Transactions on Machine Intelligence 6(1) (2023) 28-36

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/index.php/TMI

A Language-Independent Method for Extracting the Essence of a Text in The Form of Phrases

J. Davoudi Moghaddam[1,*], A. Mosallanezhad[2], A. Ahmadi[3]

[1,2,3]K. N. Toosi University of Technology, Computer Engineering Faculty, Tehran, Iran

ARTICLE INFO

Article History:
Received 10 November 2022
Received in revised form
15 December 2022
Accepted 11 March 2023
Available online 12 March 2023

Keywords:
Essence of Text; Text Main Points; Text Processing; Language Independent; Essence Phrases.

ABSTRACT

With the growing integration of Information Technology (IT) into educational environments, individual learning behaviors and preferences have evolved significantly. This shift underscores the increasing importance of optimizing Social Learning Networks (SLNs) to better support personalized and adaptive learning experiences. A fundamental component of this optimization is the ability to accurately predict learners' future educational needs, thereby streamlining the learning process and enhancing learner outcomes. In response to this need, the present study introduces a predictive interpreter designed to anticipate user learning requirements within SLNs. This system operates by analyzing users' previously engaged topics and subsequently recommending appropriate follow-up subjects. Central to our approach is a user-oriented Collaborative Filtering (CF) method, tailored to model individual learning trajectories and identify patterns among learners with similar behaviors. To evaluate the performance and practical applicability of the proposed method, we conducted experiments using data derived from a well-established SLN. The empirical findings reveal that learners with similar interaction histories tend to exhibit comparable educational needs. The proposed CF-based prediction framework achieved a recall rate of approximately 60%, indicating a promising level of accuracy in anticipating learners' next topics of interest. This research contributes to the field of educational technology by offering a data-driven, adaptive solution for enhancing learner engagement and progression in SLNs. The results affirm the value of personalized recommendation systems in supporting effective and continuous learning.

1. INTRODUCTION

In recent advancements of natural language processing, several unsupervised keyphrase extraction methods have emerged, each offering unique approaches to distilling essential information from lengthy documents. KP-USE, a method introduced by Ajallouda et al. (2022), employs semantic similarity to extract key-phrases, showcasing its prowess in deciphering document content [1]. Another notable contender, PatternRank, as

* Corresponding author: j.davoudi@mail.kntu.ac.ir
K. N. Toosi University of Technology, Computer Engineering Faculty, Tehran, Iran

http://dx.doi.org/10.47176/TMI.2023.28

presented by Schopf et al. (2022), leverages pretrained language models and part-of-speech information, surpassing previous state-of-the-art methods in precision, recall, and F1-scores [2]. Meanwhile, Wang and Li (2022) propose BWRank, a keyphrase extraction model based on Bert, which outperforms its counterparts by adeptly capturing phrases that best encapsulate the textual content [3]. These advancements underscore the continuous refinement and innovation within the field of unsupervised keyphrase extraction.

Presently, the significance of text processing and its varied applications is widely acknowledged by researchers and students alike. Given the surging volume of textual and documental materials, the need for efficient methods to store and retrieve information has become imperative. Notably, search engines such as Google, Yahoo, and Bing grapple with the task of parsing numerous web documents to pinpoint the most relevant ones in response to user queries. In this scenario, the demand for real-time capability is of paramount importance.

Keyphrase extraction, alongside other domains like information extraction, natural language processing, text summarization, query understanding, machine translation, and text similarity, collectively falls within the expansive realm of text processing [4]. This paper defines "essence phrases" as keyphrases encapsulating the core points of a text. The extraction of essence phrases shares similarities with both keyword extraction and keyphrase extraction. These essence phrases prove valuable in information retrieval, text summarization [5,6], document clustering, and classification [7]. Noteworthy applications include methods utilizing keyphrase extraction for detecting news topics, with the goal of offering users a concise portrayal of news content [8,9].

Despite the substantial strides in text processing, there persist numerous open issues, particularly in comprehending document semantics. While these topics may seem straightforward to humans, they pose intricate challenges for computers due to the absence of a standardized format for preserving documents until machines can decipher their meaning and content. In the realm of text processing, document understanding and keyphrase extraction stand out as crucial objectives, prompting the development of a diverse array of statistical and linguistic approaches to address these intricate tasks. These methodologies exhibit variations, with some relying on multiple documents and others grappling with the complexities of single-document scenarios. The incorporation of learning algorithms, coupled with training data, characterizes certain techniques, while others operate without such dependencies. The efficacy of natural language processing tools and resources, including ontologies, proves impactful in augmenting results, albeit with varying reliability across languages. Furthermore, the demand for real-time output adds an additional layer of complexity for specific methods.

