Advanced Search in Quran: Classification & Proposition of All Possible Features

Assem Chelli, Taha Zerrouki, and Amar Bala,

***Abstract*—Quran, the sacred book of Muslims, contains various information about all aspects of life: Scientific, Social, Historic, Politic...etc. It is so different of its language, structure, diversity. With a huge treasure of information, we can extract manually only a small part. The printed indexes can't help much since many search process waste the time and the force of the searcher. The simple search using exact query doesn't offer real options and still inefficient to move toward Thematic search by example. Because this limitation, we have to find out a new way to query. Our proposition is to design an information retrieval system that fits to the needs of the Quran. But to realize this objective, we must first list and classify all the search features that are possible and helpful. We wrote this paper to explain this point. The paper contains a listing for all search features that we have collected and a classification depending on the nature of feature and the way how can be implemented.**

***Index Terms*—Search Features, Quran, Query, Arabic, Information Retrieval, Search engines**

I. Introduction

QURAN, the sacred book of Muslims, contains various information about all aspects of life: Scientific, Social, Historic, Politic...etc. It is so different of its language, structure, diversity. With a huge treasure of information, we can extract manually only a small part.

By example, if you want to find a book of English grammar, you'll simply Google it, select a PDF and download it .that's all! Search engines (like Google) are utilized generally on Latin letters and for searching general information of document like content, title, author...etc. However, searching that way in the Quran is not sufficient because there is lots of information that must be extracted to fulfill Quran scholar's needs. Before computer, Quran scholars were using printed lexicons made manually. The printed lexicons can't help much since many search process waste the time and the force of the searcher. Each lexicon is written to reply to a specific query which is generally simple. Nowadays, there are applications that are specific for search needs; most of applications that were developed for Quran had the search feature but in a simply way: sequential search with regular expressions.

The simple search using exact query doesn't offer better options and still inefficient to move toward Thematic search by example. Full text search is the new approach of search that replaced the sequential search and which is used in search engines. Unfortunately, this approach isn't applied on The Quran. The Question is why we need this approach? Why search engines? Do Quran search applications really need to be implemented as search engines? The features of search that we'll mention them in this paper will answer those questions.

Our proposition is about design a retrieval system that fit the Quran search needs. But to realize this objective, we must first list and classify all the search features that are possible and helpful. We wrote this paper to explain this point. The paper contains a listing for all search features that we have collected and a classification depending on the nature of feature and the way how can be implemented.

We'll go through the problematic for a start. Secondly, we'll set out an initial classification and list all the search features that we think that they are possible. Furthermore, we'll introduce the API of the project Alfanous - Quranic Search Engine API - which demonstrates well some implemented search features.

II. Problematic

To clarify your vision about the problematic of this paper, we are describing the challenges that face the search in Quran:

- First, as a general search need;
- Second, as an Arabic search challenge;
- Third, Quran as a special source of information.

We start explaining the first point, the search in Quran is by theory has the same challenges of search in any other documents. The search in documents has passed by different phases in its evolution. At the beginning, the search was sequential based an exact keyword before the regular expressions were introduced. Full text search was invented to avoid the sequential search limitations on huge documents. The full text search introduces some new mechanisms for text analysis that include tokenization, normalization, and stemming...etc. Gathering Statistics make now a part of search process, it helps to improve the order of results and the suggestions. After the raising of the web semantic, the search is heading to a semantic approach where the to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable dataspace to generate more relevant results. To get more user experience, the search engines try to improve the behavior of showing the results by sorting it based on their relevance in the documents , more sorting criteria, Highlighting the keywords,

Assem Chelli, ESI - Algiers, e-mail: a_chelli@esi.dz.
Taha Zerrouki, ESI - Algiers, e-mail: taha.zerrouki@gmail.com.
Amar Bala, ESI - Algiers, e-mail: a_bala@esi.dz.

Pagination, Filtering and Expanding. Moreover, improve the query input by introducing different input methods (ex: Vocal input) and suggesting related keywords. Till now, most of these features are not implemented to use on Quran. And many of them need to be customized to fit the Arabic properties that what explained in the next point.

