Keyword Detection Techniques

A Comprehensive Study

Zaffar Ahmed Shaikh
Faculty of Computer Science & Information Technology
Benazir Bhutto Shaheed University
Lyari, Karachi, Pakistan
zashaikh@bbsul.edu.pk

Abstract—**Automatic identification of influential segments from a large amount of data is an important part of topic detection and tracking (TDT). This can be done using keyword identification via collocation techniques, word co-occurrence networks, topic modeling and other machine learning techniques. This paper reviews existing traditional keyword extraction techniques and analyzes them to make useful insights and to give future directions for better automatic, unsupervised and language independent research. The paper reviews extant literature on existing traditional TDT approaches for automatic identification of influential segments from a large amount of data in keyword detection task. The current keyword detection techniques used by researchers have been discussed. Inferences have been drawn from current keyword detection techniques used by researchers, their advantages and disadvantages over the previous studies and the analysis results have been provided in tabular form. Although keyword detection has been widely explored, there is still a large scope and need for identifying topics from the uncertain user-generated data.**

Keywords-keyword detection; information retrieval; topic detection; machine learning; comprehensive study

I. INTRODUCTION

Keyword extraction using manual methods is slow, expensive and bristling with mistakes [1]. In recent years, many automatic bursty keyword extraction techniques have been proposed to extract keywords from large amounts of data. These keywords are helpful in identifying themes and influential segments and framing semantic web and other applications of natural language processing [2, 3]. Automatic keyword detection research area is related to topic detection and tracking (TDT) domain which was proposed in [4]. Various applications use keyword extraction techniques for web search, report generation and cataloguing [5]. This area is intended to identify the most useful terms which include many sub-processes. Documents are introduced in MS Word, html, or pdf formats. Initially, the documents are pre-processed to remove redundant and unimportant information [6, 7]. The data is then processed through different keyword extraction approaches including statistical approach, linguistic approach, machine learning approach, network based approach and topic modelling approach [8, 9].

In statistical approach, term frequency–inverse document frequency (Tf-Idf) is the most widely used technique for keyword extraction. The researchers use Tf-Idf to give a document a score based upon some query. A change in score occurs when a query is changed or updated. Without a query, there is no score [10]. Recently many new techniques have been developed for statistical keyword extraction [11]. These include PageRank, LexRank, etc. In PageRank, the researchers assign a score to a document based upon the documents it links to, and the documents which link to it. It is a global ranking scheme [10]. Therefore, in PageRank, the score does not change (like in Tf-Idf) depending on the query used. As observed, PageRank and LexRank algorithms perform better than Tf-Idf. In linguistic approach, automatically identifying keywords is similar to semantic resemblance [12]. In machine learning approach, the keyword extraction technique is considered as classification technique [13].

Different dictionaries including WordNet, SentiNet and ConceptNet are used for keyword extraction techniques. In network based algorithms, the nature and semantics of word co-occurrence networks is studied to identify important terms. In this, nodes are considered as words and edges are considered as co-occurrence frequency [14]. Many useful insights have been obtained from these algorithms for identifying influential segments and keywords. Topic modelling techniques have been popularized in [15]. Authors introduced Latent Dirichlet Allocation technique which is used to identify which document is related to which topic and to what extent [16]. This has been further improved by Hierarchical Dirichlet Process, Pachinko Allocation Model, Relational Topic Modeling, Conditional Topic Random Fields and recently by Hierarchical Pitman–Yor–Dirichlet Language Model and Graph Topic Model [17]. Although keyword extraction is an important area of research and many researchers and practitioners gave a lot of attention to it, state of the art keyword extraction method is still not observed as compared to many other core natural language processing tasks [18]. This paper reviews existing traditional keyword extraction techniques and analyzes them to make useful insights to give future directions for better automatic, unsupervised and language independent research.

