ISSN: 2182-2069 / E-ISSN: 2182-2077

Improving Terminologies Synonym Expansion Model for Cultural Heritage Contents

Dr. Wafa' Za'al Alma'aitah[1*], Dr. Fatima N. AL-Aswadi[2], and Dr. Rami S. Alkhawaldeh[3]

[1*]Department of Intelligence Systems, Faculty of Artificial Intelligence, Al-Balqa Applied University, Al-Salt, Jordan. wafaa_maitah@bau.edu.jo, https://orcid.org/0000-0001-5183-1654

[2]Institute of Computer Science and Digital Innovation, UCSI University, Kuala Lumpur, Malaysia. Fatima.Nadeem@ucsiuniversity.edu.my, https://orcid.org/0000-0001-5413-1207

[3]Department of Computer Information Systems, The University of Jordan, Aqaba, Jordan. r.alkhawaldeh@ju.edu.jo, https://orcid.org/0000-0002-2413-7074

Received: January 07, 2024; Revised: February 28, 2024; Accepted: March 30, 2024; Published: May 30, 2024

Abstract

The cultural heritage (CH) domain possesses large volumes, necessitating users to provide more precise details regarding their requirements. Nonetheless, several formidable challenges are observed by the CH information retrieval researchers, including vocabulary issues and access points. Hence, an increasing demand for models capable of addressing these issues and professional search systems are required. These models enable users to search efficiently inside the CH domain. Many non-experts among its users are also typically attracted by CH content, necessitating improved access models to these rich contents. Therefore, this study investigated a terminologies synonym expansion (TSE) model for CH content. The proposed model combined three elements in the framework: TextRank algorithm capacity for terminology identification, comprehensive WordNet lexical database for synonym expansion, and synonym linking to their respective terminologies. Consequently, two CH collections (CHiC2013 and CHiC2013_EDE) demonstrated a noteworthy enhancement compared to the traditional information retrieval methods. This model could bridge the vocabulary disparity between non-expert users and the specialised terminology employed in the CH domain.

Keywords: Synonym, TextRank Algorithm, Cultural Heritage.

1 Introduction

Cultural heritage (CH) involves past tangible and intangible preserved elements, including traditions, languages, handicrafts, architecture, and natural ecosystems. Hence, effective and efficient CH information retrieval has become increasingly paramount in the current rapid digitisation and globalisation era. The preservation and accessibility of cultural artefacts can also impact the historical understanding, current lifestyle, and future legacy. Nevertheless, the invaluable information collection has become increasingly challenging to discover and safeguard due to the exponential growth of digital artefacts, data, and documents. This study examined the challenges associated with CH-related information retrieval while highlighting the necessity of developing innovative solutions to address these

Journal of Internet Services and Information Security (JISIS), volume: 14, number: 2 (May), pp. 237-246.
DOI: *10.58346/JISIS.2024.I2.015*

*Corresponding author: Department of Intelligence Systems, Faculty of Artificial Intelligence, Al-Balqa Applied University, Al-Salt, Jordan.

challenges. Consequently, future generations could appreciate and learn from their extensive cultural history.

A growing interest in sizeable digital library creations has been observed in recent decades. Numerous major initiatives have received national (Gallica) or international support (The European Library or Europeana) (Alma'aitah et al., 2020; Agina-Obu & Oyinkepreye Evelyn, 2023). These projects aim to preserve CH while offering global users access to priceless artefacts. The CH information has recorded higher access opportunities due to the increasing contemporary hardware capabilities and costs. Access to CH information has also become more feasible through online platforms owing to increasing internet capabilities. Therefore, an effective CH data management system can facilitate user access and information exploration on CH collections (Colace et al., 2020; Tanja & Milica, 2023).

The CH collection is considered a domain-specific collection (Chakma & Chowdhury, 2023). Conversely, users' inadequate specific terminologies or general phrase searches can generate vocabulary issues (Azad & Deepak, 2019); Gao et al., 2021; Warburton, 2023). The terminologies users use in their queries can differ from those used in the CH collection. This process results in a retrieval failure of relevant articles, lower overall recalls, and higher matching process difficulties (mismatch terms) (Jayasree et al., 2012). Hence, several methodologies have been suggested to address this concern and manage the synonymy exclusively from the query perspective (Fitzgerald et al., 2021).

