INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Identifying Terms to Represent Concepts of a Work Process Ontology

Mohamad Amin Osman[1], Shahrul Azman Mohd Noah[1]*, Saidah Saad[1]

1 Center for Artificial Intelligence Technology, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Malaysia

OPEN ACCESS

Received: 27-06-2023

Accepted: 01-04-2024

Published: 13-06-2024

Citation: Osman MA, Noah SAM, Saad S (2024) Identifying Terms to Represent Concepts of a Work Process Ontology . Indian Journal of Science and Technology 17(24): 2519-2528. https://doi.org/10.17485/IJST/v17i24.1597

* **Corresponding author**.

shahrul@ukm.edu.my

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To identify ontology concepts from text documents for the construction of work process ontology. **Methods:** This study proposes a methodology to identify terms representing a work process ontology concept from a document. The methodology encompasses several key steps: document collecting, text pre-processing, term weighting and analysis, terms mapping, and domain expert relevance judgments. A comparison between the results of three different term weighting schemes, namely the Term Frequency (TF), Term Frequency-Inverse Document Frequency (TFIDF), and Mutual Information (MI) is made with the ontology concept that the domain expert has judged. **Findings:** The approaches adopted in this study have managed to extract ontological concepts from the targeted domain knowledge source. The findings of the comparison analysis suggest that the TFIDF term weighting scheme exhibits better results compared to the TF and MI weighting schemes. **Novelty:** A work process ontology is a structured knowledge describing daily operations in the government sector. However, there has been little to no effort in building the work process ontology. This study presents an integrated approach for identifying ontology concepts from documents within the domain of the work process. To the utmost extent of our understanding, this research initiative is the initial attempt to introduce a structured methodology for the semi-automatic extraction and evaluation of concepts and relationships within this domain. The findings can be utilised as a foundation for developing an ontology in the specific field.

Keywords: Ontology; Work process; Text extraction; Natural language processing; Term weighting

1 Introduction

From the perspective of information science, ontology is a representation technique of some domains by means of concepts and the relationships between those concepts. By defining a set of concepts and categories that represent the subject, ontology is a way to demonstrate the characteristics of a subject area and how they relate to one another. Konopka & Smedley [(1)] defined ontology as:

- it is a knowledge representation tool that conceptualises, systematises, and organises knowledge, and
- has been used to represent domain-specific knowledge.

Ontologies have been, and are being, used to solve data integration problems by providing a common, agreed-upon vocabulary. Over time, ontologies have been used in many other applications to solve complex problems. For example, the Reachability Matrix Ontology (RMO)[2] has been developed in the network and cyber security domain to compute the reachability matrix by describing the network's element, network connectivity information, and access control policies. In geology, the GeoCore ontology[3] was used to define general concepts within the geology field to facilitate communication between geology users and the domain applications' integration. Ontology was also used in medicine, such as the Onto Pharma[4], which reduces prescribing errors by providing an alert when a medication error is identified. In disaster management, an ontological approach has been applied to automate the management process by helping with task distribution among relevant authorities at different stages of a crisis. It outlines a knowledge-driven decision support system for providing financial relief to victims[5]. Ontologies can also be used to manage and exchange information as suggested by B. Ben Mahria et al.[6], who transform a relational database into an ontology to provide applications based on the semantic representation of the data.

The need for ontologies has increased over time for various applications. Hence, there is an increasing demand for efficient and effective techniques to build ontologies[6]. Building an ontology is a complicated task that can be grouped into three categories: manual, automatic, and semi-automatic. Manual approaches are typically inefficient, expensive, and subjective as they require the involvement of an expert group comprising ontology and knowledge engineers[7]. In comparison to the manual approach, automatic ontology building has several benefits, including more complexity, quicker completion, and less reliance on expert knowledge[8]. On the other hand, semi-automatic systems maintain the correctness and quality of the created ontology by balancing automated methods with human expertise[8],[9].