The literature abounds with various proposed approaches, such as co-occurrence-based keyphrase extraction [10,11], statistical methods exemplified by [12-14] and [15]. Kaur and Gupta devised a methodology centered on scoring and clustering document keywords, selecting the shortest noun phrase with the highest score as document keywords (13). Onda presented a keyword extraction method through clustering related keywords derived from query frequency history (14). Matsuo and Ishizuka introduced a keyword extraction method optimizing the x2 measure (15). Khoury et al. innovatively employed a part-of-speech hierarchy for extracting keywords from English sentences, involving rule learning based on a training corpus and hierarchical structure. The extraction process utilizes the most similar rules to each sentence [19]. Paukkeri et al. contributed an independent language method for keyphrase extraction, utilizing a suitable reference corpus for the test language [20]. Additionally, Aquino et al. proposed a language-independent approach for extracting n-grams from documents.

Following the conversion of n-grams into numeric vectors, a network is trained using these vectors, aiming to generate a model proficient in extracting keywords from newly generated documents [21]. In a parallel vein, Wu et al. introduced an SVM-based method tailored for identifying keywords within scientific documents [22]. The principal focus of this paper is the extraction of all document phrases, with an emphasis on eliminating non-critical ones. Section Two delineates the initial phase of the essence phrase extraction process, elucidating the intricacies of text preprocessing. Section three delves into the statistical analysis employed for scoring and filtering phrases, while Section four elucidates the methodology for extracting the essence of a phrase. Section five introduces the concept of fortifying phrases through recreation. Subsequently, Sections six and seven expound upon evaluation tests and draw conclusions, respectively.

2. PREPROCESSING

Preprocessing is used in text processing methods in order to process the text as raw data to omit meaningless or unimportant parts. Preprocessing contains several procedures. In this research we apply some of these procedures as described below.

2.1. Text Extraction

Texts are stored in several formats like HTML, PDF, etc. Sometimes the initial source of a text includes multimedia objects like web pages. The first step is to extract plain text from one of these initial formats. The plain text then will be the input of tokenization in the next steps.

2.2. Tokenization

The primary objective of this step is to tokenize a text into phrases, defining a phrase as a sequence of words arranged consecutively in a document. The order of words within a phrase is pivotal for conveying specific concepts. Phrases can encompass one or more words, meaning that a document containing K words comprises K single-word phrases, K-1 two-word phrases, and so forth. In this study, we operate under the assumption that there are no restrictions on the structure of the document text. Our method is structure-agnostic, treating a document as a sequence of words without discrimination between components such as title pages and the body.

A phrase, by definition, is composed of uninterrupted words. To identify relevant phrases within a document, the text can be segmented into sections separated by punctuation marks, excluding white space. Put differently, the word string of the document text is split by punctuation marks, and then phrases, each consisting of up to three words, are extracted from each sub-string. The maximum length of the extracted phrases is capped at three words. For example, if a sub-string comprises five words, such as "This is a scientific paper," there will be twelve phrases, each containing up to three words, as illustrated in Table 1. This step takes a document text as input and yields a list of phrases with lengths up to three words as output.

2.3. Stop word removal

Some words are considered "stop words" in all languages, such as pronouns, adverbs, conjunctions, prepositions, and modal verbs, among others. These words lack important meaning and can be removed from the text, despite their high frequency. Our experimental findings show that phrases containing up to three words, starting or ending with a stop word, lack clear meaning and should be omitted from extracted phrases. Starred phrases have been removed from extracted phrases list in this step. Phrases remained in this section are called “Active Phrases List” which are the input of next section.

Table 1. All possible phrases within the “This is a scientific paper” string.

* Starred phrase will be removed in the next steps.

Phrases containing one word	Phrases containing two words	Phrases containing three words
this *	this is *	this is a *
is *	is a *	is a scientific *
a *	a scientific *	a scientific paper *
scientific	scientific paper	
paper		

3. STATISTICAL PROCESSING

There are two major approaches in text processing generally as bellow:

1. Linguistic approaches
2. Statistical approaches

Linguistic approaches, particularly Natural Language Processing (NLP) tools, leverage techniques such as stemming, POS tagging, parsing, tokenizing, and named entity recognition systems. These methods empower NLP tools to attain a heightened level of semantic understanding in text, revealing valuable information and elevating precision and accuracy. Despite sharing a common goal, the implementation of these techniques varies across languages due to syntactical and structural distinctions.

