Comprehensive Cluster Labels Generation for Arabic Documents

Fatma Elghannam

Abstract— The process of picking descriptive labels for a cluster of documents is called cluster labeling. A major challenge of automatic cluster labeling is the higher redundant information that can be found, whereas only few numbers of labels are required to represent a cluster. In this paper, a new technique for automatic cluster labeling is introduced. The main concepts of text documents cluster are represented by keyphrases that maximize both the relevance and coverage to the cluster. The experimental results on Arabic documents proved that the proposed technique is more efficient to generate comprehensive cluster labels than the commonly used term frequency method.

Index Terms— cluster labeling, indexing, summarization, information retrieval, Arabic documents.

1 Introduction

With the exponential growth of information on the Internet, finding and organizing relevant documents have become very important. Clustering is used in organizing textual data, where similar documents are grouped into one cluster. Text clustering may be used for different tasks, such as grouping similar documents (news, tweets, etc.) and the analysis of customer/employee feedback, discovering meaningful implicit subjects across all documents. In such settings, clusters must be labeled, so that user can interact with the cluster to identify and focus on the relevant set of results. The process of picking the most descriptive, human-readable labels from a cluster of documents is called cluster labeling. Standard clustering algorithms do not typically produce such labels. Cluster labeling algorithms extract labels that summarize the main topics of the cluster.

Cluster labeling approaches distinguish two basic categories: cluster-internal labeling selects labels based only on the contents of the cluster of interest, whereas differential cluster labeling labels a cluster by comparing the terms in one cluster with the terms occurring in other clusters [1].

The major challenge of cluster labeling is due to the multiple resources from which information is extracted. A cluster of documents that deal with the same main topic include the risk of higher supplementary topics with multiple aspects than would typically be found in a single document. So the key tasks are not only identifying frequent terms/words across documents, but also recognizing novelty and ensuring that the final extracted labels are both coverage and relevant. An additional challenge of cluster labeling is how to extract meaningful informative candidate labels. The most common approach for cluster labeling works by picking up the most frequent terms occurring in a cluster, or using top weighted cluster centroid's terms [2]. A main drawback of these methods consists in that individual words may not produce an optimal solution for extracting meaningful labels from the documents making up a cluster [3]. Existing research has reported that phrases [4] are more informative than keywords for understanding. Moreover, linguistic knowledge about words cannot be neglected; it plays an essential role in the determination of valid significant keyphrases/labels [5, 6]. It affects the resulted labeling performance because it allows a search term to focus more on the meaning of a term and closely related terms instead of specific character matches.

In this paper an approach for cluster labeling based on both linguistic and statistical perspectives is introduced. A technique for extracting informative and expressive labels that covers the main topics of a cluster of documents is proposed. The technique aims to capture labels that include the important shared common concepts along the cluster, along with the important concepts that are addressed by individual documents. In addition to statistical model, linguistic knowledge is used during the steps of the labeling process to guarantee informative and representative final extracted cluster labels.

The proposed cluster labeling is divided into two main phases, local document keyphrase extraction, and cluster topics construction. First, documents are preprocessed to extract word features and generate the lemma form and for each word in a document. Then, Indicative keyphrases of each document at lemma level are extracted based on the statistical and linguistic knowledge. Extracted local keyphrases scores and statistical processes are used in the second phase to construct cluster topics and their scores. The score of a topic carries both the shared and local topic importance. To generate output labels, two different schemes are adopted to achieve one or more goals of cluster labeling. The goals are:

- Extract the most informative labels that capture main topics.
- Eliminate the domination of main topic on output.
- Keep labels redundancy to a minimum. This is an essential requirement to allow a room for other concepts to be presented in the output.
- Cover all important topics of the document.

The work presented here is not concerned with how the clusters are generated; it extracts keyphrases from already clustered documents. The technique is applied to automatically extract labels for clusters of Arabic documents. However, the steps of the process are independent of neither the domain of the documents, nor the language used.