II. Related work

Authors in [19] developed a so-called tool 'Keyword Extractor' for automatic extraction of most likely terms that closely match experts' preferences. Their study was related to brain research which involved worldwide collaborations and exchange of information among neuroinformatics centers and portal sites. The main objective of their study was the efficient use of resources and the improvement in the quality of brain research. Each center and site developed their own set of keywords for classification of the main text and the resources. The researchers tested their tool over the abstract database of two science journals. Authors in [20] extracted keywords from a Chinese microblog. To extract keywords, they performed five steps and used three features (i.e., graph model, semantic space, and location of words). In the first step, researchers downloaded microblog API of a user. Secondly, they preprocessed the data by applying data cleaning, word segment, POS tagging, and stop word removal techniques. To extract keywords, researchers in the third step created a graph model that was based on the co-occurrence between words. They assigned sequence numbers to the words according to their location and developed weight of the words by using the score formula. In the fourth step, researchers first created a semantic space that was based on topic extraction and then computed statistical weight of the words by using Tf-Idf. In the fifth and last step, researchers first identified location of words, and then, based on the location of those words, computed the rank value of each word. Authors in [21] focused on the structure approach and graph generation. The approach used in this paper is structure based in which researchers created graph model and identified bursty topics and events. In topic clustering, twitter tweets were separated to produce homogeneous graphs and heterogeneous graphs. For homogeneous graphs, researchers used OSLOM algorithm to find interaction among users. For heterogeneous graphs, rankclus algorithm was used to construct a set of tweets ranked with number. Finally, from both graph results, the concept, theme or event of a tweet was measured by joining tweets with the same name. Researchers planned ahead to develop graph models to be used for different types of events and to construct a method that can define events.

Authors in [22] developed a keyword extraction technique for tweets with high variance and lexical variant problems. Lexical variants are examples of free variation in language. They are characterized by similarity in phonetical or spelling form and identity of both meaning and distribution. The authors used brown clustering and continuous word vector methods. In brown clustering method, they clustered words having same meaning (such as no, noo, etc.) and then found out the features for the individual cluster. In continuous word vector method, the authors defined a layer by finding its probability and then the word is changed into continuous word vector. Next, they predicted the length of the keyword by calculating the ratio between the number of keywords and the total number of words in the tweets. In the end, linear regression method was used to predict the number of keywords. Authors in [23] developed a system to detect popular keyword trend and bursty keywords. Their system detects keyword abbreviations and any typing and spacing errors. The first step they took is to collect the candidate keywords (i.e., the first word starting with the capital letter or the word enclosed in quotation mark is considered as candidate keyword). The second step was to merge keywords. To do so, they considered acronyms and typo and spacing errors, and then, found out the Tf accordingly. Finally, they detected popular keywords from the candidate keywords which were merged, and then, selected bursty keywords using the burst ratio technique. Authors in [24] gave the idea of TOPOL (a topic detection method based on topology data analysis) which identifies the irrelevant noisy data from the useful data. The first step of the authors was the preprocessing step in which the elimination of the hashtags, the URLs, and the non-textual symbols from a tweet was done. Their second step was mapping, in which a matrix was generated by applying the SVD technique. In the third step, which the authors called the topic extraction step, the topics were selected based on the interest. Finally, the results were computed based on topic recall, keyword precision, and keyword recall parameters. Authors in [25] presented and discussed different methods and approaches used in the keyword extraction task. They also proposed a graph based keyword extraction method for the Croatian language which is based on extraction of nodes. The authors used selectivity-based keyword extraction method in which text is represented in the form of vertex and edges. The result is computed on the in-degree, out-degree, closeness and selectivity. Authors in [26] developed a keyword extraction method that represents text with a graph, applies the centrality measure and finds the relevant vertices. Authors proposed a three-step based technique called TKG (Twitter Keyword Graph). The first step was the pre-processing in which stop words were removed. In the second step, a graph was developed in which nearest neighbor and all neighbors were considered. Finally, the results were computed based on the precision, recall, F-measure test scores and graph scalability.

Authors in [27] proposed an information summarization method for the large quantum of information which is disseminated everyday through tweets. Their method collects tweets using a specific keyword and then, summarizes them to find out the topics. The authors provide two algorithms: Topic extraction using AGF (TDA) and topic clustering and tweet retrieval (TCTR). The methodology first extracts tweets from twitter and then applies the Tf-Idf technique to find out weights and word frequency. The AGF is evaluated using keyword rating. Finally, the results are calculated based on the class entropy, purity, and cluster entropy. Authors in [28] proposed a technique in which a user can search using a search engine but without entering any keywords. The google similarity distance technique is used to find the keywords. A log is maintained in which user behavior and repository is saved. So, the need for the repository is abolished and everything is done online and in real time. Keyword expansion and extraction methods are used to extract relevant and accurate information. In keyword expansion, help is provided to user to enter the exact keyword and to get the exact information. In keyword extraction, the word is analyzed based on the occurrence on the length and frequency. Keyword extraction method relies on statistical approaches and machine learning approaches. The proposed methodology of the authors is composed of three parts: 1-g filtering, google similarity distance calculation, and search results filtering. Finally, the results are calculated based on the parameters of precision and recall. The relationship between top k results is evaluated. Thus, the authors proposed a system in which user just needs to browse the web page and the relevant keywords are generated. The system suits well for the science stream as the words are clear but may not be accurate for the social science.