The second approach involves utilizing statistical information derived from the text. Typically, the text is regarded either as a sequence of words or as a collection of words, and statistical analysis is employed to identify crucial elements or keywords. Word frequency in a text stands out as a notable analytical measure. The prevalent vocabulary effectively conveys the core message of this paper. Our method, grounded in statistical approaches, is presented as a versatile solution applicable to any language. It's worth noting that while our method supports multi-language texts, NLP tools with comparable outputs are not universally available. Moreover, some tools may yield incomplete or imprecise results (as referenced in [23]). The proposed method addresses this limitation, ensuring suitable results for multi-language texts where current NLP tools may fall short.

For this segment, the input is a list named "Active Phrases," and five parameters are calculated as statistical information. The first parameter, phrase frequency, indicates the frequency of each phrase in the text. Evaluating each phrase based on frequency is essential for determining its significance, particularly after the removal of stop words. However, it's important to note that this concept does not always hold true for extracting key phrases.

Second parameter -subphrase_count- shows the number of its subphrases that exist in Active Phrases List. A subphrase is a string of words that can be seen in longer phrases. Single-word phrases only contain one subphrase which is the same as the phrase. Two-word and three-word phrases contain three and six subphrases, respectively. For example, suppose that α, β and γ are three words and “α β γ” is a phrase. For this phrase, subphrases will be: {“α”, “β”, “γ”, “α β”, “β γ”, “α β γ”}. It is possible that some subphrases have been removed from Active Phrases List in the stop word removal phase, So subphrase_count for a phrase will be equal to number of subphrases that exist in Active Phrases List. Subphrase_count shows that how much a phrase is reliable. Our researches show that phrases have low usefulness probability, if their valid subphrases are less than a threshold. Threshold determines according to the length of the phrase. For a phrase, third parameter -subphrase_frequency- is the summation of frequencies of its subphrases. Better understanding of a phrase is realized by subphrase_frequency, especially for those phrases with low frequency. Maybe an important phrase occurs a few times in a text but its subphrases have high frequency. This parameter helps longer phrases get higher scores unlike their low iteration counts.

Each substring extracted in section two, contains several phrases, and for each phrase, subphrase_frequency is calculated in the past step. Forth parameter -substring_score- shows the importance of phrase among all phrases of text. Substring_score is calculated by summation of subphrase_frequency of a phrase and its subphrases devided by summation of the number of phrases and subphrases. For example, in the phrase “This is a scientific paper”, after stop word removal, three phrases remain in the Active Phrases List. Suppose that, a document contains only this phrase, so parameters for each phrase is shown in Table 2.

Table 2. Example of statistical information

Phrase	phrase frequency	Subphrase count	Subphrase frequency	Substring score
scientific	1	1	1	5/3 = 1.6
Paper	1	1	1	5/3 = 1.6
Scientific paper	1	3	3	5/3 = 1.6

4. LOGICAL PROCESSING

Based on statistical information extracted from previous section, some logical rules are illustrated to select the best phrases which show text's main points. These logical rules help us to detect unimportant phrases and omit them from Active Phrases List. For better understanding of rules, they have been divided in to two clusters. In the following, each rule is introduced in the related cluster and its reasons are explained.

4.1. Static Thresholds Cluster

First cluster contains two rules which use static thresholds for the phrase_frequency and subphrase_count parameters as below:

- phrase_frequency static threshold satisfaction
- subphrase_count static threshold satisfaction

It is unlikely for a key phrase to appear only once in a document. Typically, an important phrase should be referenced multiple times. Therefore, we set the phrase frequency threshold to one. If a phrase occurs only once in the text, it will be omitted. A two-word phrase contains three subphrases, while a three-word phrase contains six subphrases (see section II). Analytical experiments indicate that phrases without sufficient subphrases in the active phrases list are unlikely to be significant. If the subphrase count falls below a certain threshold for each phrase, it is considered a less meaningful phrase and should be excluded from the active phrases list.