Secondly, Quran's language is considered as the classical Arabic. Arabic is a specific language because its morphology and orthography, and this must be taken into consideration in text analyzing phases. For instance, letters shaping (specially the Hamza -ء-), the vocalization, the different levels of stemming and types of derivations...etc. That must be taken into consideration in search features by example: the regular expressions are badly misrepresenting the Arabic letters since the vocalization diacritics are not distinct from letters. The absence of vocalization issues some ambiguities in understanding the words:

- الملك؟ المَلَك, المَلِك، المُلْك
- وعد؟ وَعَدَ ، وَ+عَدَّ
- وله؟ وَلَه، وَلِّ+ه ، وَ+لَ+هُ

Trying to resolve these problems as a generic Arabic problem is really hard since it hasn't sufficient linguistic resources to make strict lexical analyzers. By the contrary, Quran has a limited count of words and that means that it's possible to write manually morphological indexes and use it to replace lexical analyzers. Finally, we explain in this point what the specific challenges faced in search in order of the particular characteristic of Quran. El-Mus-haf , the book of Quran, is written on the Uthmani script . This last is full of recitation marks and spells some words in a different way than the standard way. By example, the word "بسطة" is spelled "بصطة" in Uthmani. The Uthmani script requires considering its specifications in Text analyzing phases: Normalization, Stemming. El-mus-haf is structured in many different ways: Juz'/Hizb/Sura/Aya, pages and themes...etc. the users may need to search, filter results or group them based on one of those structures. There are many sciences related to Quran, named Quranic Sciences: Tafsir, Translation, Recitation, Similitude and Abrogation…etc. Next, we'll propose an initial classification for the search features that we have inspired from the problematic points.

III. Classification

To make the listing of search features easier, we classified them in many classified based on their objectives.

1) **Advanced Query**: This class contains the modifications on simple Query in order to give the user the ability of formulating his query in a précised way. By example: Phrase search, Logical relations, Jokers.
2) **Output Improvement**: Those are to improve the results before showing it to users. The results must pass by many phases: Scoring, Sorting, Pagination, Highlighting...etc.
3) **Suggestion Systems**: This class contains all options that aims to offer a suggestion that help users to correct, extend the results by improving the queries. By example, suggest correction of misspelled keywords or suggest relative-words.
4) **Linguistic Aspects**: This is about all features that are related to linguistic Aspects like stemming, Selection&filtring stop words, normalization.
5) **Quranic Options**: It's related to the properties of the book and the information included inside. As we mentioned in the problematic, the book of Quran (al-mushaf) is written in uthmani script full of diacritization symbols and structured in many ways.
6) **Semantic Queries**: Semantic approach is about to allow the users to pose their queries in natural language to get more relevant results implicitly.
7) **Statistical System**: This class covers all the statistical needs of users. By example, searching the most frequented word.

This is an initial classification; we have to improve it for a well exploit of all possible search features.

IV. Proposition

In this point, we enlist all possible search features based on the classification we mentioned before. These entire features express a search need: general, related to Arabic or related to Quran. We have collected the basic ideas from:

- Classic & Semantic search engines: Google,
- Arabic search engines: Taya it,
- Quranic search tools: Zekr application, al-monaqeb alqurany,
- Indexing/Search programming libraries: Whoosh, Lucene
- Quranic Paper lexicons: المعجم - محمد فؤاد عبدالباقي المفهرس لألفاظ القرآن الكريم

We have manipulated those ideas to fit the context of Arabic and Quran. There are many features that are totally new , we propose them to fulfill a search need or resolve a specific problem. In addition to simple search, these are our propositions:

1) **Advanced Query**
   a) **Fielded search**: uses fieldname in the query to search in a specific field. Helpful to search more information like surah names.
      - سورة:الفاتحة
   b) **Logical relations**: to force the existence or absence of a keyword. The most known relations are: AND for conjunction, OR for disjunction and NOT for exception. The relations can be grouped using parenthesis.
      - (الصلاة - الزكاة) + سورة:البقرة
   c) **Phrase search**: is a type of search that allows users to search for documents containing an exact sentence or phrase.
      - "الحمد لله"

d) **Interval search:** used to search an interval of values in the numerical field. Helpful in fields like: Ayah ID, Page, Hizb, statistic fields.

- رقم_الآية:[١ إلى ٥]

e) **Regular expressions (Jokers):** used to search a set of words that share some letters. This feature can be used to search a part of a word. In latin , there is two Jokers used largely : ? replaces one letter, * replaces an undefined number of letters. These jokers are inefficient in Arabic because the existence of vocalization symbols which they are not letters, and Hamza(ء) letter that has different forms (different Unicode emplacements) .

- ب؟طة = بسطة، بصطة
- *نبي* = نبي ، النبيين ، الأنبياء ...

f) **Boosting:** used to boost the relevance factor of any keywords.

