ISSN 1870-4069

A Hybrid Approach to Extract Key Phrases from Arabic Text

Reda Ahmed-Zayed, Mohamed Farouk Abdel-Hady,
Hesham A. Hefny

Cairo University, Institute of Statistical Studies and Research,
Egypt

Reda_fcis@yahoo.com,
mohamed.abdel-hady@alumni.uni-ulm.de,
hehefny@ieee.org

Abstract: Key phrases are the phrases, consisting of one or more words, representing the important concepts in the document. This paper presents a hybrid approach to key phrase extraction from Arabic text. The proposed approach is an amalgamation of three methods: The first one is assigning weights to candidate key phrases based on term frequency and inverse document frequency, the second one is assigning weights to candidate key phrases using some knowledge about their similarities to the structure and characteristics of key phrases available in the memory (stored list of key phrases), and the third one is assigning weights to candidate key phrases using some knowledge about fatwa label (class or fatwa area). Also, an efficient candidate key phrase identification method has been introduced in this paper. The experimental results show that the proposed hybrid approach gives good performs.

Keywords: Arabic text mining, information retrieval, key phrase extraction automatic indexing, Arabic Islamic religion domain.

1 Introduction

The task of extracting key phrases from free text documents is becoming increasingly important as the uses for such technology expands. Key phrases are listed of phrases or key words composed of about five to fifteen important words and phrases that express the main topics discussed in a given document or article. Key phrases are useful for a variety of tasks such as text summarization, automatic indexing, clustering or classification, text mining.

It can provide the automation of generating metadata that gives a high-level description of a document's contents, highlighting important topics within the body of the text, summarizing documents for prospective readers, measuring the similarity between documents, making it possible to cluster and categorize documents, and searching more precise upon using them as the basis for search indexes or as a way of browsing a collection of documents [3].

Fig. 1. Key phrase extraction system overview.

When a Muslim has a question that they need to be answered from an Islamic point of view, they ask an Islamic scholar this question, and the answer is known as a" fatwa". It is similar to the issue of legal opinions from courts in common-law systems. A fatwa in the Islamic religion represents the legal opinion or interpretation that a qualified jurist or mufti can give on issues related to the Islamic law.

The fatwa request has to be directed to the most relevant mufti, and this task is first task we make in our proposed application [5]. Mufti start to answer the fatwa and manual assign key phrases for it. The main contribution of this paper is applying a hybrid approach for key phrase extraction using domain knowledge base and statistical information about the fatwa details depend on term frequency and inverse document frequency.

This paper and points to future work the most key phrase extraction systems which are proven to be successful have used supervised machine learning techniques. The main advantages of supervised machine learning techniques are that they can adapt to the specific nature of documents.

2 Related Work

Arora et al. [12] they proposed a new clustering algorithm based on the Kea Key phrase algorithm that used here to extract several Key phrases from source Text documents by using machine learning techniques.

The show that The Kea bisecting K-means clustering algorithm gives easy and efficient way to extract text documents from large amount of Text documents, there results showed that kea can an average match between one and two of the given key phrases chosen. The consistently good quality of the clustering that it produces, bisecting K-means is an excellent algorithm for clustering a large number of documents.

El-Beltagy et al. [2] presented the KP-Miner system, and demonstrated through experimentation and comparison with widely used systems that it is effective and efficient in extracting key phrases from both English and Arabic documents of varied length. Unlike other existing key phrase extraction systems, the KP-Miner system does not need to be trained on a particular document set in order to achieve its task. It also has the advantage of being configurable as the rules and heuristics adopted by the system are related to the general nature of documents and key phrases.

Gollapalli et al. [11] they explore a basic set of features commonly used in NLP tasks as well as predictions from various unsupervised methods to train their taggers. In addition to a more natural modeling for the key phrase extraction problem, they showed that tagging models yield significant performance benefits over existing state-of-the-art extraction methods.