Authors in [29] produced a facility based on Bayesian text classification approach called high relevance keyword extraction (HRKE) to extract the keywords at the stage of classification without the use of pre-classification process. The facility uses a posterior probability value to extract keywords. The HRKE first extracts the words from the text. Next, the posterior probability is calculated. Finally, the Tf-Idf method is used to assign weights to words. Authors claim that the HRKE facility improves the performance and accuracy of the Bayesian classifier and reduces time consumption. The experiment was conducted on three dataset-featured article datasets. In the end, the corresponding threshold and accuracy graph is plotted. Authors in [30] address the problem of part-of-speech (POS) tagging from the richer text of twitter. Authors developed a POS tagset first. Secondly, they performed manual tagging on the dataset. Afterwards, the features for the POS tagger were developed. Finally, the experiments were conducted to develop the annotated dataset for the research community. The hashtags, URLs, and emotions were considered. The results were obtained with 90 percent accuracy. Authors concluded claiming that the approach can be applied to linguistic analysis of social media, and the annotated data can be used in semi supervised learning. Authors in [31] gave a solution to the problem of statistical keyword extraction from the text by adapting entropic and clustering approaches. Authors made changes in these approaches and proposed a new technique which detects keywords as per user’s needs. The main objective of the authors was to find and rank important words in the text. The two approaches were applied on short texts (such as web pages, articles, glossary terms, generic short text etc.) and long texts (such as books, periodicals etc.). Results were evaluated and the clustering approach proved to be better for both cases, while the entropic approach suited well for the long text and did not perform well for the partitioned text.

Authors in [32] proposed a metric called entropy difference (ED) for the ranking of the words on a Chinese dataset. Authors used Shannon’s entropy method which is the difference between intrinsic and extrinsic modes. The idea of intrinsic and extrinsic modes is that meaningful words are grouped together. Therefore, the words are extracted and ranked according to the entropy difference. Authors calculated mean, mode and median on entropy differences. Their ED metric proved to be a good choice in word ranking. The method differentiates between the words that define authors’ purpose and the irrelevant words which are present randomly in the text. This method is well suited for single document of which no information is known in advance. Authors in [33] provided a solution to the inherent noisy and short nature tweet problem of Twitter streams called HybridSeg. Authors incorporated local context knowledge of the tweets with global knowledge bases for better tweet segmentation. The tweet segmentation process was performed on two tweet datasets. The tweets were split into segments to extract meaning of the information conveyed through the tweet. Results show that HybridSeg significantly improved tweet segmentation quality compared with other traditional approaches. Authors claim that the segment based entity is better than word based entity.

Author in [34] provided a unique solution to the keyword extraction problem called ConceptExtractor. The ConceptExtractor do not decide on the relevance of a term during the extraction phase, instead, it only extracts generic concepts from texts and postpones the decision about relevant terms based on the needs of the downstream applications. Authors claim that unlike other statistical extractors, ConceptExtractor can identify single-word and multi-word expressions using the same methodology. Results were evaluated based on three languages. Precision and recall were used for the result evaluation. Authors also defined a metric to specificity both single and multi-word expressions usable in other languages. Authors in [35] considered various Chinese keyword extraction methods. In this paper, extended Tf approach has been defined which considers Chinese characteristics with Tf method. Authors also developed a classification model based on support vector machine (SVM) algorithm. Many improvement strategies were defined and four experiments were performed to evaluate the results. Results showed that SVM optimized the keywords. Precision and recall rate improved much better. Authors concluded that the improved Tf method is much better than the traditional Tf method in terms of accuracy and precision. Authors in [36] discovered and classified terms that are either document title or ‘title-like’. Their idea was that the terms that are title or title like should behave in the same way in a document. The classifier was trained using distributional and linguistic features to find the behavior of the terms. Different features were considered such as location, frequency, document size etc. The rating was calculated on the basis of topical, thematic and title terms. After this the evaluation was performed based on recall and precision. The recall rate of finding the title terms was high but the precision rate was low because some of the words which were not titles were also identified in title terms. Authors in [37] developed a sensitive text analysis for extracting task-oriented information from unstructured biological text sources using a combination of natural language, dynamic programming techniques and text classification methods. Using computable functions, the model finds out matching sequences, identifies effects of various factors and handles complex information sequences. Authors pre-processed the text contents and applied them with entity tagging component to find out the causes of diseases related to low-quality food. Results show that the bottom-up scanning of key-value pairs improves content finding which can be used to generate relevant sequences to the testing task. The method improves information retrieval accuracy in biological text analysis and reporting applications.