- سميع ٨ ٢ بصير

g) **Combining features:** search using a combination of the previous elementary features.

- "*حمد لله" ٨ ٢

2) **Output Improvments**

a) **Pagination:** dividing the results on pages.

- 10, 20,50... results per page

b) **Sorting**: sort the results by various criteria such as:

- Score
- Mushaf order
- Revelation order
- Numerical order of numeric fields
- Alphabitical or Abjad order of alphabetic fields
- A combination of the previous orders

for the alphabetical order ,we must consider the real order of:

- Hamza forms : ؤ ئ ء أ
- Ta' forms: ة ت
- Alef forms : ى ا

c) **Highlight:** used for distinction of searched keywords in Ayah.

- الحمد <style>لله</style>رب العالمين

d) **Real time output**: used to avoid the time of user waiting ,and show the results directly when retrieved.

e) **Results grouping:** the criteria can be used for grouping are:

- by similar Ayahs
- by Surahs
- by subjects
- by taffsir dependency
- by revelation events
- by Quranic examples

f) **Uthmani script with full diacritical marks:**

- [۞ لَّقَدْ كَانَ فِى يُوسُفَ وَإِخْوَتِهِۦٓ ءَايَـٰتٌ لِّلسَّآئِلِينَ]

3) **Suggestion Systems**

a) **Vocalized spell correction**: offers alternatives for keywords when misspelled or appeared with a different form in Quran.

- أبرَاهَام: إبْرَاهِيم

b) **Semantically related keywords**: used as hints, generally based on domain ontologies

- يعقوب: يوسف، الأسباط، نبي...

c) **Different vocalization**: suggest all the possible vocalizations sorted by the frequencies.

- الملك : المَلِك ، المُلْك، المَلَك ...

d) **Collocated words**: provides a completion based on the collocation statistics.

- سميع : سميع عليم، سميع بصير
- الحمد: الحمد لله

e) **Different significations:** used to limit a keyword on only one of their meanings.

- رب: معنى ١ (إله)، معنى ٢ (سيد)

4) **Linguistic Aspects**

a) **Partial vocalization search:** gives user the opportunity to specify some diacritics and not all.

i) مَـلـك_ to find مَلِك, مَلَك ... and ignore مُلْك

b) **Multi-level derivation**: the Arabic words derivations can be devided to four levels : exact word فأسقيناكموه , Word affixes removed (Lemma) أسقينا , Stem: أسقى , Root : سقي .

- (word:سقي , level: root) to find يَسْقُونَ, نَسْقِي, وَيَسْقِينِ , يُسْقَوْنَ , وَيُسْقَوْنَ, وَسُقْيَاهَا , وَأَسْقَيْنَاكُمْ , وَيُسْقَى , لَأَسْقَيْنَاهُمْ , وَنُسْقِيَهُ , فَأَسْقَيْنَاكُمُوهُ , فَسَقَى , السِّقَايَةَ, نُسْقِيكُمْ, فَيَسْقِي, سِقَايَةَ, تَسْقِي, سَقَيْتَ, يُسْقَى, اسْتَسْقَى, اسْتَسْقَاهُ, وَسُقُوا.
- (Word: أسقينا, level: lemma) to find وَأَسْقَيْنَاكُمْ , لَأَسْقَيْنَاهُمْ, فَأَسْقَيْنَاكُمُوهُ .

c) **Specific-derivations**: this is specification of the pervious feature .Since Arabic is fully flexional , it has lot of derivation operations .

- Conjugation in Past tense of قال to find قالت, قال, قالوا, قلن ...
- Conjugation in Imperative tense de قال to find قولوا , قل ...

d) **Word properties embedded query:** offers a smart way to handle words families by filtering using a set of properties like: root , lemma, type, part-of-speech, verb mood, verb form, gender, person, number, voice...etc.

- {جذر:ملك، نوع:اسم، عدد: مفرد }

e) **Numerical values substitution:** this helps to retrieve numbers, even if appeared as words.

- 309 replaces ثلاثمائة وتسعة

f) **Consideration/Ignoring spell faults:** especially for the letters that usually misspelled like Hamza(ء); The Hamza letter is hard to write its shape since its writing is based on the vocalization of both of it and its antecedent.

- مءصدة replaces مؤصدة
- ضحي replaces ضحى
- جنه replaces جنة

g) **Uthmani writing way:** offers the user the ability of writing words not as it appeared in Uthmani Mushaf .