An ontology can be constructed from various sources. Domain-specific articles, existing ontologies, domain knowledge bases, terminological databases, dictionaries, lexical resources, HTML/XML documents, and text collections are among the primary sources for ontology learning[10],[11]. Among the sources, textual documents are the important sources for building an ontology due to their rich knowledge. However, extracting knowledge from these sources is difficult due to their unstructured nature. As such, most of the approaches that use textual documents employ a semi-automatic approach.

The daily operations in the government sector primarily focus on delivering services to customers, who are the community members who deal with the respective agency. A structured and effective work process must be integrated across the entire service delivery process to produce exceptional service. Among the measures deemed capable of improving the level of service is the implementation of knowledge sharing among the organisation's members. Knowledge sharing in the work process domain can significantly enhance the organisation's daily operations.

Ontology is one of the most suitable approaches for assisting knowledge-sharing initiatives. The work process ontology will include all relevant concepts as well as the relationships between concepts in this domain. In addition, ontologies can give people a shared knowledge of the subject matter, in our case, the execution of work processes within an organisation. These unique features may facilitate efficient communication among members of the organisation and provide a platform for inferring new knowledge through the development of queries.

Government agencies own a substantial array of reference documents that serve as guidance for organisational members in the execution of their daily duties. These documents define the procedural framework, defining and elaborating upon the steps required to execute a given task. The work process can be conceptualised as a sequential set of procedures or actions undertaken to complete a task or achieve a certain purpose. The process of manually extracting the key information from the document proved to be difficult due to the large number and variety of available documents. Hence, text extraction can be employed as a method to extract the main concepts from the given text. Text extraction involves obtaining reduced expressive contents from text documents[12]. Text extraction can be carried out using numerous techniques and tools[13]; one of them is the Natural Language Processing (NLP) technique. NLP enables computers to process and analyse massive volumes of natural language data to achieve human-like language processing for various tasks and applications[14],[15].

Previous research has investigated the possibility of using text extraction methods as a means of identifying an ontology concept from textual documents. Research by M. Javed et al.,[16] proposes the role of the NLP method for extracting data from standard documents and using ontologies to translate that data to semantics. Both syntactic features captured by NLP tasks (POS tags, phrasal tags, etc.) and semantic aspects (ontology concepts and relations) are stored in the rules for information extraction that are created using regular expressions. Utilising NLP allows for a more objective and consistent interpretation of regulatory regulations while reducing human intervention in document processing and information extraction.

B. Fawei et al.[17] have developed a semi-automated mechanism for generating legal ontologies, which is intended to facilitate legal question answering. The tool utilises a building model that incorporates NLP and human intervention to extract structured information from criminal law and legal procedures. This is achieved by utilising source material from exam preparation

material, domain experts, and bar exam questions. The methodology encompasses a series of 15 sequential processes, including the production and analysis of competency questions and the use of NLP tools to extract textual information. The methodology also incorporates the utilisation of OWL and SWRL to represent legal concepts and norms and derive logical conclusions from the ontology.

Another study that utilises NLP in its methodology is the research conducted by A. Ayadi et al. [18]. This study employs a deep learning-based NLP technique to enhance the Biomolecular Network Ontology (BNO) by identifying, extracting, classifying, and integrating novel concepts and specialised relationships from biological literature. The methodology presented has three primary stages: data collecting, knowledge extraction, and ontology population. The initial stage of data acquisition entails the exploration of online documents and local files that pertain to the specific domain of the ontology. Additionally, it comprises the preparation of biological documents to be processed in the subsequent phase, known as the pre-processing phase. The knowledge extraction phase aims to identify valuable knowledge from the biological documents provided as input. This involves extracting potential instances of concepts, relations, and attributes. To accomplish this task effectively, the BNO ontology and deep learning-based NLP techniques are employed together. The objective of the ontology population phase is to assess the redundancy and consistency of the extracted instances with respect to the pre-existing knowledge contained in the BNO ontology. Subsequently, the filtered instances are transferred to populate the BNO ontology.