Although previous studies have addressed this issue, incorporating synonyms into the collection during indexing to prevent additional processing time at query execution has not been focused. These studies also do not consider the synonym correlations while designing retrieval models (Zamani et al., 2020). Thus, this study addressed the necessity for an enhanced IR system that effectively managed CH content for users. A document synonym model called the terminologies synonym expansion (TSE) model was created to assist non-expert users in accessing information from collections with greater efficiency. The proposed model combined three elements in the framework: TextRank algorithm for terminology identification, WordNet lexical database for synonym expansion, and synonym linking to the respective terminologies.

Overall, the vocabulary gap between non-expert users and the specialised terms used within the CH domain could be bridged using this model. This process facilitated efficient retrieval and content understanding of the collections. The remainder of this study is organised as follows: Section 2 presents the literature review. Section 3 describes the proposed TSE model, while Section 4 provides the experimental results and discussions. Finally, Section 5 concludes this study.

2 Literature Review

Individuals often encounter comprehension and complicated vocabulary access issues regarding CH. Particularly, specialised terminologies and jargon in CH databases are concerning. One example is a museum catalogue entry vaguely referring to the portolan chart concept. This search challenges someone unfamiliar with the subject to comprehend it as a distinct navigational map type. A crucial access point is produced, in which proficiency-lacking individuals encounter issues finding and engaging with relevant resources. Therefore, a potential solution involves user interface development with higher comprehension and better navigation (Shaik, 2020). This solution is achieved by using explanations or alternative keywords to aid non-proficient users in overcoming their knowledge gap.

A significant challenge can arise from language barriers due to the multiple language characteristics of CH collections. These repositories often contain a wide range of materials in several languages, such as Arabic, Greek, and Latin. The primary issue in this scenario is the difficulty inexperienced users

encounter when trying to access and understand resources written in unfamiliar languages. Hence, language translation tools or contextual explanations can significantly contribute to overcoming this issue while improving the accessibility and inclusion of CH. A more seamless interaction between users and the vast array of CH items is facilitated by recognizing and addressing the vocabulary difficulties and points of entry. This outcome enhances the understanding and generates highly accessible resources for a wider demographic.

Previous studies documented that vocabulary presented a significant obstacle to CH information access (Dragoni et al., 2017). Two studies (Binding & Tudhope, 2016; Davis & Heravi, 2021) reported the importance of creating effective access points to bridge the divide between expert and non-expert users. Hence, terminology comprehension and synonyms are crucial for effective information retrieval in a specialised field (such as the CH domain) (Križaj, et al., 2022). The vocabulary issues in such scenarios can also result in misinterpretations (Fitzgerald et al., 2021; Jorge, et al., 2017).

Synonym expansion has emerged as a potential solution to address the vocabulary issue. A study (Vlachidis & Tudhope, 2016) assessed the synonym expansion mode using the semantic expansion approach. The semantic expansion of a mode encompasses synonyms for glossary terms in thesauri structures. Moreover, the overlapping terms in the glossary and thesaurus contain common word senses due to the semantic alignment between ontological entities and terminology resources (AL-Aswadi et al., 2023; Jain et al., 2021).

The retrieval outcomes can be enhanced using enriching texts in domain-specific corpora from WordNet information and highly precise word embedding. A study (Yusuf, et al., 2019) introduced a query expansion approach using word synonyms as explicit relevant feedback. The study highlighted that accurate judgement and word synonyms could improve search quality in search engines. Hence, efficient keyword extraction is beneficial for structuring and understanding literature, mainly when vocabulary-related issues are widespread (Smith, 2021). Likewise, (Wei & Ding, 2023) discussed that the TextRank algorithm was commonly employed for keyword extraction (Salem & Stolfo, 2010). The study revealed that the algorithm functioned by considering word associations as nodes within a graph. Words acquiring more connections to other words were considered more significant and likely to be selected as keywords (Qiu & Zheng, 2022; Wei & Ding, 2023).