The rest of the paper is organized as follows: Section 2 introduces existing related work briefly. Section 3 presents the details of the proposed technique. Experimental results and analysis are reported in Section 4 and the last section concludes this paper.

2 Related work

Clustering algorithms have been introduced to automatically group similar documents into subsets (clusters), the obtained clusters need to be analyzed to help understand what clusters are about. However, while clustering techniques represent an important tool to categorize documents, they have limits to produce such labels. A list of words or short phrases is assigned to the cluster to describe their contents. In such settings, cluster labeling techniques come into the scene. Several cluster labeling algorithms have been proposed for this purpose.

Several proposals on cluster labeling use a naive method to identify the cluster by selecting the most salient terms that characterize the cluster. Salient terms are extracted by using statistical feature selection, e.g., the most frequent terms in this [7, 8, 9]. More advanced approaches select as labels the most weighted terms in the cluster's centroid. One example of algorithms that use this representation is Scatter/Gather centroid [2], which represents a cluster with a list of documents near the cluster's centroid and a list of topical terms. The topical terms are the terms with the highest weights in the cluster centroid.

There have been another works to identify cluster labels from word distribution in the hierarchy. Popescul and Glover [10], [11] proposed statistical methods in selecting cluster descriptors, based on the context of the surrounding clusters (parent cluster and sibling clusters). Popescul proposed to use the statistical test $\chi 2$ to detect difference in word distribution across the hierarchy.

Keyphrase extraction, which is a text mining task, extracts highly relevant phrases from documents. Literature lists over a dozen applications that utilize key phrase extraction. For example, providing mini-summaries of large documents, highlighting keyphrases in text, text compression, indexing, document clustering, and document classification are few use cases. Keyphrase Extraction from single document is often treated as supervised learning task while keyphrase extraction from a set of documents is often treated as unsupervised learning task. Unsupervised task tries to discover the topics rather than learn from examples. Several researchers adopted keyphrase extraction technique in cluster labeling. They aim at extracting important phrases from sentences or documents [12, 13]. The titles of articles are used in Li's research to improve the quality of keyphrases (words) based on an assumption that the words occurring in the titles should have higher significance. Hammouda et al. [14] introduced an algorithm called "CorePhrase " for topic discovery using keyphrase extraction from multi-document sets and clusters based on frequent and significant shared phrases between documents. CorePhrase works by extracting a list of candidate keyphrases by intersecting documents using a graph-based model of the phrases in the documents. In the work of Li et al. [5] cluster labeling is divided into two steps and a hybrid approach which tries to produce labels from both linguistic and statistical perspectives. They used Linguistic knowledge to make sure the extracted phrases are readable and informative. A context sensitive scoring method is proposed to model the influence of words over the ranking of candidate labels.

Other works use external resources in their cluster labeling techniques. The work of Lalitha et al. [15] considered embedding external knowledge to terms using WordNet. They provided an approach to derive a theme in the group of documents and label that group with the most appropriate phrase. The work of Qureshi at el. [16] used external resources to ensure the readability of selected labels. They used the titles or categories in Wikipedia pages as cluster label candidates. Labels produced in this way are readable, but they are usually high-level concepts and cannot describe small size clusters precisely [17]. For example, there is a page ''Armenian Genocide" in Wikipedia, but there is no page for ''the 78th anniversary of the Armenian Genocide" which is a small topic in 20-NewsGroup dataset. Therefore, sometimes they may even hurt the labeling accuracy due to their irrelevance to the documents' topics [18].

3 Generating cluster labels

The cluster labeling algorithm is divided into two main phases, local document keyphrase extraction, and cluster topics construction. First, each document in the cluster is applied to Arabic lemmatizer module to extract the lemma form and word features for each word. Then, Indicative keyphrases of each document at lemma level are extracted based on statistical model and linguistic knowledge. The extracted local keyphrases scores and statistical processes are used in the second phase to construct cluster topics and their scores. The cluster topic score carries both the shared and local topic importance. Finally, the top n cluster topics scores are picked as cluster labels from the list of candidate topics. In this regard, two different schemes are proposed to determine important cluster topics. The following sections illustrate the details of the technique.