Y. Aleman [19] conducted a study with the aim of proposing a semi-automatic framework for the generation of pedagogical domain ontologies using techniques from NLP and information retrieval. The process comprises three primary parts: resource compilation, ontology construction, and evaluation. During the phase of resource compilation, the focus is on collecting and examining the relevant resources. The available resources encompass a range of scholarly articles, books, and other relevant publications related to the specific subdomain within the field of pedagogy under investigation. Subsequently, the gathered resources undergo pre-processing procedures, including tokenisation, stop-word removal, and stemming approaches. Following the pre-processing stage, the resources undergo analysis utilising NLP methodologies, including part-of-speech tagging, named entity recognition, and dependency parsing. During the ontology creation phase, the ontologies that correspond to each class are developed using a semi-automatic process. The method has seven general steps, which can be categorised into two main sections: the examination of the theoretical components and the subsequent organisation. The selection of the theoretical approach for each class is determined by the quantity of information accessible for analysis. The process of organising the ontologies is then carried out on a class-by-class basis. In the evaluation phase, the ontologies that have been developed are assessed and improved to ensure their precision and comprehensiveness.

M. Gomez Suta et al. [20] suggested a semi-automated approach for converting Spanish texts into ontology structures involving identifying terms, concepts, and their relationships. This process also includes the participation of human experts in validating the identified terms. The approach employed in this study consists of two distinct stages for the ontology learning process. These stages encompass the development of vocabulary through the identification and validation of concepts and relations. During the phase of vocabulary development, the authors employed four statistical weighting techniques to construct a lexicon consisting of relevant terms that accurately depicted the corpus. The authors additionally employed named entity recognition to aid in the detection of collocations. At the extraction and validation stage, the authors employed the Directed Louvain method at the third hierarchical level to extract and validate abstract ontological structures. The validation task utilised resources created by humans, indicating that experts were engaged in the process of validating the extracted phrases.

Table 1. Existing Approaches for Building Domain Ontologies Using Text Extraction Methods

Authors	Domain	Language	Sources	Term-Weighting
M. Javed et al., [16]	Space System Software Engineering Processes	English	System Software Engineering Standards	
B. Fawei et al. [17]	Criminal Law and Legal Procedures	English	Collection of Legal Documents	
A. Ayadi et al. [18]	Biomolecular Network	English	Biological Documents	
Y. Aleman [19]	Pedagogy	Spanish	Articles Related to the Types of Intelligences, Learning Strategies, and Learning Styles	
M. Gomez Suta et al. [20]	Colombian Armed Conflict	Spanish	Domain Related Online Articles	TF-IDF, TF-Entropy

Table 1 summarizes existing approaches for building domain ontologies using text extraction methods. It was evidence that the researchers have worked on various domains with a reasonable amount of textual content. Integrating NLP techniques with

ontological concept identification methods, such as term weighting, is one of the options. In general, it may be noted that the use of combination methods of text extraction, term weighting and analysis, and verification by domain experts remains very limited.

J. Watrobski[11] asserts that ontology learning from text is difficult due to its arduous and time-intensive nature, typically requiring human analysis of textual documents. The process of creating and maintaining ontology typically involves manual intervention and is commonly carried out by multiple domain experts. In the realm of scientific research, it is common for domain experts to focus on distinct areas of study, leading them to develop unique conceptualisations of their knowledge[21]. Consequently, our methodology incorporates the addition of domain expert decisions inside our methodological framework.

Through the review, it was also discovered that the creation of an ontology for the work process domain has received little to no attention. Work process procedures in government-published documents can be highly domain-specific and thus pose challenges in extracting them. The documents may also contain a mix of Malay and English languages. This study aims to address the issues by proposing an integrated method that combines three distinct term weighting schemes, namely, Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), and Mutual Information (MI).

Apart from that, it becomes evident that the utilisation of Malay text document extraction for ontology concept identification remains rather limited in its application. This occurrence is anticipated due to the predominant emphasis of current extraction methods on English text. The Malay language exhibits limitations in the field of NLP and has been classified as an under-resourced language[22]. This technology might not develop well for Malay text[23] as most of the NLP resources are mainly available in English[24]. As almost all documents related to work processes are in Malay, this approach will be challenging to execute. Furthermore, extensive research on text extraction techniques for documents containing a mixture of Malay and English languages is lacking. Hence, this research paper presents a novel approach for extracting textual information to identify phrases that indicate concepts related to work processes in government documents that combine Malay and English.