- بصطة replaces بسطة
- نعمت replaces نعمة

h) **Resolving pronoun references:** Pronouns usually refer to other words, called their antecedents because they (should) come before the pronoun. This feature gives the oppurtunity of search using the antecedents.

- لا اله إلا هو, هو الله

5) **Quranic Options**

a) **Recitations marks retrieving:** helpful for Tajweed scholars.

- سجدة : نعم
- نوع _سكتة:واجبة
- قلقلة:نعم

b) **Structural option**: since Quran is devided to parts () and the part to Hizbs and Hizb to Halfs ... till we get Ayas.There is also other structures of Quran such as to pages ...etc.

- صفحة: ١
- حزب :٦٠

a) **Translation embedded search:** helps users to search using words translations to other langugaes instead of the native arabic text.

- { text: mercy , lang: english , author: shekir }

b) **Similitudes search** (المتشابهات)

- متشابهة _ {55,13}
  [فَبِأَيِّ آلَاءِ رَبِّكُمَا تُكَذِّبَانِ] , 31 exact similitudes

c) **Examples search** (الأمثال)

- مثل في سورة:البقرة
  [مَثَلُهُمْ كَمَثَلِ الَّذِي اسْتَوْقَدَ نَارًا فَلَمَّا أَضَاءَتْ مَا حَوْلَهُ ذَهَبَ اللَّهُ بِنُورِهِمْ وَتَرَكَهُمْ فِي ظُلُمَاتٍ لَا يُبْصِرُونَ]

6) **Semantic Queries**

a) **Semantically related words**: there are several different kinds of semantic relations: Synonymy, Antonymy, Hypernym, Hyponymy, Meronymy, Holonymy, Troponymy.

- Syn(جنة) to search for جنة, نعيم, فردوس ...
- Ant(جنة) to search for جحيم, سعير , جهنم, سقر ...
- Hypo(جنة) to search for عدن، فردوس ...
- ...

7) a) **Natural questions:** this means to offer the option of forming the search query as a form of an Arabic natural question . the most used Arabic question words are: هل؟ من؟ ما؟ أين؟ متى؟ لم؟ كم؟ أيّ؟ لمن؟

- من هم الأنبياء؟
  (Who are the prophets?)
  [إِنَّا أَوْحَيْنَا إِلَيْكَ كَمَا أَوْحَيْنَا إِلَى نُوحٍ وَالنَّبِيِّينَ مِنْ بَعْدِهِ وَأَوْحَيْنَا إِلَى إِبْرَاهِيمَ وَإِسْمَاعِيلَ وَإِسْحَاقَ وَيَعْقُوبَ وَالْأَسْبَاطِ وَعِيسَى وَأَيُّوبَ وَيُونُسَ وَهَارُونَ وَسُلَيْمَانَ وَآتَيْنَا دَاوُودَ زَبُورًا] - النساء ١٦٣
- ما هي الحطمة؟
  (What is Al-hottamat?)
  ؟ [نَارُ اللَّهِ الْمُوقَدَةُ] - الهمزة ٦
- أين غلبت/هزمت الروم؟
  (Where was Rome defeated?)
  [فِي أَدْنَى الْأَرْضِ وَهُمْ مِنْ بَعْدِ غَلَبِهِمْ سَيَغْلِبُونَ] - الروم ٣
- كم مكث أصحاب الكهف؟
  (How much of time did People of Cave stay?)
  [وَلَبِثُوا فِي كَهْفِهِمْ ثَلَاثَ مِائَةٍ سِنِينَ وَازْدَادُوا تِسْعًا] - الكهف ٢٥
- متى يوم القيامة؟
  (When is the Day of Resurrection?)
  [يَسْأَلُكَ النَّاسُ عَنِ السَّاعَةِ قُلْ إِنَّمَا عِلْمُهَا عِنْدَ اللَّهِ وَمَا يُدْرِيكَ لَعَلَّ السَّاعَةَ تَكُونُ قَرِيبًا] - الكهف ٢٥
- كيف يتشكّل الجنين؟
  (How has the embryo be formed?)
  [ثُمَّ خَلَقْنَا النُّطْفَةَ عَلَقَةً فَخَلَقْنَا الْعَلَقَةَ مُضْغَةً فَخَلَقْنَا الْمُضْغَةَ عِظَامًا فَكَسَوْنَا الْعِظَامَ لَحْمًا ثُمَّ أَنْشَأْنَاهُ خَلْقًا آخَرَ فَتَبَارَكَ اللَّهُ أَحْسَنُ الْخَالِقِينَ] - المؤمنون ١٤

b) **Automatic diacritization :** the absence of diacritics lead to the ambiguities that we've mentioned them in Problematic. This feature helps to pass over these ambiguities and try to resolve them before executing the search process.