3 System Overview

The architecture of the proposed key phrase extraction system is shown in Figure 1. The aim of this system is to automatically generate key phrase for a fatwa (legal opinion) requests. Each fatwa is associated with a category (Fatwa areas) by Muslim Scholar or our Proposed Routing System [5].

The Key phrase extraction in the proposed system is a two main phases. The first phase is text preprocessing and feature engineering for fatwa text to select the feature vector which represent each class this step is called Generating Knowledge Base and domain context. The second phase is a three-step process: candidate key phrase selection, candidate key phrase weight calculation and finally key phrase refinement. Each of these steps, is explained in more details about the employed algorithm.

4 Generating Knowledge Base and Domain Context

We need to build a bag of words for each fatwa class (category), to elect the list of words than can be considered as a feature vector for each category we need some process to extract this feature vector, these details explained in details in the next section.

4.1 Text Preprocessing

The nature of the Arabic text is different than the English text, preprocessing of the Arabic text is more challenging. A huge number of features or keywords in the documents lead to a poor performance in terms of both accuracy and time. Therefore, preprocessing is a very important step before training the text classifiers to get knowledge from massive data and reduce the computational complexity. Before Arabic word stemming step, fatwa requests are normalized as follows:

- Remove Fatwa Question introduction from start to word "المتضمن:".
- Remove punctuation.
- Remove special characters and remove any HTML tags.
- Remove diacritics (primarily weak vowels).
- Remove non-Arabic letters.
- Replace Arabic letter ALEF with hamza below, Arabic letter ALEF with madda above, and Arabic letter ALEF with hamza above with a Arabic letter ALEF.
- Replace final Arabic letter Farsi YEH with Arabic letter YEH.
- Replace final Arabic letter TEH marbuta with Arabic letter HEH.
- Stop-word removal: we determine the common words in the documents which are not specific or discriminatory to the different classes.
- Stemming: different forms of the same word are consolidated into a single word. For example, singular, plural and different tenses are consolidated into a single word.

Table 1. Number of fatwa belong to main fatwa class in the KP data set.

	Remove prefixes	**Remove Suffixes**
Light 1	ال،و ال،بال،كال،فال	None
Light 2	ال،و ال،بال،كال،فال،و	None
Light 3	،،	ه،ة
Light 8	،،	ها ، ان ، ات ، ون ، ين ، يه ، ية ، ه ، ة ، ي
Light 10	ال،و ال،بال،كال،فال،و ،مل،و	

Light Stemmer Larkey et al. [13] developed several light stemmers for Arabic, and assessed their effectiveness for information retrieval using standard TREC data. They have compared light stemming with several stemmers based on morphological analysis. The light stemmer, Light10, outperformed the other approaches. It has been included in the Lemur toolkit, and is becoming widely used for Arabic information retrieval. They tried several versions of light stemming, all of which followed the same steps:

– Remove Arabic letter WAW (and) for Light2, Light3, and Light8 if the remainder of the word is three or more characters long. Although it is important to remove Arabic letter WAW, it is also problematic, because many common Arabic words begin with this character, hence the stricter length criterion here than for the definite articles.
– Remove any of the definite articles if this leaves two or more characters.
– Go through the list of suffixes once in the (right to left) order indicated in figure below, removing any that are found at the end of the word, if this leaves 2 or more characters. The strings to be removed are listed in Fig. 2. The prefixes are actually definite articles and a conjunction.

4.2 Feature Engineering and Class Representation

Before any key phrase task or classification task, we need to represent each class (fatwa category) by feature vector. One of the most fundamental tasks that need to be accomplished is that of document representation and feature selection. We try to build lexicon for each class, each class can be reprinted by feature of vector. this vector is set of words; each text instance has to be represented as a fixed-length numeric feature vector which are mostly the text words.