3.1 Local Document Keyphrase Extractor

First, each document in the cluster is applied to our previous Arabic lemmatizer module to extract the lemma form and word features for each word. The lemmatizer splits the document into sentences and words, removes punctuations and strange characters, extracts the necessary Part Of Speech (POS) features for each word, and generates the corresponding lemma form. The lemmatizer algorithm is based on Morphological and syntactic rules in addition to limited sized auxiliary dictionaries to generate lemma form and word category. For more details about the Arabic lemmatizer refer to [19].

Then, based on the lemmatizer output, each document of the cluster is passed to the keyphrase extractor LBAKE module to extract indicative keyphrases of each document at lemma level [20]. LBAKE is a supervised learning system for extracting keyphrases of single Arabic document. Linguistic processing step extracts Part of Speech (POS) Tagging to tag each single word from the document with its part-of-speech, and also extracts the lemma form of each word. The information results of the linguistic processing are used to extract keyphrases. The extractor is supplied with linguistic knowledge as well as statistical information to enhance its efficiency. All possible phrases of one, two, or three consecu-

tive words that appear in a given document are generated as n-gram terms. These n-gram words are accepted as a candidate keyphrase if they follow syntactic rules. To hide inflectional variations, words are represented in their lemma forms in all computation processes. The importance of a keyphrase (score) within a free-text document is based on seven features:

Number of words in each phrase.

Frequency of the candidate phrase.

Frequency of the most frequent single word in a candidate phrase.

Location of the phrase sentence within the document.

Location of the candidate phrase within its sentence.

Relative phrase length to its containing sentence.

Assessment of the phrase sentence verb content.

Weights of these features were learned during building the classifier. The output of LBAKE is a set of scored keyphrases normalized to their maximum, representing the input document. Each document is replaced by the features illustrated in Table (1).

TABLE 1
FEATURES REPRESENTING A DOCUMENT

Feature	Description
d_i	Document number [1≤ i ≤D]
$S_{i,j}$	Set of sentences in document j
Ld_i	Length of the document i expressed as the number of sentences in the document.
NP_i	Total number of extracted local keyphrases for a document i
$P_{i,j}$	Set of Keyphrases in a document. P is represented in lemma form, 1≤ i ≤D, and 1≤ j ≤ NP_i
$LS_{i,j}$	Set of Normalized local Keyphrases score in a document. Their ranges: 0≤LS≤1, 1≤ i ≤D, and 1≤ j ≤ NP_i
	$LS_{i,j}$ is the local score divided by a maximum local score of a given document.

3.2 Cluster Topics Construction

The next step of the algorithm was to construct the cluster topics T_k, and their Cluster scores TS_k for all documents. Extracted local keyphrases have rich information, and can be used in various scoring schemes for cluster topic construction. To realize this process, all local kephrases features of the set of documents were combined together, and each keyphrase was assigned a new global cluster topic score based on its importance on the local document as well as the relevance to all documents of the cluster. The next subsections illustrate steps to construct cluster topics and their scores.

3.2.1 Maximum Coverage Score

A direct solution to construct the cluster topics (T) is to union all local keyphrases.

$T = \cup P_{i,j}$ $1 \le i \le D$, *and* $1 \le j \le NP_i$

Since a keyphrase T may appear in many documents, we set the maximum coverage score MCS equal to the maximum local keyphrase score that match T.

$MCS_k = max(LS_{i,j})$, *and* $T_k = P_{i,j}$

Top ranked non-duplicated keyphrases were then selected, which guaranteed the inclusion of all important local keyphrases in the global labels. All important topics in local documents will be included in the cluster labels with this technique; hence it tends to maximize the coverage of the labels.