Another study (Pan et al., 2019) demonstrated that the TextRank algorithm outperformed TF-IDF in keyword extraction tasks. Similarly, (Martinez-Romo, et al., 2016) explored semantic relationship graphs for keyword extractions from a corpus of texts. The initial step in graph creation is merging word co-occurrences and related senses. Specifically, word associations are formed by analysing a substantial number of occurrences inside a document representing a semantic unit (rather than solely relying on basic co-occurrence). The word relations of the created graph are subsequently improved with WordNet data. Considering that vocabulary issues are addressed using WordNet by providing synonyms and related words, word comprehension and suitable phrase selection in text processing can be enhanced.

WordNet facilitates the comprehension and elucidation of word meanings by offering synonyms and word correlations. A study (Wei et al., 2015) applied the WordNet lexicon to determine semantically related words, which presented a practical illustration of how WordNet could be employed. The study was highly beneficial for keyword extraction, as a significant idea-based paragraph or page included terms or synonyms semantically relevant to the phrase or keywords. Alternatively, (Jain et al., 2021) introduced WordNet to produce synonyms from textual description terms. The study then incorporated the synonyms into the training dataset. Overall, the capacity of the TextRank algorithm to identify terminology with the comprehensive WordNet lexical database for synonym extension can significantly enhance information retrieval effectiveness (Zhou et al., 2022).

3 The Proposed TSE Model

The linguistic concerns between non-expert users and CH-related field terms could be addressed by the proposed TSE Model in this study. This discrepancy hindered the successful data retrieval from the CH collection, reducing the overall content retrieval efficiency. Given that the CH domain was highly specialised, the vocabulary issue persisted. This study proposed the TSE Model to facilitate CH content-related access and study for non-expert users, bridging the lexical gap between them and their users. The proposed TSE model contained three elements in the framework, including TextRank, WordNet, and synonym mapping, which effectively addressed the CH information retrieval issue. Thus, this model could resolve the vocabulary concerns while enhancing the search results and user experience.

The TextRank algorithm is commonly used in content rating by applying natural language processing and graph theory to detect significant keywords and phrases inside articles. This algorithm outperforms other approaches (such as TF-IDF) in effectively managing synonyms and capturing semantic linkages for addressing vocabulary issues. Meanwhile, synonym analysis uses the WordNet database to provide an extensive semantic network of words and their interconnections. This database enhances query expansion and comprehension of user intent, which is efficient due to its widespread adoption and comprehensive coverage. Alternatively, user searches are improved by Thesaurus mapping through synonym and related term integrations, aligning better with CH terms. Compared to complex ontological systems or deep learning-based techniques, the mapping is simple, scalable, and interpretable. Figure 1 depicts the proposed workflow for the TSE model, which consists of three sequential stages: terminology definition, synonym term extractions, and bag of term creation. The subsequent subsections explain each stage.

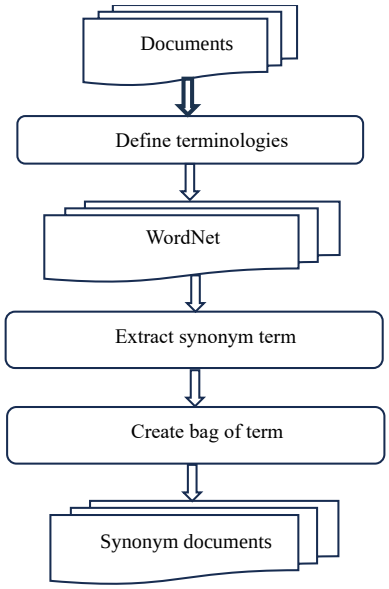

Fig. 1: The workflow of the proposed TSE model in this study

3.1 Terminology Definition

The initial stage of the proposed TSE model involved establishing precise definitions for the terminologies used in each article and outlining the specific workflow elements (see Figure 2). This process was achieved by employing data cleaning and pre-processing procedures while attempting to

analyse text data, which was conducted before data vectorisation. Table 1 tabulates the data-cleaning strategies used in this study. The text graph is represented as an undirected graph (G) as follows:

$$G = (V, E)$$

Where V and E are the sets of words and edges connecting the words, respectively.