This kind of text representation typically leads to high dimension input space. While feature selection is also desirable in other classification tasks, it is especially important in text classification due to the high dimensionality of text features and the existence of irrelevant (noisy or not important) features. Several methods are used to reduce the dimensionality of the feature space by choosing a subset of features in order to reduce the classification computational complexity without scarifying the accuracy. In this

paper, Chi-Squared ($\chi 2$) statistics [1] used as a scoring function to rank the features based on their relevance to the categories.

In general, text can be represented in two separate ways. The first is as a bag of words in which a document is represented as a set of words, together with their associated frequency in the document. Such a representation is essentially independent of the sequence of words in the document (context independent).

The second method is to represent each document as strings of words (called N-grams such as bigrams and trigrams), in which each document feature represents a sequence of words (it takes the context into consideration). In this paper, the bag-of-words representation is used as it has shown good key Phrase extraction performance.

5 Proposed Approach to Key Phrase Extraction

Proposed key phrase extraction consists of three primary components: document pre-processing we discuss at previous section, candidate key phrase identification and assigning scores to the candidates for ranking.

5.1 Candidate Key Phrase Identification

We follow a simple and knowledge poor approach to candidate key phrase identification is adopted as the first step of the proposed system. This approach is a variant of the candidate key phrase identification approach presented in [4]. A candidate key phrase is considered as a sequence of words containing no punctuations and stop words. A list of common verbs is also added to the stop word list because it is observed that the author assigned key phrases rarely contains common verbs. The process of candidate key phrase extraction has two steps:

Step1: extraction of candidate key phrases considering punctuations and stop words as the phrase boundary, Step2: Breaking further the phrases selected at the step one into smaller phrases using the following rules:

i. If a phrase is L -word long, all n-grams (n varies from 1 to L-1) are generated and added to the candidate phrase list,

ii. If a phrase is longer than five words, it is discarded.

Figure 3 shows a sample sentence and the candidate Key phrases identified from this sentence. Some candidate phrases generated using the above-mentioned method may not be meaningful to human readers. For example, in figure1, the candidate phrase "بصفة دائمة" is less meaningful. After computing phrase frequency and phrase weight, such kind of candidate key phrases are filtered out. For this purpose, some conditions are applied.

Condition one is to choose threshold on the phrase weight (Phrase weighting scheme has been presented in the next subsection which is a function of phrase frequency, inverse document frequency, domain knowledge etc. The second condition is related to the first appearance of the phrase in the document. Previous works [9] have suggested that key phrases appear sooner in an article. The works in El-Beltagy [2] states that a phrase occurring the first time after a predefined threshold is less likely a key phrase.

Sample Fatwa

بسم الله الرحمن الرحيم. الحمد لله وحده والصلاه والسلام علي من لا نبي بعده سيدنا محمد رسول الله وعلي اله وصحبه ومن تبعه باحسان الي يوم الدين. اطلعنا علي الطلب المقدم من/ محمود عزمي احمد ابو العزم المقيد برقم 126 لسنه 2006 م المتضمن: أُجريت لي عمليه جراحيه في البروستاتا والمثانه مما ادي بعد الشفاء من الجراحه الي خروج قطرات بول مني بصفه دائمه وعدم التحكم فيه بعد الاستنجاء، مما يضع النفس في حيره وشك في الوضوء والصلاه . نرجو الافاده ، وكيف يصح الوضوء والصلاه ؟

Fatwa Question without introduction

أُجريت لي عمليه جراحيه في البروستاتا والمثانه مما ادي بعد الشفاء من الجراحه الي خروج قطرات بول مني بصفه دائمه وعدم التحكم فيه بعد الاستنجاء، مما يضع النفس في حيره وشك في الوضوء والصلاه . نرجو الافاده ، وكيف يصح الوضوء والصلاه ؟

Initial list of candidate key phrases (after step1).

اجريت,عملية جراحيه ,البروستاتا والمثانه ,ادي ,الشفاء ,الجراحه,خروج قطرات بول ,بصفة دائمة , التحكم ,الاستنجاء,يضع النفس,حيرة وشك ,الوضوء.

The list of candidate phrases (after step2).