4.2. Dynamic Threshold Cluster

This cluster uses dynamic thresholds for subphrase_frequency and substring score. It is very hard and confusing task to assign static thresholds to these parameters because of dynamic statistical information extracted from text. This cluster also contains two rules like previous cluster as mentioned below:

- Subphrase_frequency dynamic threshold satisfaction
- Substring_score dynamic threshold satisfaction

The paragraph assumes that a phrase's significance is contingent on its context, emphasizing the need to ascertain whether an important phrase can exist within an unimportant context for a comprehensive understanding of the text's main points. In essence, the central question revolves around whether a phrase retains significance even when situated in an unremarkable context. Addressing this inquiry gives rise to a pivotal hypothesis known as textual correlation. This hypothesis suggests that phrases within unimportant contexts may be dispensable. Our analysis of 100 articles substantiates the validity of this hypothesis, revealing that less important phrases generally fail to convey significant concepts. Consequently, it becomes imperative to intentionally position important phrases within meaningful contexts, as they are not random occurrences in the text.

To operationalize this hypothesis, we leverage subphrase_frequency and substring_score. Specifically, we calculate the average of both the subphrase_frequency and substring_score parameters across all document phrases. These resulting averages then serve as the subphrase_frequency threshold and substring_score threshold, respectively. This approach allows for a nuanced evaluation of the interplay between phrases and their contexts, shedding light on the importance of strategic placement for effective communication of key concepts within the text. In order to use these thresholds some rules are established as follows:

- If a phrase contains only one word and its subphrase_frequency parameter or substring_score parameter be less than their thresholds then it should be omitted.
- If a phrase occurred less than two times and its subphrase_frequency parameter is less than threshold then it should be omitted.
- If both subphrase_frequency and substring_score parameters are less than their threshold for a phrase then it should be omitted.
- If phrases which are before or after a phrase, have substring_score near to zero, then the context for that phrase is known as unimportant and that phrase should be omitted.

It should be mentioned that these rules are not the only ones and some other rules can be added to increase the accuracy and precision. After these sections, remaining phrases named as “Important Phrases List”, will be the input of the next section.

5. PHRASE RECREATION

As elucidated in Section 2, the extraction process involves capturing all phrases with up to three words from the text. However, it is plausible that an important phrase in the text may extend beyond three words. The objective of this section is to construct longer important phrases that might not be initially extracted. To achieve this, we consider combinations of each pair of extracted keyphrases in the important phrases list, excluding one-word phrases. In each pair, if the end of one phrase aligns with the start of another, they can be amalgamated to create a longer phrase. The acceptability of this combination is contingent upon the existence of the longer phrase in the text. If validated, the output phrase is then replaced with these two combined phrases. While the pursuit of shorter phrases is crucial in some contexts, experimental analysis reveals that longer, meaningful phrases encapsulate more useful concepts. Thus, combining phrases offers two significant advantages:

a) Decreasing key phrase number

b) Longer phrases contain more information

6. EVALUATION

To evaluate proposed method, 100 articles are collected as a database. Most of these articles discus the text processing and the remaining are in other fields like image processing, electronic and etc. Articles have a page range number between four to twelve pages. Each article text set as input data. The output of this method for an article is some phrases that show the article's focuses. For each result, someone asked to classify each phrase in two classes. Some of these persons were article's author and some other were familiar with article and its aspect. Phrases are determined whether they are meaningful or not or whether they are related to the article concepts or not? Phrases are categorized into two distinct groups: those that convey meaning, constituting the first category, and those that offer crucial insights related to the article, falling into the second category. It is essential to note that if a phrase does not belong to the first category, it cannot belong to the second category. However, a phrase may belong to the first category without being a member of the second. For instance, Table 1 illustrates some phrases extracted from [23], accompanied by the corresponding questions they answer. Each article, on average, yields 42 phrases, with two questions posed for each phrase. The results reveal that 97.49 percent of the phrases are classified as meaningful, and 90.06 percent of the phrases pertain to the article's overarching concept.

The method's accuracy is gauged by the percentage of meaningful phrases, showcasing that the proposed approach excels in selecting accurate phrases while avoiding less relevant ones. Furthermore, the precision of the method can be ascertained by the percentage of phrases aligning with the article's concept. The essence phrases identified through this method play a crucial role in pinpointing the article's main points. The high accuracy and precision achieved by the proposed approach underscore its efficacy in keyphrase extraction, emphasizing its utility in distilling meaningful and conceptually relevant information from the text.