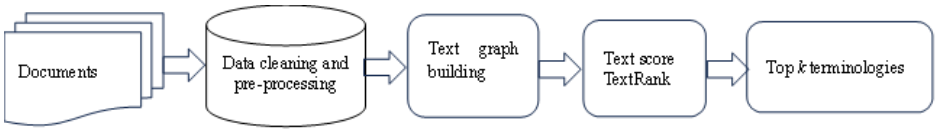

Fig. 2: The workflow of the terminology definition process

Table 1: Summary of the data cleaning techniques used in this study

Data cleaning and pre-processing techniques
Stop word removal
Lemmatisation
Tokenisation
Wordpiece

The text graph production involved two distinct processes. Initially, candidate keywords were selected from the text and considered as nodes in V. Subsequently, edges were created inside a defined window of size w. A co-occurrence relation generated these edges of a text graph, while the sliding window size for the input text was modified (Mihalcea et al., 2004). If the words (v_i and v_j) appeared in a sentence in the content of window size w, they were treated as nodes and added to V. Finally, an edge was produced between v_i and v_j.

The primary stage in the proposed TSE model was to recognise and establish the terminologies in each CH-based article. This process involved primary term extractions with distinct meaning and relevance to the domain. Any redundant terminologies were removed to provide a clear and succinct depiction using the TextRank algorithm. A study (Mihalcea et al., 2004) proposed the TextRank algorithm as a graph-based ranking algorithm inspired by PageRank. Generally, TextRank has been extensively utilised for several natural language processing tasks, such as keyword extraction, text summarisation, and automatic terminology extraction. Previous studies demonstrated that the TextRank Algorithm was excellent in keyword extractions compared to other measurement approaches (Huang & Xie, 2022; Li, et al., 2019; Pan et al., 2019; Rani & Bidhan, 2021). Other studies (Kazemi et al., 2020; Zuo et al., 2017) also denoted that the TextRank Algorithm was a common relevance measuring tool.

The fundamental approach of TextRank was creating a graph that represented the words and their correlations inside an article. Subsequently, the most crucial word vertices were identified using recursively computed importance scores over the entire graph. Sentence and word parsing were also utilised to extract candidates from the text and provide a list of terms for evaluation. Each word and its corresponding correlation with other words were then incorporated into a sliding window enclosing the word on the graph. An iterative ranking algorithm was used for each vertex, continuously updating the word scores (depending on the scores of related words) until the values reached a stable state. Table 2 lists the TextRank parameter settings (Zhang et al., 2020).

Table 2: Summary of the TextRank parameter settings

TextRank parameter	Parameter setting
Co-occurrence window size (w)	3
Iteration number (t)	20
Damping factor (d)	0.9
Rank (k)	10

The TextRank algorithm utilises the following equation 1 to update the TextRank score of a node repeatedly. This equation relies on a graphical representation consisting of nodes V_i and edges connecting nodes with a weight w_{ji} as follows:

$$TextRank(V_i) = (1-D) + D * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in out(V_j)} w_{jk}} * TextRank(V_j) \quad (1)$$

Where D is the damping factor. The collection of nodes directed by v_j was represented by out(v_j), and the weight of the edge from node w_j to node w_i was represented by w_{ji}.

3.2 Synonym Term Extractions

The second stage entailed vocabulary expansion by establishing connections between each phrase and its corresponding terms. This process was achieved using the WordNet lexical database. Considering that WordNet was previously effective in information retrieval tasks, certain researchers employed this method for improving query terms with related concepts. This study applied the WordNet database due to its accessibility (free access) and several mappings to WordNet inside the utilised ontologies. An inherent benefit of WordNet was its hierarchical organisation. The words in WordNet were classified as synsets (collections of synonyms with a shared definition), and this structure facilitated the process of identifying correlated terms.

