this strategy brings an interesting performance for languages (such as Arabic and Chinese) in which the words are not explicitly defined and different words are not separated by spaces, so a sentence is composed of many consecutive characters

VI. CONCLUSION AND FUTURE WORK

We have presented in this paper an approach for multilingual document representation which combines surface analysis and an LSA statistical algorithm for the detection of concepts in order to

create a semantic space that will serve as a pivot language for MIR. We have proposed a process of feature extraction founded on the variable length N-gram extraction from the relevant word groups located between the borders. This strategy brings an interesting performance for languages such as Arabic in which the words are not explicitly defined and different words are not separated by spaces.

The work described here opens up a number of new directions for future research. A first direction is to use the constructed pivot language for retrieval systems or machine translation to or from Arabic, French and English. The next step is to use our indexing process in a retrieval system to compare the speed and quality of results with other retrieval systems.

REFERENCES

[1] Emad E., Eissa A., and Hatem A.. Semantic Boolean Arabic Information Retrieval. The International Arab Journal of Information Technology, Vol. 12, No. 3, May 2015.

[2] SALTON, G., WONG, A., YANG, C. S., A Vector Space Model for Automatic Indexing. In Communication of the ACM, novembre 1975, Vol. 18, N°11. p 613-620.

[3] M. Baziz, M. Boughanem, N. Aussenac-Gilles. A Conceptual Indexing Approach based on Document content Representation. Dans : CoLIS5 : Fifth International Conference on Conceptions of Libraries and Information Science, Glasgow, UK, 4 juin 8 juin 2005. F. Crestani, I. Ruthven (Eds.), Lecture Notes in Computer Science LNCS Volume 3507/2005, Springer-Verlag, Berlin Heidelberg, p. 171- 186.

[4] Boulaknadel S., Béatrice D., Driss A.. A Multi-Word Term Extraction Program for Arabic Language. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco.

[5] Harish, B.S., Guru, D.S., Manjunath, S.. Representation and classification of text documents: a brief review. IJCA, Special Issue on RTIPPR 2, 110–119. 2010.

[6] Zhang, S., Boukamp, F., Teizer, J.. Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for job hazard analysis (JHA). Autom. Constr. 52, 29–41. 2015.

[7] Yen-HsienLee. Paul J.. Wan-JungTsao. , LiangLi. Use of a domain-specific ontology to support automated document categorization at the concept level: Method development and evaluation. Expert Systems with Applications. Volume 174, 15 July 2021, 114681.

[8] Kanaris, I., Kanaris, K., Houvardas, I., Stamatatos, E., 2007. Words versus character ngrams for anti-spam filtering. Int. J. Artif. Intell. Tools 16 (06), 1047–1067.

[9] McNamee and Mayfield J.. Character NGram Tokenization for European Language Text Retrieval. Information Retrieval 7, 73-97, 2004.

[10] Peter A. Chew, Brett W. B., Abdelali A.. Latent Morpho-Semantic Analysis: Multilingual Information Retrieval with Character N-Grams and Mutual Information. Proceedings of the 22nd International Conference on Computational Linguistics, August 2008.

[11] Laroum S., Béchet N., Hamza H. and Roche M., Classification automatique de documents bruités à faible contenu textuel, Manuscrit auteur, publié dans "RNTI : Revue des Nouvelles Technologies de l'Information. 2009.

[12] Rammal M., Sanan M.. Improving Arabic Information Retrieval System using n-gram method. Journal WSEAS Transactions on Computers. Volume 10 Issue 4, April 2011.

[13] Elghannam F., Text representation and classification based on bi-gram alphabet, Journal of King Saud University –Computer and Information Sciences, 33, 2021, 235–242.

[14] Balpe, J.P., Lelu, A. Papy, F., Techniques avancées pour l'hypertexte. Paris, Hermes. 1996.

[15] Shannon C. E.. Prediction and entropy of printed English. Bell System Technical Journal 30, pages 50 - 64.

[16] Cuna Ekmekcioglu F., Michael F. Lynch, and Peter W. . Stemming and N-gram Matching For Term Conflation In Turkish Texts, 1996.

[17] Yaël Ch.. Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information, 2009.