This paper aims to use text extraction techniques to extract the concept of the work process from relevant documents obtained from public sector agencies in Malaysia. The method being proposed comprises a series of sequential steps. The process commences with collecting documents, followed by text pre-processing, tokenisation, and the use of term weighting and analysis techniques. This research employed the Python programming language, specifically utilising the Natural Language Toolkit (NLTK) and a specialised library designed for Malay text processing called Malaya NLTK[25].

2 Methodology

Different methods have been used to extract ontology concepts from textual documents. These documents typically contain a wealth of useful information that may be utilised to build ontologies. In the public sector organisation, various documents have been published to provide information and guidance in the operation of services in government agencies. As a result, a set of procedures must be developed to enable the extraction of information and identify the main concepts to be used for ontology construction.

This article proposed the text extraction method as an improved approach to tackle long-standing challenges faced by current methods. These challenges encompass the substantial quantity of documents, which creates difficulties for manual text extraction and the formal language commonly applied in official government documents. The suggested approach consists of several stages to address these problems, as illustrated in Figure 1. The procedure starts with collecting documents, followed by text pre-processing, assigning weights to terms, mapping the terms, and ultimately obtaining relevant judgments from domain experts. The Python Programming language was utilised as a tool to automate text pre-processing and term weighting and analysis operations, resulting in a streamlined procedure and improved efficiency.

By integrating various stages, the proposed method aims to streamline text extraction from documents while addressing key issues such as document volume and language complexities. The sequential approach ensures a systematic and comprehensive extraction process. Automation using Python for text pre-processing and term weighting expedites the workflow and enhances accuracy and consistency in handling diverse document types, thus offering a robust solution to text extraction challenges.

2.1 Document Collection

The first step is to create a corpus. It is a collection of texts or spoken language data that can be created for a variety of purposes, including studying language patterns, analysing discourse, investigating language variation, and developing language resources[26]. For the development of this corpus, the document's source must be identified. Generally, one can access government-published documents online via the respective agency's website. This method is the fastest method to collect documents involved in the text extraction, which will be carried out later. There are, however, some circumstances in which online searches cannot be used to find the necessary papers. To access the required documents, researchers must contact the

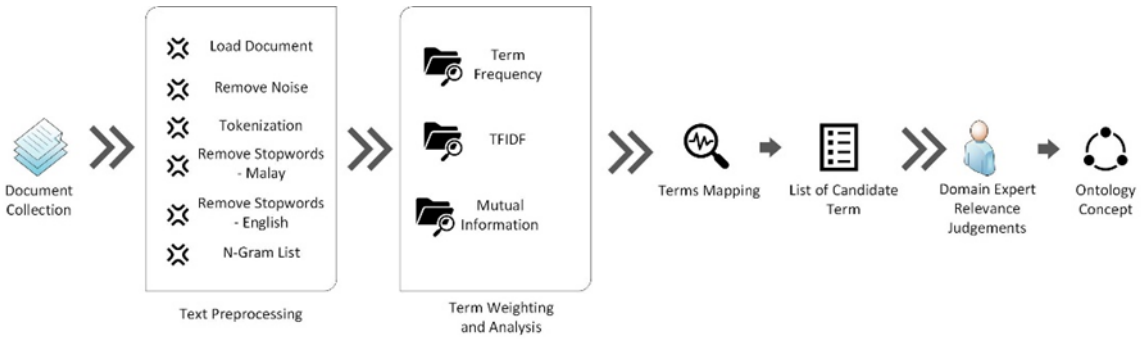

Fig 1. The framework of the Proposed Method

responsible persons in that agency. For this research, the entire documents were found via a search on the relevant government agencies' websites. A copy of the document in the form of a soft copy is collected and checked to ensure that all pages are available and complete for the preparation of the following process.