3.2.2 Centroid Topic Score

In spite of its simplicity, the previous scoring ignores the relevance aspect of selected keyphrases. In cluster labeling, importance should be given to common information that maintained by many documents. For example if there are two different keyphrases with the same local scores in two different documents in a cluster, and only one of these keyphrases could be repeated multiple times in other documents. To provide a fair assessment of the keyphrase importance, repetitions of the keyphrase in other documents must be considered. This is represented by the relevance feature which reflects the importance of a keyphrase for the set of documents. The relevance of a local keyphrase (P) can be found by its frequency (F_P) among the cluster documents. The concept is that the importance of a keyphrase increases as it appears in more documents. The frequency F of a keyphrase P is given by:

$$F_P = count(P \cap P_{i,j}), \quad \textit{for all i,j}$$

Note that F also represents the number of documents that contain P. The use of frequency as a sole representation of the importance of a keyphrase is not always an accurate representation of importance. For example a minor topic that is repeated in many documents will gain a false importance. Therefore, we considered a 'Centroid Topic Score' as a solution to overcome this. Centroid topic is defined as the topic that is important in its local document and relevant to document cluster. Therefore, Centroid Topic Score CTS is given by multiplying the two factors:

$$CTS_k = NF_k \, MCS_k \qquad (1)$$

where $NF_k = F_k / max(F)$

Where NF_k : is the normalized frequency of T_k among T.

3.2.3 Centroid Document Score

An important feature of the proposed technique, is the ability to reject (or at least reduce) the effect of non-related documents. For example, if there is a cluster containing nine documents concerned with 'Tsunami', and the tenth article is strongly related to 'terrorist incident'. Since the tenth article is strongly related to 'terrorist incident', its keyphrases still have top scores. This will mislead the extractor to include unim-

portant topics. In our approach, we exploit a 'Centroid Document Score CDS' to evaluate the relevance of the document to the cluster. Keyphrases extracted from centroid documents get a bonus by CDS values. CDS is ranked by the number of links of a document to other documents. A Link Score between two documents A and B is the count of their matched keyphrases.

CDS of a document k is calculated as the summation of link scores between document k and all other documents divided by the number of keyphrases of k. Since the keyphrases are guaranteed not to be repeated within a local document, CDS is set to:

$$CDS_K = \left(\frac{\sum_{i=1}^{NP_K}(F_i)}{NP_K}\right) \quad (2)$$

In multiple documents, the extracted keyphrase must be important in its local document in addition to having strong relevance to the main concepts of the cluster. Finally, to have a balance between maximum coverage and relevance, we included the centroid document score to equation (1) to represent the Maximum Centroid Topic Score.

$$MaxCR\ TS_K = CDS_K\ CTS_K \quad (3)$$

3.4 Cluster Labels Extraction

Once extracted cluster topic scores have been computed, the cluster labels become ready for extraction. In this regard, two different schemes are adopted to achieve one or more goals of cluster labeling. The first heuristic prefers top scores to generate important labels. While the second favors phrases well covering top topics. The following subsections describe the two heuristics for extracting cluster labels based on cluster topics. For each cluster of documents, the top ten extracted cluster topics are employed through the evaluation experiments.

3.4.1 Uppermost Heuristic

Once extracted cluster topic scores have been computed, labels are then ranked based on their Normalized Maximum Centroid Topic Score, and an (n) percentage of uppermost topics are extracted into the output. The algorithm is greedy since it favors a topic that contains important concepts. The algorithm succeeds to capture labels that carry most important topics of the cluster.