III. Analysis of Keyword Extraction Approaches

Table I provides inferences drawn from modern keyword detection techniques, their advantages and disadvantages over previous studies, and result analysis.

TABLE I. ANALYSIS OF EXISTING KEYWORD DETECTION TECHNIQUES

Paper	Techniques used	Advantages	Disadvantages	Results / Analysis
[20]	a) Graph model b) Semantic space	a) Can detect the words which are wrongly segmented. b) Extracts keywords from a micro blog.	a) Not suitable for large texts. b) Some terms will not be distinguished.	Best performance obtained is 0.6972
[21]	a) OSLOM algorithm b) Page rank algorithm	a) Able to identify the topics of twitter event. b) Less expensive.	Not able to identify the events based on graph clusters.	Best result obtained from structured based approach.
[22]	a) Brown clustering b) Continuous word vector	a) Improved state of the art for keyword extraction. b) Automatically keyword extraction.	Not suitable for Facebook text keyword extraction.	Accuracy for precision obtained is 72.05, recall 75.16.
[24]	TOPOL	a) Suitable for noisy data. b) Reduces computation time and improves topic extraction result.	Suffers from data fragmentation.	The result obtained is 0.5380 for recall,0.7500 for precision.
[26]	a) Tf-Idf b) KEA c) Proposed TKG	a) TKG proved to be robust and superior compared to other approaches b) TKG is simpler to use than KEA	The best configuration of TKG was not found	TKG results better compared to KEA and Tf-Idf.
[28]	a) Statistics approach b) Machine learning approach	a) Search engine which can automatically extract important keywords b) System works well	Not suitable for business management domain	High recall rate
[29]	Bayesian approach	a) Low cost, simple and efficient method. b) Handles raw data without text preprocessing.	a) Presence of noisy data may degrade the performance. b) Feature selection method degrades the efficiency of classification task.	Improved accuracy
[31]	a) Entropic b) Clustering approach	a) Suitable for both long and short texts. b) Reliable obtained results.	Median and mode did not give the correct result.	Good clustering results for both short and long texts.
[32]	Shannon entropy	a) Suitable for text with no information known in advance. b) Easy to numerically implement.	Median and mode did not give the correct result.	Better results for single document.
[33]	Hybrid segmentation	High quality tweet segmentation.	Manual segmentation is expensive.	Improved precision.
[34]	Statistical language independent	Good for extracting single and multi-word expressions.	Not suitable for long text	Improved precision and recall.
[36]	a) Decision tree classifier b) Pattern recognition	Easy title determination	a) Not easy to determine the best document size. b) Precision was not significant than recall.	Recall 85% was achieved for title like terms.
[37]	a) Sensitive text analysis b) Context-based extraction method	a) Category-oriented approach for extraction of task-specific information b) Investigations into recall and precision were carried out.	Not tested on generic data.	a) Food safety is analyzed to prevent future consequences. b) Improved classification accuracy by utilizing optimization constraints. c) Causes of diseases related to low-quality food were identified.

IV. CONCLUSIONS

This paper extends understanding of widely used `existing approaches to keyword detection in the identification of influential segments from a large amount of textual data or documents. Therefore, extant literature on existing traditional TDT approaches to automatic identification of important words was reviewed and discussed. Techniques reviewed include collocation, word co-occurrence networks, topic modelling and other machine learning approaches. Results show that the majority of these techniques is domain dependent and language dependent. It was observed that although traditional keyword extraction techniques have been performing satisfactorily, a need exists to propose unsupervised, domain independent and language independent techniques which use statistically computational methods. Keyword extraction task has been widely explored, but there is still a large scope and gap for identifying topics from the uncertain user-generated data.