A synset was formed by merging all the senses and contextual variations of a term with many meanings. The WordNet then enhanced lexical representation by establishing connections between synsets instead of individual words, resulting in a more complete and cohesive structure. WordNet was also utilised to select the suitable synset for each term identified in the previous stage. Once the synset was identified, all terms inside were considered potential synonyms for the provided terminology. Consequently, fewer supplementary measures were necessary.

3.3 Bag of Term Creation

The final stage involved establishing a connection between the terms in the collection and their respective terminology. This procedure presented a direct correlation between the synonyms in the word bag and the original represented terms. Therefore, various synonyms linked to each term could be conveniently retrieved and comprehended by the users, facilitating efficient information retrieval from the CH collection.

4 Results and Discussions

This study assessed the synonym impact on the vocabulary of the CH-related articles. The process involves two collections to examine the effectiveness of the proposed TSE model as follows:

1. CHiC2013: The CH collections in the Europeana database have been compiled (since 2014) to aid IR-related researchers in the practical assessment of information access involving CH contents (Alma'aitah et al., 2020).
2. ECHiC2013_EDE (Talib & Osman, 2019).

Three metric types (MAP, P@10, and F-score) were applied to evaluate the effectiveness, which was vital for assessing the proposed TSE model efficiency concerning CH information retrieval. A language model was also included in this study as the retrieval model for the CH collections based on the methodology of (Alma'aitah et al., 2021) study. Subsequently, the ECHiC2013_TSE and

ECHiC2013_EDE_TSE collections were expanded by the proposed model. Table 3 presents the MAP, P@10, and F-score metric outcomes for CHiC2013, ECHiC2013_DE E, CHiC2013_TSE, and ECHiC2013_EDE_TSE. The collective measurements of the proposed TSE model could effectively retrieve culturally significant articles. This process was followed by appropriate ranking and satisfactory user experience delivery. Therefore, these metrics provided a comprehensive study regarding the performance of the model. Well-informed choices could also be made regarding the improvement and implementation of the model.

Table 3: The MAP, P@10, and F-score metric outcomes for CHiC2013, ECHiC2013_EDE, CHiC2013_TSE

Collections	**MPA**	**P@10**	**F-score**
CHiC2013 (Baseline)	0.502	0.419	0.434
ECHiC2013_EDE (Baseline)	0.550	0.488	0.462
ECHiC2013_TSE	0.540	0.482	0.480
ECHiC2013_EDE_TSE	0.613	0.565	0.547
Improvement (ECHiC2013_TSE, CHiC2013)	7.6%	15.0%	10.6%
Improvement (ECHiC2013_EDE_TSE,ECHiC2013_EDE)	11.5%	15.8%	18.4%
Improvement (ECHiC2013_EDE_TSE, ECHiC2013_TSE)	13.5%	17.2%	14.0%

The ECHiC2013_TSE demonstrated 7.6% (MAP), 5.0% (P@10), and 10.5% (F-score) improvements compared to CHiC2013. Meanwhile, the ECHiC2013_EDE_TSE presented 11.5% (MAP), 15.8% (P@10) and 18.4% (F-score) improvements to ECHiC2013_EDE. Likewise, the ECHiC2013_EDE_TSE improved by 13.5% (MAP), 17.2% (P@10), and 14.0% (F-score) compared to ECHiC2013_TSE. Previous studies highlighted that incorporating synonyms into CH terminologies could improve information retrieval. The capacity to generate queries from these articles was also enhanced by employing synonym terminology. Consequently, the quantity of articles obtained increased while the probability diminished to zero.