3.4.2 Comprehensive Heuristic

In uppermost heuristic, many labels that describe same (focus) topic dominate the selection. All main topic individual grams constituent will get higher scores. For example a cluster of documents that deals with (انترنت الأشياء -*internet of things*), extracted labels could include 3 individual entity grams of the main topic, i.e. ("انترنت ", " الأشياء", "انترنت الأشياء " – "*internet* ", "*things* " , "*internet of things* "). For Multi-topic documents, this limits the opportunity for other concepts to be presented in the output and therefore low coverage of elementary cluster concepts could occur. Moreover, for documents that are rich in terminologies, this could results in generating uninformative labels.

To achieve one or more goals of cluster labeling, the comprehensive heuristic neglects the unigram topics that are already included in other bi/tri-gram top scores topics. The intent is to provide a room for other topics to appear in the output list, and to produce informative and cohesive readable labels as well. The coverage heuristic is given by

Find % N top scores cluster topic T_k *, where N> n number of required cluster labels*
For each lemma form of cluster topic T_{ki} *with n-gram=1*
Find the first cluster topic T_{kj} *that includes* T_{ki} *, add* T_{kj} *to* nT_k
else, add T_{ki} *to* nT_k

4 Performance Evaluation

The proposed technique is applied to automatically extract labels for clusters of Arabic documents. To assess the performance, precision, recall, and F-measure are calculated and compared against TF method. To assess the performance of the proposed technique, two experiments were applied. The first measures the validity of the proposed technique and make a comparison against TF method, while the second tests other characteristics of the extracted cluster labels.

4.1 Dataset

One of the major limitations faced by Arabic research is the lack of adequate resources and gold standards that could help in evaluating the performance of different systems. To evaluate features of the proposed technique, two datasets are adopted. In the first (DatSet1), we have collected 200 news web articles that deal with different 20 events. Each group/cluster includes 10 articles, and has on the average 180 words per document. Each cluster of related articles deals with news about a single main event. Each article in a certain cluster includes a limited number of topics (mostly one or two topics) about the event. Some topics are dealt with by many documents, while others are addressed in a single document. Human expert is asked to label the clusters manually. The human labels are used as a gold standard to assess the automatically extracted labels.

However dataset1 can test the basic features of the proposed algorithm. It is not enough to test multi-topic feature and informative richness of the produced labels. This is because all documents are concerned with a single event that carries a limited number of topics. Therefore, we adopted a second dataset (DatSet2) that contains five collections of related web articles in social, tutoring, science, geology, and geophysics domains. Each collection contains a cluster of 10 related articles with multiple numbers of topics for each domain subject. Each cluster of related articles deals with a particular general main topic. Each article in a certain cluster includes multiple topics about the main topic. The average number of words per article is (340) words. The labeling task was to produce 10 labels.

4.2 Evaluation and Results

4.2.2. Experiment1 Validity and Comparison

The first experiment measures the validity of the proposed technique and comparison against other method. The proposed technique is applied to automatically extract labels for clusters of Arabic documents (Dataset1) mentioned above. The top scores N automatically extracted labels are compared to the human labels. Precision, recall, and F-measure are evaluated.

For comparison of the proposed method, we used a baseline term frequency TF method over the same dataset. In the calculation of TF method, after removing stop words, term frequency for each word in the cluster is calculated, and the top scores N words are considered as labels for a cluster.

Table 2 illustrates the results of precision, recall, and F-measure for both of the proposed technique and TF method. The experimental results demonstrated satisfactory results of the proposed technique. F-measure=0.52 compared to 0.26 in TF method.

The results prove that the proposed technique is a more efficient technique than using TF as a sole representation of the term importance. The proposed technique succeeds to capture labels that carry most important topics of the cluster.