2.2 Text Pre-processing

Before the text can be processed, a corpus of documents or collection must undergo some preparation steps. Text pre-processing is the process of converting unstructured data into structured data by applying specific criteria or rules to facilitate future keyword extraction [27]. Some common pre-processing steps include tokenisation, lemmatization, lowercasing, stopwords removal, punctuation mark removal, and reduction of replicated characters [28]. Since Python NLTK is well-known for its user-friendly interface and ability to efficiently process huge amounts of text data [29], it was utilised to facilitate text pre-processing. All documents that have been collected will be loaded so that they can be read and arranged into a list. Then the list will go through a noise removal process in which, in this step, all unwanted characters, digits, and pieces that can interfere with the analysis are removed.

At this point, one of the initial challenges that must be overcome is the implementation of the text pre-processing procedure. This is related to the fact that most of the document contents are written in Malay. Considering this situation, a specialised Python library, Malaya NLTK [25], was used for the tokenisation process. Aside from tokenisation, Malaya NLTK provides many capabilities, such as creating a knowledge graph, sentiment analysis, and knowledge extraction. The tokenisation process will tokenise the sentences into individual words or phrases. The following step is the removal of stopwords. In this study, we use both Malay and English stopword lists, considering that some document corpus was published in Malay and English mixed language. Finally, the n-gram token will be generated, and we consider the unigram, bigram, and trigram as the result of the extracted term. This is based on a recommendation by L. Hickman et al. [30] who recommend that researchers carry out the tokenising procedure up to trigrams to boost the text's validity by gathering some semantic information.

2.3 Term Weighting and Analysis

Term Weighting is an approach for assigning appropriate weights to terms for the weights to represent the relevance of terms in a document [31]. In this study, we implemented three different term weighting approaches, being that these approaches might suggest different important terms. On the other hand, we would also want to find terms that are unanimously agreed upon by all three approaches to be important or relevant to represent ontology's concepts of the work process. The three approaches are Term Frequency (TF), Term Frequency-Inverse Document Frequency (TFIDF), and Mutual Information (MI).

Term frequency represents how many times a word appears in a document [32]. To calculate the term frequency, we must first calculate the number of times the term (t) appears in document (d) and the total number of terms in document (d). The equation for term frequency is as below:

$$TF_{t,d} = \frac{\#\ of\ times\ term\ t\ appear\ in\ document\ d}{\#\ of\ terms\ in\ document\ d}$$

TFIDF is a statistical method to evaluate the significance of a word in a document or corpus [33]. Terms with higher TF-IDF scores are more relevant and can be utilised in tasks such as keyword extraction, document ranking, and information retrieval [34]. As

for TFIDF, the value of IDF must be calculated first before the final calculation can be executed. The following is the equation for IDF and TFIDF:

$$IDF_t = log\frac{N}{n_t}$$

- N – Total number of documents in the corpus
- n_t – Total number of documents that have the term t

Thus, $TFIDF_{t,d} = TF_{t,d} \times IDF_t$

MI can be used to measure the dependency value of two variables (words) [35] the higher MI might show a strong semantic relationship and a high possibility that the correlation between X and Y is to be an ontology concept [36], [37]. The equation for the MI of two events, X and Y are as follows:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

where, $H(X,Y) = -\sum p(x,y)log(p(x,y))$and, $MI(w_1,w_2) = log\frac{p(w_1,w_2)}{p(w_1)\times p(w_2)}$

$p(w_1)$ – the probability that w_1 is concentrated in the whole training text

$p(w_2)$ – the probability that w_2 is concentrated in the whole training text

3 Results and Discussion

3.1 Terms Mapping

The previous step produces three lists of terms corresponding to the three term weighting approaches for each n-gram. It is necessary to reconcile and combine the extracted terms into a single, integrated list. To do that, a threshold point must be established to provide a cut-off point for term selection for each term weight. This filtering step needs to be applied as thousands of extracted terms exist for each of them.

There are different thresholds to consider for each set of documents and it can be set freely [38]. In the case of our study, once the term weights are nearly constant, it is when the threshold is reached. (i.e., the difference between the term's weight is close to 0). Figure 2 shows an example of how the threshold point was set for normalized TF on the unigram list, whereby the threshold is set to 0.02. The same procedure was applied to other term weighting. Table 2 shows the threshold values for all the terms-weighting used in this research.