Also, we use this method for text classification and compare it with Naive Bayesian classification. For this purpose, six news groups of twenty news groups from [24] are selected as below:

1. Comp.graphics
2. Misc.forsale
3. Rec.autos
4. Sci.crypt
5. Soc.religion.christian
6. Taik.politics.guns

The initial step involves classifying news groups using the Naive Bayesian classification algorithm, and to ensure result accuracy, we employ both 10-fold and 5-fold cross-validation. It is crucial to emphasize that while Naive Bayes considers all words in the text for classification, our proposed method achieves superior results by classifying news groups based on the essence phrases it extracts, rather than utilizing all words. Surprisingly, the

accuracy of classification using both methods is comparable. This evaluation underscores the effectiveness of our proposed method in selecting essential phrases, as Naive Bayes, despite considering all text words, fails to significantly outperform our approach. The implication is that the proposed essential phrases encapsulate the most crucial concepts within documents, rendering other phrases less likely to convey important or useful meanings.

To make a relative comparison with other articles, assuming the overall average from the six news groups represents the entire 20 news groups in Reuters's dataset, direct comparisons should be approached cautiously due to potential discrepancies in conditions, such as selecting the same instruction and evaluation data. Nonetheless, our proposed method exhibits high precision in selecting keyphrases from the passage, further validating its efficacy in capturing essential concepts from news articles.

Table 1. Evaluation of some phrases extracted from [24]

Phrase	Meaningful	Related to concept
based fuzzy inference systems	No	No
Information	Yes	No
Method	Yes	No
emotional learning	Yes	Yes
fuzzy neural network	Yes	Yes
output membership functions	Yes	Yes
Predict sunspot time series	Yes	Yes
Complex fuzzy sets	Yes	Yes
Solar activity forecasting	Yes	Yes
quantum neural networks	Yes	Yes
time delay line recurrent fuzzy inference systems	Yes	Yes

Table 4. Classification rate for 6 news groups

	5-fold	10-fold	Mean
Naive Bayes	86.50	87.70	87.10
Proposed method	82.60	84.80	83.70
DF + SVM [25]	-------	-------	84.71
PCA + SVM [25]	-------	-------	83.97
LDA + SVM [25]	-------	-------	88.96

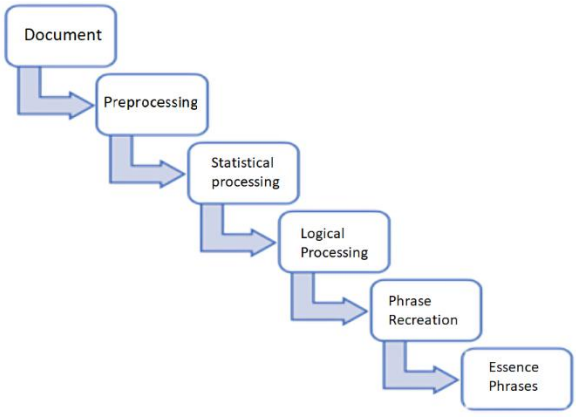

Fig. 1. Essence phrase extraction steps

7. CONCLUSION

This paper introduces a novel method designed to extract the essence of a text in the form of key phrases from a single document. The primary objective of our research is to gain a comprehensive understanding of the main

points within a text by leveraging the extracted essence phrases. The method presented herein adopts a language-independent approach, utilizing a combination of statistical analysis and logical rules. The key steps of our proposed keyphrase extraction algorithm are outlined, featuring an overall time complexity of O(n2) in the worst-case scenario. Notably, this approach stands out as it operates without the reliance on Natural Language Processing (NLP) tools or specialized resources such as ontologies or corpuses. Furthermore, it exhibits versatility, being applicable to any document across various fields without the need for specific domain training data.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Ajallouda, L., Fagroud, F., Zellou, A., & Lahmar, E. (2022). KP-USE: An Unsupervised Approach for Key-Phrases Extraction from Documents. International Journal of Advanced Computer Science and Applications. https://doi.org/10.14569/ijacsa.2022.0130433.

[2] Schopf, T., Klimek, S., & Matthes, F. (2022). PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction, 243-248. https://doi.org/10.5220/0011546600003335.

[3] Wang, H., & Li, J. (2022). Unsupervised Keyphrase Extraction from Single Document Based on Bert. 2022 International Seminar on Computer Science and Engineering Technology (SCSET), 267-270. https://doi.org/10.1109/scset55041.2022.00068.

[4] Bracewell, D. B., Ren, F., & Kuriowa, S. (2006). Multilingual single document keyword extraction for information retrieval. 2005 International Conference on Natural Language Processing and Knowledge Engineering. Wuhan, China. doi:10.1109/nlpke.2005.1598792

[5] Wan, X., Yang, J., & Xiao, J. (2007). Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In Proceedings of the 45th annual meeting of the association of computational linguistics (pp. 552-559).