TABLE 2
RESULTS AGAINST TF METHOD

	Precision	Recall	F-measure
Proposed technique	0.41	0.72	0.52
TF method	0.21	0.36	0.26

4.2.3 Experiment2 Multi-topic Features

In the second experiment, the proposed technique is applied to automatically extract labels for the clusters of multi-topic Arabic documents (dataset2) described above. Human evaluator was asked to assess the features of generated labels. Cluster labels were tested on the bases of measuring their coverage, and informative richness. The maximum score was 10 per measure with a total 20 for a cluster. Figure 1 shows the average results of the human evaluation for the TF method and the proposed comprehensive heuristic. The proposed technique is a more efficient for its high coverage and informative richness. In the comprehensive heuristic, only one label at most, is extracted for each topic. The unigram topics that are already included in other high score topics are excluded from the output list. It extracts the unigram topic, if it is unique and does not include in other topics. Thus, it gives the opportunity for other important topics to be presented in the output label list. The algorithm intent is to comprehend all the major topics of the document, and at the same time keeping redundancy to a minimum. Furthermore, the algorithm favors bi/tri gram topics over the unigram ones; this in turn contributes extracting terminologies and informative readable topics. Table 3 shows samples of the output results for both TF and the proposed comprehensive method. To improve the results and reduce redundant concepts in the TF method, the lemma form is considered during the calculations.

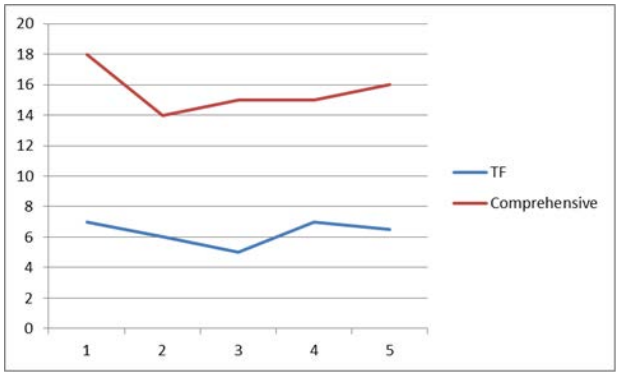

Figure 1 Average results of the human evaluation for the tf method and the proposed comprehensive heuristic

TABLE 3
SAMPLES OF THE OUTPUT LABELS FOR THE TF AND THE PROPOSED COMPREHENSIVE HEURISTIC

TF	Comprehensive
الماء،السطح، المستويات،خزان،الآبار،عملية،ا لأرض،منطقة،المائية ،الخزانات	المياه الجوفية،الخزانات ، ،مستوي الماء ،الموارد المائية ،المياه السطحية ،الضغط الجوي ،نتائج الامراض ،أرتفاعاً فى تكاليف ،المياه الأرضية، المنطقة المشبعة ، التغذية المائية
البيانات.، قاعدة، معلومات، جدول ، السجل، الموظفين،الجدول، عمليات، بالإنجليزية،طريقة	قواعد البيانات ، المعلومات الخاصة ، جدول البيانات ، أنواع قواعد البيانات ، البيانات التجارية (بالإنجليزية ،) مدير قاعدة البيانات ، تكامل البيانات ، تصميم قاعدة البيانات ، Access المعلومات، طبيعة تشكيل البيانات

5 CONCLUSION

In this paper we proposed a technique for extracting labels for a cluster of documents. Keyphrases that are based on both linguistic knowledge and statistical model are adopted to represent cluster labels. To obtain maximum coverage and relevance, labels are selected based on the topic importance in its local document in addition to the relevance to the main concepts of the cluster. Two different heuristics were adopted to extract cluster labels. The first prefers top scores to generate important labels. While the second favors phrases well covers top topics. To assess the performance of the proposed technique, two experiments were applied. The proposed technique was compared against TF method to automatically extract labels for clusters of Arabic documents. The results proved that the proposed technique is more efficient. In the proposed

technique, the main concepts are represented by keyphrases that maximize both the relevance and coverage to the cluster. We explored balances between concept importance and coverage of all topics.

REFERENCES

[1] Manning, C.D., Raghavan, P. and Schütze, H., 2008. Text classification and naive bayes. Introduction to information retrieval, 1, p.6.