اجريت,عمليه,جراحيه ,عمليه جراحيه,البروستاتا,والمثانه,البروستاتا والمثانه,ادي,الشفاء,الجراحة,خروج,قطرات, بول ,خروج قطرات,خروج قطرات بول,قطرات بول,بصفه,دائمه,بصفه دائمه,التحكم,الاستنجاء,يضع,النفس,يضع النفس,حيره,وشك,حيره وشك,الوضوء

Fig. 2. A sample fatwa and the candidate key phrases identified from this fatwa.

A threshold is set on the position of the phrases where the phrases are numbered sequentially and the first phrase in the document is numbered as 1 and the last phrase is numbered as N. If a phrase appears first after the given threshold it is ignored, that is, if a phrase X appears first at pos i and the threshold value is set to T_{pos} and $i > T_{pos}$, the phrase X is discarded.

5.1.2 Assigning Scores to Candidate Key Phrases

In general, a fatwa has a few numbers of author assigned key phrases. To select a small subset of candidates as the key phrases requires assigning weights to the candidates and raking them based on these weights.

5.1.2.1 Assigned Score Phrase Frequency, Inverse Document Frequency

The weight of a candidate key phrase is computed using three important features: phrase frequency, inverse document frequency and domain specificity. Weighting using phrase frequency (PF) and inverse document frequency (IDF). The score for a candidate key phrase due to PF and IDF features is computed using the following formula:

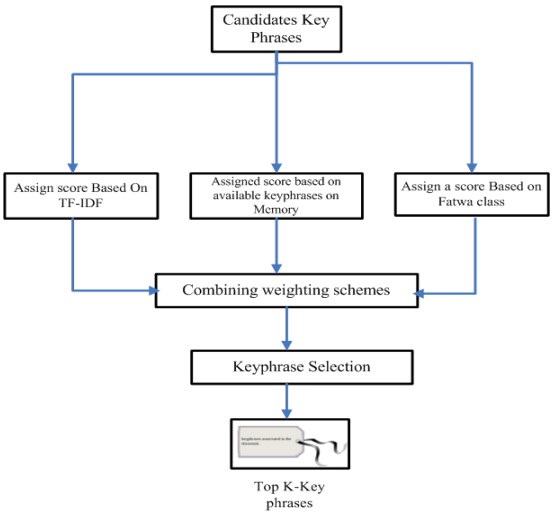

Fig. 3. Assign score to candidate key phrases.

$$Score PF \times IDF = \begin{cases} PF \times IDF, if\ plength = 1, \\ PF \times log(N), if\ plength > 1, \end{cases}$$

where:

$plength$= length of the phrases in terms of words PF = phrase frequency which is counted as number of times a phrase occurs in a document.

IDF= log(N/DF), where N is the total number of documents in the corpus (a collection of documents in a domain under consideration) and DF is the number of documents in which a phrase occurs at least once. Equation (1) shows that for multi-word phrases, phrase score is computed using PF × log (N), which is basically PF * IDF with DF set to 1. This is due to the fact that multi-word phrases do not occur as frequently within a relatively small collection of documents as do single-word phrases.

5.1.2.2 Using Domain Knowledge for Weighting Vandidate Key Phrases

We assigned score to each candidate key phrase using two condition, the first condition depend on if the key phrase word appear in the list of words that represent the class of fatwa as we mention the previous section, and the second condition depend on if the key phrase word appear in the list of key Phrase in the memory. This list contains manual key phrase extracted from fatwa and all possible key phrases generated from fatwa title.at next section we will discuss how assign weight to candidate key phrase in details.

5.1.2.3 Assigned Score based on Available Key Phrases on Memory

A score is assigned to a candidate key phrase based on how much it is similar to the structure and characteristics of key phrases available in the memory (a stored list of key phrases for the domain under consideration). For this purpose, a key phrase list is created with readily available author assigned key phrases collected from predefined key phrase for Fatwa. A predefined key phrase list is used to create a domain specific

glossary database giving some knowledge about the structure and characteristics of key phrases.