[6] D'Avanzo, E., Magnini, B., & Vallin, A. (2004). Keyphrase extraction for summarization purposes: The LAKE system at DUC-2004. In Proceedings of the 2004 document understanding conference.

[7] Tonella, P., Ricca, F., Pianta, E., & Girardi, C. (2004). Using keyword extraction for Web site clustering. Fifth IEEE International Workshop on Web Site Evolution, 2003. Theme: Architecture. Proceedings. Amsterdam, Netherlands. https://doi.org/10.1109/WSE.2003.1234007

[8] Lee, S., & Kim, H.-J. (2008). News keyword extraction for topic tracking. Fourth International Conference on Networked Computing and Advanced Information Management. (NCM), Gyeongju, South Korea. https://doi.org/10.1109/NCM.2008.199

[9] Wang, C., Zhang, M., Ru, L., & Ma, S. (2008). An automatic online news topic keyphrase extraction system. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Sydney, Australia. https://doi.org/10.1109/WIIAT.2008.225

[10] Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. International Journal of Artificial Intelligence Tools: Architectures, Languages, Algorithms, 13(01), 157-169. https://doi.org/10.1142/S0218213004001466

[11] Ohsawa, Y., Benson, N. E., & Yachida, M. (1998, April). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In Research and Technology Advances in Digital Libraries,

1998. ADL 98. Proceedings. IEEE International Forum on (pp. 12-18). IEEE.

[12] Hulth, A. (2003, July). Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 conference on Empirical methods in natural language processing, 216-223. Association for Computational Linguistics. https://doi.org/10.3115/1119355.1119383

[13] Turney, P. D. (2000). Learning algorithms for keyphrase extraction. Information retrieval, 2(4), 303-336. https://doi.org/10.1023/A:1009976227802

[14] Islam, M. R., & Islam, M. R. (2008, December). An improved keyword extraction method using graph based random walk model. In Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on, 225-229. IEEE. https://doi.org/10.1109/ICCITECHN.2008.4802967

[15] Kaur, J., & Gupta, V. (2010). Effective approaches for extraction of keywords. International Journal of Computer Science Issues (IJCSI), 7(6), 144.

[16] Chien, L. F. (1997, July). PAT-tree-based keyword extraction for Chinese information retrieval. In ACM SIGIR Forum , 31, 50-58. ACM. https://doi.org/10.1145/278459.258534

[17] Onoda, T., Yumoto, T., & Sumiya, K. (2008, December). Extracting and Clustering Related Keywords based on History of Query Frequency. In Universal Communication, 2008. ISUC'08. Second International Symposium on,162-166. IEEE. https://doi.org/10.1109/ISUC.2008.22

[18] Khoury, R., Karray, F., & Kamel, M. S. (2008). Keyword extraction rules based on a part-of-speech hierarchy. International Journal of Advanced Media and Communication, 2(2), 138-153. https://doi.org/10.1504/IJAMC.2008.018504

[19] Paukkeri, M. S., Nieminen, I. T., Pöllä, M., & Honkela, T. (2008). A language-independent approach to keyphrase extraction and evaluation. Coling 2008: Companion volume: Posters, 83-86.

[20] Aquino, G. O., Hasperué, W., Estrebou, C. A., & Lanzarini, L. C. (2013). A novel, language-independent keyword extraction method. In XVIII Congreso Argentino de Ciencias de la Computación.

[21] Wu, C., Marchese, M., Wang, Y., Krapivin, M., Wang, C., Li, X., & Liang, Y. (2009, December). Data preprocessing in SVM-based keywords extraction from scientific documents. In Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on, 810-813. IEEE. https://doi.org/10.1109/ICICIC.2009.155

[22] Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. Web Intelligence and Agent Systems: An International Journal, 7(2), 195-207. https://doi.org/10.3233/WIA-2009-0162

[23] Moghaddam, J. D., Mosallanezhad, A., & Teshnehlab, M. (2013, August). Sunspot prediction by a Time Delay line Recurrent Fuzzy Neural Network using emotional learning. In Fuzzy Systems (IFSC), 2013 13th Iranian Conference on ,1-5. IEEE.

[24] Joachims, T. (1996). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization (No. CMU-CS-96-118). Carnegie-mellon univ pittsburgh pa dept of computer science.

[25] Luo, L., & Li, L. (2014). Defining and evaluating classification algorithm for high-dimensional data based on latent topics. PloS one, 9(1), e82119. https://doi.org/10.1371/journal.pone.0082119