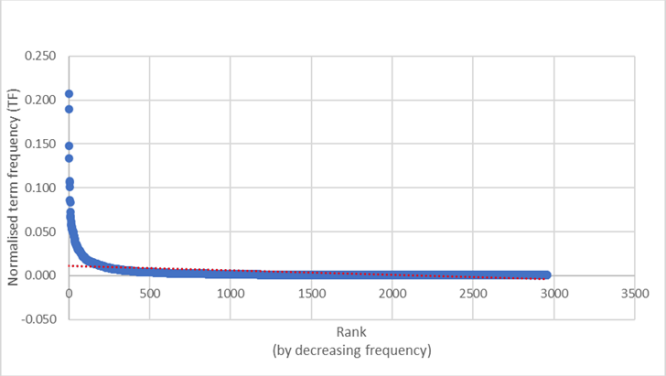

Fig 2. Example of Threshold Point Setting

Figure 3 shows the number of terms extracted from each weighting approach. Figure 4, on the other hand, shows the number of terms after all the results of the three weighting approaches have been mapped (i.e., unique terms and redundant terms were removed). For instance, the initial term extracted for each term weighting method is 2959 tokens for unigram. Filtering based on the threshold values of Table 2 produced 89, 111, and 87 tokens for TF, TFIDF, and MI, respectively. For Bigram, the initial extracted terms were 17296, and a total of 50, 46, and 53 tokens have been filtered for the TF, TFIDF, and MI approaches, respectively. Lastly, for the trigram, out of the 23903 terms extracted, 25, 27, and 24 tokens were filtered for each weighting approach.

Table 2. Threshold Values for Each Term-Weighting Approach

	Unigram	Bigram	Trigram
TF	0.0200	0.0001	0.00004
TFIDF	0.0100	0.0001	0.00020
MI	0.0010	0.0001	0.00020

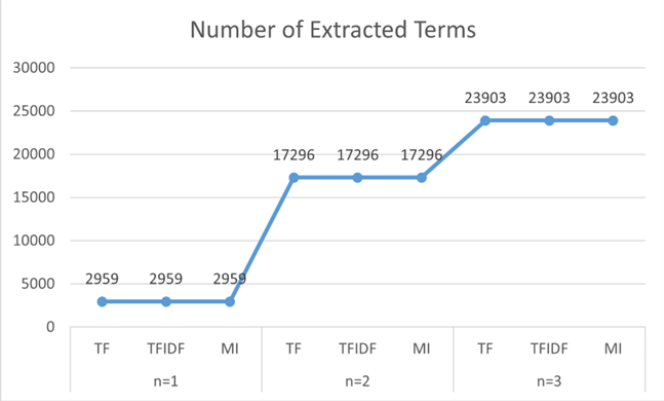

Fig 3. Extracted Term for Each n-Gram and Term Weighting

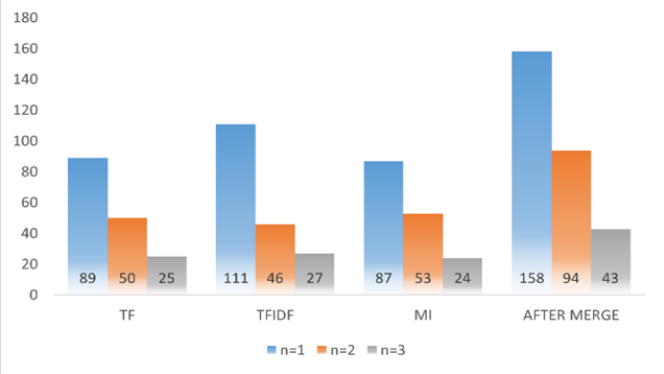

Fig 4. Number of Terms After Mapping

Overall, we can observe that based on the threshold values, only an average of 3.2% of the terms extracted were selected to be candidates of the ontological concepts, whereas for the bigram and trigram, only 0.3% and 0.1% were considered, respectively. The filtered list for each term weight was then compared, and duplicate terms were removed, leaving a single list for each n-gram term. The list for each n-gram was then combined, and the merged list consisted of 295 terms. The list contains terms that are potential candidate concepts for the work process ontology domain.