- رسول من الله == رَسُول مِنَ اللهِ

c) **Proper nouns search** :lot of proper nouns are mentions clearly in Quran but some of them are just referred to implecitly . This feature is to search for proper nouns however mentioned : explicitly or implicitly.

- بنيامين؟
  [إِذْ قَالُوا لَيُوسُفُ وَأَخُوهُ أَحَبُّ إِلَى أَبِينَا مِنَّا وَنَحْنُ عُصْبَةٌ إِنَّ أَبَانَا لَفِي ضَلَالٍ مُبِينٍ] - المؤمنون ١٤

- أبو بكر/ الصديق؟
[ثَانِيَ اثْنَيْنِ إِذْ هُمَا فِي الْغَارِ إِذْ يَقُولُ لِصَاحِبِهِ لَا تَحْزَنْ إِنَّ اللَّهَ مَعَنَا] - التوبة ٤٠

8) **Statistical System**

a) **Unvocalized word frequency:** this feature is about gathering the frequency of each word per ayahs ,per surahs , per hizbs.
- How many words of "الله" in Sura "المجادلة"?
- What the first ten words which are the most frequently cited words in the whole Quran?

b) **Vocalized word frequency:** the same as the presvios feature but with consideration of diacritics.that will make the difference for instance between مَنْ ,مِنْ and مَنَّ .

c) **Root/Stem/Lemma frequency:** the same as the previous feature but use Root, Stem or Lemma as unit of statistics instead of the vocalized word.
- How many the word of "بحر" and its derivations are mentioned in the whole Quran? the word "بحار" will be considered also.

d) **Another Quranic units frequency:** the statistics could be gathered based on many units otherwise the words like letters,ayas, Recitation marks...etc.
- How many letters in the Sura "طه"?
- What's the longest Ayah?
- How many Marks of Sajda in the whole Quran?

V. Alfanous API as a prototype

To validate our vision about search features on Quran, we work in implementing them through the API of Alfanous. Alfanous is an open source project that aims to build an indexing/search API in Quran and use it to develop different interfaces: Web, Desktop...etc. Alfanous API is pure python, prototyped from the library whoosh which is a fast, featureful full-text indexing and searching library implemented in pure Python. Programmers can use it to easily add search functionality to their applications and websites. Every part of how Whoosh works can be extended or replaced to meet specific needs exactly. The API helps the researchers and developers that working on Quran applications and want to take the benefits of search. Alfanous API is:

- Easy to use, by virtue of a JSON output system that interacts with the API. This last can be used as a wrapping to other programming languages
- Easy to customize(prototyping), because the simplicity and clarity of the code
- Free Open Source, developed under the license AGPL (Affero General Public License)

In this second table, we'll mention some specific queries already implemented in the API:

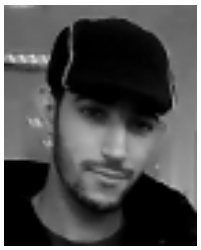
Table I: Advanced Query in Alfanous

Feature	Syntax
simple search	سماكم
derivation level 1	>سماكم
derivation level 2	>>سماكم
joker ?	ال؟لك
joker *	*نبي*
phrase	"رب العالمين"
logical relation : AND	الزكاة + الصلاة
logical relation : OR	الزكاة | الصلاة
logical relation : ANDNOT	الزكاة - الصلاة
synonyms	~السعير
fields	sura:الفاتحة
interval	sura_id:[112 to 114]
vocalization	'المَلك'
word properties tuple	{ملك,فعل}
transliteration: buckwalter	fawoqa

VI. Related Works

VII. Conclusion

In this paper, we have enlisted the search features in Quran that it's helpful. To facilitate the enlisting, we have classified those features depending on the nature of the problem and the way how can be fixed. For that, we designed a small prototype for a retrieval system to explain why we have done this classification. That list will help us to make a detailed retrieval system that fit perfectly the needs of the Quran. Each feature has a different level of complexity: could be implemented easily or may lead to vast problem that need a deeper study.



Assem Chelli A Master student from National higher school of computer science (ESI) - Algiers. Doing his research about the indexing & search in Arabic Generally, Quran Specifically. To validate his work, he implemented a Quranic search engine API that is easily-costumizable. He launched this API as an Open source project called Alfanous.