The *t*-test findings (*t* and *p*-values) for each method pair suggested that the *p*-value was less than 0.05. This observation implied that the performances of the ECHiC2013_TSE and TSE, ECHiC2013_EDE were significantly superior to the corresponding comparative benchmarks (ECHiC2013_EDE and ECHiC2013_EDE). Considering that the *p*-value was less than 0.05 at a 95% confidence level, the proposed ECHiC2013_TSE and TSE, ECHiC2013_EDE statistically outperformed ECHiC2013_EDE and ECHiC2013_EDE. The *p*-value was also less than the α level ($p < 0.05$), indicating significant differences between the means of the comparison approach and the metrics improvements (MAP, P@10, and F-score). Overall, the low *P*-values in all the presented models proved that the results were not obtained by chance and that the improvements were statistically significant.

5 Conclusion and Future Works

This study successfully investigated significant concerns regarding terminologies in CH information retrieval applications. The proposed TSE model effectively addressed the knowledge gap between specialists and non-experts. This model facilitated a more comprehensive engagement with CH content. When the TextRank algorithm, WordNet lexical database, and intentional coupling of synonyms to related terminologies were incorporated into this model, the vocabulary issue often hindering efficient information retrieval could be resolved. Two CH collections (CHiC2013 and CHiC2013_EDE) were

successfully validated using the proposed TSE model, revealing the potential of the model to enhance information retrieval performance.

Compared to previous approaches, the proposed TSE model effectively addressed the CH domain complexities while adapting to the diverse requirements of users (specialists and non-experts) due to its improved outcomes. Nevertheless, further studies should involve expanding the functionalities of the model to accommodate multilingual settings for better user accessibility and global relevance. The proposed TSE model can also integrate user feedback mechanisms and customised recommendations based on individual search patterns. Hence, modifying the proposed TSE model to accommodate user preferences can enhance personalised and efficient information retrieval experiences.

References

[1] Agina-Obu, R., & Oyinkepreye Evelyn, S.G. (2023). Evaluation of Users' Satisfaction of Information Resources in University Libraries in Nigeria: A Case Study. *Indian Journal of Information Sources and Services, 13*(1), 1–5.

[2] AL-Aswadi, F.N., Chan, H.Y., & Gan, K.H. (2023). Enhancing relevant concepts extraction for ontology learning using domain time relevance. *Information Processing & Management, 60*(1), 103140. https://doi.org/10.1016/j.ipm.2022.103140

[3] Alma'aitah, W.Z.A., Talib, A.Z., & Osman, M.A. (2020). Opportunities and challenges in enhancing access to metadata of cultural heritage collections: a survey. *Artificial Intelligence Review, 53*(5), 3621-3646.

[4] Alma'aitah, W.Z.A., Talib, A.Z., & Osman, M.A. (2020). The digital resources objects retrieval: Concepts and figures. *In International Conference of Reliable Information and Communication Technology*, Cham: Springer International Publishing, 430-438.

[5] Alma'aitah, W.Z.A., Talib, A.Z., & Osman, M.A. (2021). Towards adaptive structured Dirichlet smoothing model for digital resource objects. *Multimedia Tools and Applications, 80*, 12175-12194.

[6] Azad, H.K., & Deepak, A. (2019). Query expansion techniques for information retrieval: a survey. *Information Processing & Management, 56*(5), 1698-1735.

[7] Azad, H.K., Deepak, A., Chakraborty, C., & Abhishek, K. (2022). Improving query expansion using pseudo-relevant web knowledge for information retrieval. *Pattern Recognition Letters, 158*, 148-156.

[8] Binding, C., & Tudhope, D. (2016). Improving interoperability using vocabulary linked data. *International Journal on Digital Libraries, 17*, 5-21.

[9] Chakma, K.S., & Chowdhury, M.S.U. (2023). CSA Implementation Using Novel Methodology: RTL Development. *Journal of VLSI Circuits and Systems, 5*(2), 22-28.

[10] Colace, F., De Santo, M., Lombardi, M., Mosca, R., & Santaniello, D. (2020). A Multilayer Approach for Recommending Contextual Learning Paths. *Journal of Internet Services and Information Security, 10*(2), 91-102.

[11] Davis, E., & Heravi, B. (2021). Linked data and cultural heritage: a systematic review of participation, collaboration, and motivation. *Journal on Computing and Cultural Heritage (JOCCH), 14*(2), 1-18.