[2] Cutting, D.R., Karger, D.R., Pedersen, J.O. and Tukey, J.W., 2017, August. Scatter/gather: A cluster-based approach to browsing large document collections. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 148-159). ACM.

[3] Pourvali, M., 2017. Improving the quality of text clustering and cluster labeling.

[4] Nguyen, C.T., Phan, X.H., Horiguchi, S., Nguyen, T.T. and Ha, Q.T., 2009. Web search clustering and labeling with hidden topics. ACM Transactions on Asian Language Information Processing (TALIP), 8(3), p.12.

[5] Li, Z., Li, J., Liao, Y., Wen, S. and Tang, J., 2015. Labeling clusters from both linguistic and statistical perspectives: A hybrid approach. Knowledge-Based Systems, 76, pp.219-227.

[6] El-Shishtawy, T. and Al-Sammak, A., 2012. Arabic keyphrase extraction using linguistic knowledge and machine learning techniques. arXiv preprint arXiv:1203.4605.

[7] Beil, F., Ester, M. and Xu, X., 2002, July. Frequent term-based text clustering. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 436-442). ACM.

[8] Chuang S., and Chien L., 2004. A practical web-based approach to generating topic hierarchy for text segments. In Proceedings of the 20th International Conference on Information and Knowledge Management.

[9] Fung, B.C., Wang, K. and Ester, M., 2003, May. Hierarchical document clustering using frequent itemsets. In Proceedings of the 2003 SIAM International Conference on Data. Mining (pp. 59-70). Society for Industrial and Applied Mathematics.

[10] Popescul, A. and Ungar, L.H., 2000. Automatic labeling of document clusters. Unpublished manuscript, available at http://citeseer. nj. nec. com/popescul00automatic. html.

[11] Glover, E., Pennock, D.M., Lawrence, S. and Krovetz, R., 2002, November. Inferring hierarchical descriptions. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 507-514). ACM.

[12] Li, D., Li, S., Li, W., Wang, W. and Qu, W., 2010, July. A semi-supervised key phrase extraction approach: learning from title phrases through a document semantic network. In Proceedings of the ACL 2010 conference short papers (pp. 296-300). Association for Computational Linguistics.

[13] Li, D. and Li, S., 2011, March. Hypergraph-based inductive learning for generating implicit key phrases. In Proceedings of the 20th international conference companion on World wide web (pp. 77-78). ACM.

[14] Hammouda, K.M., Matute, D.N. and Kamel, M.S., 2005, July. Corephrase: Keyphrase extraction for document clustering. In MLDM (Vol. 2005, pp. 265-274).

[15] Lalitha, Y.S., Raju, N.G. and Rao, O.S., 2017. Labeling Document Clusters with Thematic Phrases. International Advanced Research Journal in Science, Engineering and Technology IARJSET,Vol.4.

[16] Qureshi, M. Atif, Colm O'Riordan, and Gabriella Pasi. "Short-text domain specific key terms/phrases extraction using an n-gram model with wikipedia." In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 2515-2518. ACM, 2012

[17] Niu, N., Reddivari, S., Mahmoud, A., Bhowmik, T. and Xu, S., 2012, June. Automatic labeling of software requirements clusters. In Search-Driven Development-Users, Infrastructure, Tools and Evaluation (SUITE), 2012 ICSE Workshop on (pp. 17-20). IEEE.

[18] Carmel, D., Roitman, H. and Zwerdling, N., 2009, July. Enhancing cluster labeling using wikipedia. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (pp. 139-146). ACM.

[19] El-Shishtawy, T. & El-Ghannam, F. (2012), An Accurate Arabic Root-Based Lemmatizer for Information Retrieval Purposes, International Journal of Computer Science Issues, Volume 9, Issue 1, pp. 58-66.

[20] El-Shishtawy, T. & El-Ghannam, F. (2012), Keyphrase Based Arabic Summarizer (KPAS), 8th International Conference on Informatics and Systems (INFOS).