REFERENCES

[1] E. Landhuis, "Neuroscience: Big brain, big data", Nature, Vol. 541, No. 7638, pp. 559-561, 2017

[2] G. Ercan, I. Cicekli, "Using lexical chains for keyword extraction", Information Processing & Management, Vol. 43, No. 6, pp. 1705-1714, 2007

[3] R. S. Ramya, K. R. Venugopal, S. S. Iyengar, L. M. Patnaik, "Feature extraction and duplicate detection for text mining: A survey", Global Journal of Computer Science and Technology, Vol. 16, No. 5, pp. 1-20, 2016

[4] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic detection and tracking pilot study final report, DARPA Broadcast News Transcription and Understanding Workshop, 1998

[5] P. Eckersley, G. F. Egan, S. Amari, F. Beltrame, R. Bennett, J. G. Bjaalie,T. Dalkara, E. De Schutter, C. Gonzalez, S. Grillner, A. Herz, K. P. Hoffmann, I. P. Jaaskelainen, S. H. Koslow, S.-Y. Lee, L. Matthiessen, P. L. Miller, F. M. da Silva, M. Novak,V. Ravindranath, R. Ritz, U. Ruotsalainen, S. Subramaniam, A. W.Toga, S. Usui, J. van Pelt, P. Verschure, D. Willshaw, A. Wrobel, Tang Yiyuan, "Neuroscience data and tool sharing", Neuroinformatics, Vol. 1, No. 2, pp. 149-165, 2003

[6] D. Kuttiyapillai, R. Rajeswari, "Insight into information extraction method using natural language processing technique", International Journal of Computer Science and Mobile Applications, Vol. 1, No. 5, pp. 97-109, 2013

[7] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, Text Mining: Applications and Theory, John Wiley & Sons, 2010

[8] J. Wu, S. R. Choudhury, A. Chiatti, C. Liang, C. L. Giles, "HESDK: A hybrid approach to extracting scientific domain knowledge entities", In ACM/IEEE Joint Conference on Digital Libraries, pp. 1-4, 2017

[9] D. B. Bracewell, F. Ren, S. Kuriowa, "Multilingual single document keyword extraction for information retrieval", IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 517-522, 2005

[10] D. Kuttiyapillai, R. Rajeswari, "Extended text feature classification with information extraction", International Journal of Applied Engineering Research, Vol. 10, No. 29, pp. 22671-22676, 2015

[11] S. C. Watkins, The young and the digital: What the migration to social-network sites, games, and anytime, anywhere media means for our future, Beacon Press, 2009

[12] I. M. Soboroff, D. P. McCullough, J. Lin, C. Macdonald, I. Ounis, R. McCreadie, "Evaluating real-time search over tweets", International Conference on Weblogs and Social Media, pp. 943-961, 2012

[13] H. L. Yang, A. F. Chao, "Sentiment analysis for Chinese reviews of movies in multi-genre based on morpheme-based features and collocations", Information Systems Frontiers, Vol. 17, No. 6, pp. 1335-1352, 2015

[14] J. Yang, J. Leskovec, "Patterns of temporal variation in online media", 4rth ACM international conference on Web search and data mining, pp. 177-186, 2011

[15] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research, Vol. 3, No. Jan, pp. 993-1022, 2003

[16] D. M. Blei, J. D. Lafferty, "Dynamic topic models", 23rd international conference on Machine learning, pp. 113-120, 2006

[17] M. Habibi, A. Popescu-Belis, "Keyword extraction and clustering for document recommendation in conversations", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 4, pp. 746-759, 2015

[18] S. Beliga, Keyword extraction: A review of methods and approaches, University of Rijeka, Department of Informatics, 2014

[19] S. Usui, P. Palmes, K. Nagata, T. Taniguchi, N. Ueda, "Keyword extraction, ranking, and organization for the neuroinformatics platform", Biosystems, Vol. 88, No. 3, pp. 334-342, 2007

[20] H. Zhao, Q. Zeng, "Micro-blog keyword extraction method based on graph model and semantic space", Journal of Multimedia, Vol. 8, No. 5, pp. 611-617, 2013