[12] Dragoni, M., Tonelli, S., & Moretti, G. (2017). A knowledge management architecture for digital cultural heritage. *Journal on Computing and Cultural Heritage (JOCCH), 10*(3), 1-18.

[13] Fitzgerald, K.A., La Harpe, A.C.D., Uys, C.S., & Bytheway, A.J. (2021). Information retrieval: Solving mismatching vocabulary in closed document collections. *South African Journal of Libraries and Information Science, 87*(2), 42-54.

[14] Gao, L., Dai, Z., & Callan, J. (2021). COIL: Revisit exact lexical match in information retrieval with contextualised inverted list. arXiv preprint arXiv:2104.07186.

[15] Huang, Z., & Xie, Z. (2022). A patent keywords extraction method using TextRank model with prior public knowledge. *Complex & Intelligent Systems, 8*(1), 1-12.

[16] Jain, N., Bartz, C., Bredow, T., Metzenthin, E., Otholt, J., & Krestel, R. (2021). Semantic analysis of cultural heritage data: Aligning paintings and descriptions in art-historic collections. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III, Springer International Publishing*, 517-530.

[17] Jayasree, V., Nithya, M., & Prabaharan, S. (2012). Cloud Data Retrieval for Multi related keyword based on Clustering Technology. *International Journal of Communication and Computer Technologies (IJCCTS), 1*(1), 60-66.

[18] Jorge, N., Alves, J.R., Medeiros, F., & Medina, S. (2017). *Controlled vocabularies in the organisation and management of cultural heritage: practical guidelines*.

[19] Kazemi, A., Pérez-Rosas, V., & Mihalcea, R. (2020). *Biased TextRank: Unsupervised graph-based content extraction*. arXiv preprint arXiv:2011.01026.

[20] Križaj, L., Zlodi, G., Stublić, H., Majer, I., & Telišman, A.L. (2022). Controlled terminology for monuments, museum and gallery objects: preliminary research on vocabularies reconciliation. *In 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, 660-665.

[21] Li, J., Huang, G., Fan, C., Sun, Z., & Zhu, H. (2019). Key word extraction for short text via word2vec, doc2vec, and textrank. *Turkish Journal of Electrical Engineering and Computer Sciences, 27*(3), 1794-1805.

[22] Martinez-Romo, J., Araujo, L., & Duque Fernandez, A. (2016). S em G raph: Extracting keyphrases following a novel semantic graph-based approach. *Journal of the association for information science and technology, 67*(1), 71-82.

[23] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. *In Proceedings of the conference on empirical methods in natural language processing*, 404-411.

[24] Mihalcea, R., Tarau, P., & Figa, E. (2004). PageRank on semantic networks, with application to word sense disambiguation. *In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 1126-1132.

[25] Pan, S., Li, Z., & Dai, J. (2019). An improved TextRank keywords extraction algorithm. In *Proceedings of the ACM Turing Celebration Conference-China*, 1-7.

[26] Qiu, D., & Zheng, Q. (2022). Improving TextRank algorithm for automatic keyword extraction with tolerance rough set. *International Journal of Fuzzy Systems*, 1-11.

[27] Rani, U., & Bidhan, K. (2021). Comparative assessment of extractive summarization: textrank tf-idf and lda. *Journal of scientific research, 65*(1), 304-311.

[28] Salem, M.B., & Stolfo, S.J. (2010). Detecting Masqueraders: A Comparison of One-Class Bag-of-Words User Behavior Modeling Techniques. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 1*(1), 3-13.

[29] Shaik, S. (2020). A coplanar wave guide fed compact antenna for navigational applications. *National Journal of Antennas and Propagation (NJAP), 2*(1), 7-12.

[30] Smith, C. (2021). Controlled Vocabularies: Past, Present and Future of Subject Access. *Cataloging & Classification Quarterly, 59*(2-3), 186-202.

[31] Talib, A.Z., & Osman, M.A. (2019). Document expansion method for digital resource objects. *In IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT)*, 256-260.