A list of key phrases is collected for creating glossary database are not included in the set of fatwas (the test set) on which the proposed key phrase extraction system is tested. However, using such a list of key phrases stored in the memory for weighing the candidate key phrases can be considered as some sort of partial supervision provided to the key phrase extraction system. The use of this kind of knowledge base in key phrase extraction task has previously been investigated in Sarkar [6, 10]. A variant of the method presented in Wu [10] is used for the proposed domain specific key phrase extraction task.

From the key phrase list, two tables are created: table1 is the keyword table, which is created by splitting the key phrases belonging to the key phrase list into words that can be called as keywords. This table has two columns (keyword, weights) and table2 is key sub-phrase table which consists of all sub phrases generated from the key phrases in the key phrase list. For any manual key phrase in the key phrase list, all possible n-grams (n varies from 2 to n) are generated and included in key sub-phrase table. The key sub-phrase table has also two columns (sub-phrase, weights) [6].

Weights for keywords in the keyword table are assigned using the following rules:

- If a keyword appears always alone independently in the key phrase list it is assigned a score of 1.
- If the keyword appears always as part of another key phrase, that is, if it has no independent existence in the key phrase list, it is assigned a score, which is computed as 1/log(c), where c is the number of times the keyword appears as the part of key phrases. Here it is assumed that a keyword, which has no independent existence and repeats many times in the key phrase-list only as the parts of other key phrases, is less domain specific.
- If the keyword appears independently in the key phrase list in some cases and also appears as part of key phrases in some other cases, it is assigned a score which is computed based on the formula: 0.5 × (1 + 1 / log (c)), where c is the number of times the keyword occurs as the part of key phrases.

The Weight of a sub-phrase or a phrase in the key sub-phrase table is computed by summing up the weights of the keywords of the sub-phrase or the phrase. The keyword table and the key sub-phrase table are used as domain knowledge in computing a score for a candidate key phrase. The score for a candidate phrase is computed using the following equation:

$$score_{D} = \sum_{i=1}^{m} K_i + \sum_{j=1}^{pc}, \tag{1}$$

where:

K_i = the weight of the $i-th$ keyword in the candidate keyphrase.
P_j= weight of the $j-th$ sub-phrase associated with candidate keyphrase.
M = the number of keywords in a candidate keyphrase.
PC= the number of sub-phrases generated from the candidate keyphrase.

The sub-phrases of the candidate key phrase are generated by computing all possible n-grams, where n varies from two to length of the phrase. When n is set to the length of the candidate key phrase, the n-gram is the candidate key phrase. The reason for taking the weights of all possible sub-phrases in calculating the candidate key phrase score, in addition to the weights of individual words, is to decide whether a sub-phrase is a manual key phrase in the key phrase table.

If it is, this candidate key phrase is assumed more important. This feature will favor those candidate key phrases which itself or whose parts are found in the knowledge base. Availability of a phrase or its sub-phrases in the knowledge base provides some evidence in support of key phrase worthiness of a phrase. Thus, with this knowledge base, the proposed key phrase extraction system is provided with some sort of partial supervision.

5.1.2.4 Assign a Score based on Fatwa Class

A score is assigned to a candidate key phrase based on fatwa class. For this purpose, a list of words is selected as feature vector which represent each class. Every class repented by 100 words, we make a feature selection using TF×IDF method. Each word repents presenting the class, for example word"المياة " represent class "الطهارة" by 100% and "الصلاة" by 80% and so on. Weights for keywords in the keyword table are assigned using the following rules:

– For each word in a candidate key phrase if the word appears in fatwa the weight formula ass following:

$$Score_C = \sum_{i=1}^{m} \frac{w[i]_{idf(c)}}{max\ w_{idf(c)}}, \tag{2}$$

where w_{tf} : the idf of word of i in the given class.

max w_{tf} : The max idf value for words in the given class words

5.1.2.5 Combining Weighting Schemes

The three weighting schemes have already been discussed in the previous subsections. These three types of scores should be combined to assign a unique score to each candidate key phrase. The combined score for a candidate key phrase is computed using the following linear combination of three scores:

$$SCORE = \alpha \times Score_{pf \times idf} + (1 - \alpha) Score_D\ \alpha \times Score_K +, \tag{3}$$

where:

$Score_{pf*idf}$ The score based on phrase frequency and inverse document frequency computed using the equation (1).