[21] H. Hromic, N. Prangnawarat, I. Hulpus, M. Karnstedt, C. Hayes, "Graph-based methods for clustering topics of interest in Twitter", International Conference on Web Engineering, pp. 701-704, Springer, 2015

[22] L. Marujo, W. Ling, I. Trancoso, C. Dyer, A. W. Black, A. Gershman, D. M. de Matos, J. P. Neto, J. G. Carbonell, "Automatic keyword extraction on Twitter", ACL (2), pp. 637-643, 2015

[23] D. Kim, D. Kim, S. Rho, E. Hwang, "Detecting trend and bursty keywords using characteristics of Twitter stream data", International Journal of Smart Home, Vol. 7, No. 1, pp. 209-220, 2013

[24] P. Torres-Tramon, H. Hromic, B. R. Heravi, "Topic detection in Twitter using topology data analysis", International Conference on Web Engineering, pp. 186-197, 2015

[25] S. Beliga, A. Mestrovic, S. Martincic-Ipsic, "An overview of graph-based keyword extraction methods and approaches", Journal of Information and Organizational Sciences, Vol. 39, No. 1, pp. 1-20, 2015

[26] W. D. Abilhoa, L. N. De Castro, "A keyword extraction method from Twitter messages represented as graphs", Applied Mathematics and Computation, Vol. 240, pp. 308-325, 2014

[27] A. Benny, M. Philip, "Keyword based tweet extraction and detection of related topics", Procedia Computer Science, Vol. 46, pp. 364-371, 2015

[28] W. Chung, H. Chen, J. F. Nunamaker Jr, "A visual framework for knowledge discovery on the web: An empirical study of business intelligence exploration", Journal of Management Information Systems, Vol. 21, No. 4, pp. 57-84, 2005

[29] D. Isa, L. H. Lee, V. P. Kallimani, R. Rajkumar, "Text document preprocessing with the bayes formula for classification using the support vector machine", IEEE Transactions on Knowledge and Data engineering, Vol. 20, No. 9, pp. 1264-1272, 2008

[30] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, N. A. Smith, "Part-of-speech tagging for Twitter: Annotation, features, and experiments", 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, Vol. 2, pp. 42-47, 2011

[31] P. Carpena, P. A. Bernaola-Galvan, C. Carretero-Campos, A. V. Coronado, "Probability distribution of intersymbol distances in random symbolic sequences: Applications to improving detection of keywords in texts and of amino acid clustering in proteins", Physical Review E, Vol. 94, No. 5, pp. 052302, 2016

[32] Z. Yang, K. Gao, K. Fan, Y. Lai, "Sensational headline identification by normalized cross entropy-based metric", The Computer Journal, Vol. 58, No. 4, pp. 644-655, 2014

[33] C. Li, A. Sun, J. Weng, Q. He, "Exploiting hybrid contexts for tweet segmentation", 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 523–532, 2013

[34] J. M. J. Ventura, Automatic extraction of concepts from texts and applications, Diss. Universidade Nova de Lisboa, 2014

[35] B. Hong, D. Zhen, "An extended keyword extraction method", Physics Procedia, Vol. 24B, pp. 1120-1127, 2012

[36] C. W. Wong, R. W. Luk, E. K. Ho, "Discovering 'title-like2 terms", Information Processing & Management, Vol. 41, No. 4, pp. 789–800, 2005

[37] D. Kuttiyapillai, R. Rajeswari, "A method for extracting task-oriented information from biological text sources", International Journal of Data Mining and Bioinformatics, Vol. 12, No. 4, pp. 387-399, 2015

AUTHOR PROFILE

Dr. Zaffar Ahmed Shaikh received his PhD in Computer Science from the Institute of Business Administration, Karachi (IBA-Karachi) in 2017. He is currently working as an Assistant Professor at Benazir Bhutto Shaheed University, Lyari, Karachi, Pakistan. He has twenty-three research publications to his credit and has received several research grants from EPFL (Switzerland), Higher Education Commission (Pakistan), Ministry of Higher Education (KSA) and IBA-Karachi. His research interests include Data Sciences, Knowledge Management, Language & Technology, Learning Environments, MOOCs, Social Software, Technology Enhanced Learning etc. Dr. Shaikh is a professional member of ACM and IEEE.