[32] Tanja, M., & Milica, V. (2023). The Impact of Public Events on the Use of Space: Analysis of the Manifestations in Liberty Square in Novi Sad. *Arhiv za tehničke nauke, 2*(29), 75-82.

[33] Vlachidis, A., & Tudhope, D. (2016). A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. *Journal of the association for information science and technology, 67*(5), 1138-1152.

[34] Warburton, K. (2023). *Terminology management Routledge encyclopedia of translation technology*. Routledge, 750-765.

[35] Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with applications*, *42*(4), 2264-2275.
[36] Wei, Y., & Ding, Y. (2023). Application of Text Rank Algorithm Fused with LDA in Information Extraction Model. *IEEE Access*.
[37] Yusuf, N., Yunus, M.A.M., & Wahid, N. (2019). Query expansion based on explicit-relevant feedback and synonyms for English Quran translation information retrieval. *International Journal of Advanced Computer Science and Applications*, *10*(5), 227-234.
[38] Zamani, H., Dumais, S., Craswell, N., Bennett, P., & Lueck, G. (2020). Generating clarifying questions for information retrieval. *In Proceedings of the web conference*, 418-428.
[39] Zhang, M., Li, X., Yue, S., & Yang, L. (2020). An empirical study of TextRank for keyword extraction. *IEEE access*, *8*, 178849-178858.
[40] Zhou, N., Shi, W., Liang, R., & Zhong, N. (2022). Textrank keyword extraction algorithm using word vector clustering based on rough data-deduction. *Computational Intelligence and Neuroscience*, *2022*.
[41] Zuo, X., Zhang, S., & Xia, J. (2017). The enhancement of TextRank algorithm by using word2vec and its application on topic extraction. *In Journal of Physics: conference series*, *887*(1). https://doi.org/10.1088/1742-6596/887/1/012028

Authors Biography

Dr. Wafa' Za'al Alma'aitah received his B.S. degree in software engineering from Hashemite University, Zarqa, Jordan, in 2005. She continued his education and obtained an MSc. Degree in Computer science from the Al Albayet University, Mafraq Jordan in 2010. Dr. **AlMa'aitah** received her PhD from the school of computer sciences, Universiti Sains Malaysia (USM), Malaysia, in 2020. From 2011 to 2022. She served as a lecturer at the Hashemite university. Since October 2022, she has held the position of assistant professor in the Department of Intelligence Systems, Faculty of Artificial Intelligence, at Al-Balqa Applied University, Jordan

Dr. Fatima N. AL-Aswadi is an assistant professor in the Institute of Computer Science and Digital Innovation, UCSI University, Malaysia. She received her PhD from the School of Computer Sciences, Universiti Sains Malaysia (USM), Malaysia, in 2023. She obtained her M.Sc. degree in Computer Sciences from King AbdulAziz University in Jeddah, Saudi Arabia, in 2014 and her B.Sc. degree in Computer Sciences from Hodeidah University, Yemen, in 2005. Her areas of specialization encompass ontology learning, deep learning, machine learning, NLP, knowledge mining, and knowledge engineering. Her research interests span both foundational and applied aspects, including text representation, concept extraction, relation discovery, information retrieval, sentiment analysis, data mining, ontology, knowledge graphs, artificial neural networks, classification, and intelligent systems.

Dr. Rami S. Alkhawaldeh received his B.S. degree in Computer Information Systems from Yarmouk University, Irbid, Jordan, in 2007. He continued his education and obtained an MSc. Degree in Computer Information Systems from the University of Jordan, Amman, Jordan, in 2010. Dr. Alkhawaldeh further pursued his academic journey and successfully completed his PhD degree in computing science from Glasgow University, the UK, in 2017. From 2010 to 2012, Dr. Alkhawaldeh served as a Lecturer at the University of Jordan, where he contributed to the education and development of students in the field of Computer Information Systems. Since February 2021, he has held the position of associate professor in the Computer Information Systems Department at the University of Jordan, further enriching the academic environment and sharing his expertise with students and colleagues.