$Score_D$ The score based on domain specificity of a phrase computed using the equation (2).

$Score_c$ The score based on domain specificity of a phrase computed using the equation (3).

α Is the tuning parameter whose value is decided through experimentation the best results are obtained when α is set to 0.6 [6].

5.1.2.6 Extracting and Select Key Phrases

After assigning scores to the candidate key phrases, the next step is to select K top-ranked candidate key phrases as the final list of key phrases. The user can specify the value of K.

5.1.2.7 Performance Evaluation

The dataset used in the experiments was provided by the Egyptian Dar al-Ifta. Dar al-Ifta al Misryyah[1] started as one of the divisions of the Egyptian Ministry of Justice. In view of its consultancy role, capital punishment sentences among others are referred to the Dar al-Ifta al- Misryyah seeking the opinion of the Grand Mufti concerning these punishments. The dataset contains about 100,000 text instances

In order to further reduce the computational complexity of classification; feature selection was applied as follows: We used $Chi - square$ feature-ranking method. Separately for each main label (fatwa class) in order to obtain a ranking of all features for that label.

We then selected the top 100 features for each label. After the aforementioned preprocessing, and the removal of empty examples (examples with no features or labels, or multi question fatwa) the final version of the dataset included 1215 instances. The following table represents each main label and related fatwa. Each label has numbers of child we ignore the child labels and combine fatwa to the parent node of the leaf.

We merge the leaf nodes that does not contain any child in the parent node and consider the parent is the fatwa label. When we try to extract the key phrase, we have the fatwa label from the fatwa routing system. according this label, we get the label words.

Each word has a value that value represents a certain percentage in a certain class. This value we make consideration when we calculate the word weight in the candidate key phrase.

6 Experiments and Results

For extracting the key phrases from a test fatwa, the proposed system identifies first the candidate key phrases, computes phrase weight using the equation (3), filters out noisy phrases based on two conditions discussed in the previous section and assigns scores to the remaining candidate key phrases. Finally, the top-ranked K candidate key phrases are selected as key phrases.

To filter out noisy phrases, two conditions discussed in subsection 3.2 are applied here in the pre-specified order as follows: (1) Phrases whose position of the first occurrence in the document is greater than $Tpos$ are discarded. The value of Tpos is set to 100 to obtain the best results on the dataset used for the proposed work, (2) the threshold value on the phrase weight is adjusted to keep those candidate key phrases which occurs at least twice in a document or which has higher similarity to the phrases in the manually created knowledge base.

[1] http://dar-alifta.org

Table 2. Number of fatwas belong to main fatwa class in the KP data set.

Main Class Name	Arabic Name	Fatwa count
Purity(alttahara)	الطهارة	77
Prayer(alssala)	الصلاة	224
Funerals(aljanayiz)	الجنائز	123
Zakat(alzzaka)	الزكاة	537
Fasting(alssiam)	الصيام	74
Hajj and Umrah(alhajj waleumra	الحج والعمرة	104
Dikher and pray(aldhdhikr walddiea)	الذكر والدعاء	146

Result of Top 5 key Phrase.

Class	Purity	Prayer	Funeral	Zakat	Fasting	Hajj	Dikher & pray
Precision	90.15%	91.35%	85.32%	90.15%	74.1%	82.41%	60.23%
Recall	89.34%	87.21%	82.53%	92.65%	86.85%	83.73%	67.62%

Result of Top 10 key Phrase.

Class	Purity	Prayer	Funeral	Zakat	Fasting	Hajj	Dikher & pray
Precision	95.2%	92.5%	90.12%	96.25%	78.36%	85.32%	75.15%
Recall	96.5%	90.25%	88.23%	95.1%	80%	86.23%	70.32%

Result of Top 15 key Phrase.

Class	Purity	Prayer	Funeral	Zakat	Fasting	Hajj	Dikher & pray
Precision	97.35%	96.1%	93.62%	96.95%	84.56%	92.45%	86.15%
Recall	97.3%	93.1%	91.33%	96.63%	86.72%	89.11%	76.42%

We compare the output result to the mufti assigned key phrase the similarity between automated key phrase extraction and manual key Phrase and the flowing tables show the result of Performance over combined keywords when extracting, 5, 10, and 15 key phrases.

7 Conclusion

This paper discusses a hybrid key phrase extraction approach in the Islamic fatwa domain. The proposed approach combines domain knowledge with the features namely phrase frequency, inverse document frequency and phrase position in a more effective way. The proposed approach results in an easy-to-implement key phrase extraction system that outperforms some state-of-the art key phrase extraction systems. The experimental results also suggest that the proposed key phrase extraction method is effective in Islamic fatwa domain and incorporation of domain knowledge as partial supervision boosts up the system performance.

References

1. Mesleh, A.: Chi square feature extraction based SVMs arabic language text categorization system. Journal of Computer Science, vol. 3, no. 6, pp. 430–435 (2007)
2. El-Beltagy, S. R., Rafea, A.: KP-miner: A keyphrase extraction system for english and arabic documents. Information Systems, vol. 34, no. 1, pp. 132–144 (2009) doi: 10.1016/j.is.2008.05.002
3. El-Shishtawy, T., Al-Sammak, A.: Arabic keyphrase extraction using linguistic knowledge and machine learning techniques. arXiv preprint arXiv, 1203.4605 (2012) doi: 10.48550/arXiv.1203.4605
4. Kumar, N., Srinathan, K.: Automatic keyphrase extraction from scientific documents using N-gram filtration technique. In: Proceedings of the eighth ACM symposium on Document engineering, 199–208 (2008) doi: 10.1145/1410140.141018
5. Zayed, R. A., Hady, M. F. A., Hefny, H.: Islamic fatwa request routing via hierarchical multi-label arabic text categorization. In: Proceedings of First International Conference on Arabic Computational Linguistics (ACLing), IEEE pp. 145–151 (2015) doi: 10.1109/ACLing.2015.28
6. Sarkar, K.: A hybrid approach to extract keyphrases from medical documents. arXiv preprint arXiv, pp. 1303–1441 (2013) doi: 10.5120/10565-552
7. Turney, P. D.: Extraction of keyphrases from text: evaluation of four algorithms (1997) doi: 10.48550/arXiv.cs/021201
8. Turney, P. D.: Learning algorithms for keyphrase extraction. Information retrieval, vol. 2, no. 4, pp. 303–336 (2000) doi: 10.1023/A:1009976227802
9. Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C. G.: KEA: Practical automatic keyphrase extraction. In: Proceedings of the fourth ACM conference on Digital libraries (1999)
10. Wu, Y. F. B., Li, Q.: Document keyphrases as subject metadata: incorporating document key concepts in search results. Information Retrieval, vol.11, no.3, pp. 229–249 (2008) doi: 10.1007/s10791-008-9044-1
11. Das-Gollapalli, S., Li, X. L.: Keyphrase extraction using sequential labeling. arXiv preprint arXiv:1608.00329 (2016) doi: 10.48550/arXiv.1608.00329
12. Arora, A., Er-Abhishek, C.: Keyphrase extraction algorithm, PARIPEX-Indian Journal of Research, vol. 5, no. 4 (2016)
13. Larkey, L. S., Ballesteros, L., Connell, M. E.: Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Rerieval (SIGIR 2005), pp. 275–282 (2005) doi: 10.1145/564376.5644