3.2 Domain Expert Relevance Judgments

There are several approaches to select ontology concepts from a list of candidate terms. Using domain experts is one of the most commonly use strategies. Domain experts play an important role in concept identification by identifying key concepts related to the domain [39]. Domain experts who are able to comprehend and analyse the semantics contained in texts, tables, and images of articles and their experimental parts are needed as computers need fine-grained metadata annotations to do so [40]. In this study, domain experts were engaged to select the most relevant concepts from a list of potential terms extracted from the previous processes. The domain experts that we chose are the personnel attached to public sector agencies with more than five years of experience with organisational work processes.

During this step, the domain experts were provided with a list of 295 terms. An explanation was provided to each of the domain experts, including the objective of the research and the things that they need to do. The domain experts were given time to thoroughly review the list and select only the pertinent terms related to the work process concept, with no limit on the number of terms they could choose.

The list of selected concepts was then analysed and compared, and only a term that all five domain experts agreed on was retained as the final list. The findings of the analysis show that 56 terms can be used to represent work process domain ontology concepts.

3.3 Discussion

Several steps or processes are required to identify terms from documents that represent the work process ontology concepts. The first task is to collect documents to create a document corpus. The next step is to perform a text pre-processing procedure on the documents to convert the entire text into tokens of words that can be analysed. The term weighting and analysis procedure are then carried out to produce a single list that will be used as the candidate term list. The next phase requires the involvement of a domain expert to select an ontology concept related to the ontology domain to be built.

During the pre-processing step, the number of terms that were successfully extracted for each n-gram varies significantly. The number of terms for a unigram is around 2000, whereas the bigram and trigram produced 17,000 and 23,000 terms, respectively. For the filtering based on the threshold, it is observed that the number of terms above the threshold point value decreases for every n-gram. This suggests that the significant differences in terms of weight value between the terms will decrease with each increment in the number of n-grams.

At the final stage, the total number of terms obtained after all the procedures were completed was 56. These terms were all chosen by the five participating domain experts. Table 3 shows the number of terms selected as work process ontology concepts according to term-weighting methods. The results demonstrate that the TFIDF term weighting scheme outperforms the TF weighting scheme by a small margin.

Table 3. The Number of Terms (With Percentages) Selected by Domain Experts (DE) for Each Weighting Method

Weighting Method	# of terms selected by DEs	% (compared to the overall selected terms)
TF	40	71
TFIDF	42	75
MI	28	50

On the application side, the framework used in this study, which was to identify work process ontology concepts, can be reused to identify ontology concepts from text-based knowledge sources. This is especially true for documents published in Malay or documents containing mixed content of Malay and English. The findings or results of this study can also be employed to identify ontology concepts in other knowledge domains, which is an additional benefit of the study. Researchers who wish to utilise the approach in ontology concept identification can also use the various word weighting methods employed in this study as a reference.

4 Conclusion

One of the crucial processes in identifying meaningful terms from a corpus is text extraction and term weighting. In this study, the corpus refers to a government-published document associated with the work process knowledge domain. Most scholars have studied and suggested a method of extracting, considering the various languages and knowledge domains available. For government-published documents, the content is usually prepared in their native language. Since most NLP techniques and resources are meant for English, this condition occasionally causes an issue for non-English documents. With some adjustments, this study performed text extraction on a mix of Malay and English text documents and finalized the candidate terms for the ontology concept by executing the extraction method and term weighting and analysis. To the utmost extent of our understanding, this research initiative is the initial attempt to introduce a structured methodology for the semi-automatic extraction and evaluation of concepts and relationships within the work process domain. Future research may investigate the viability of fully automated term extraction on Malay text documents. Furthermore, the chosen term weighting scheme used in this study is not claimed to be the best method. Approaches from the aspect of machine learning may be a fruitful area for future research.

References

1) Konopka T, Smedley D. A pan-ontology view of machine-derived knowledge representations and feedback mechanisms for curation. *bioRxiv preprint*. 2021;p. 1–29. Available from: https://doi.org/10.1101/2021.03.02.433532.
2) Scarpato N, Cilia ND, Romano M. Reachability Matrix Ontology: A Cybersecurity Ontology. *Applied Artificial Intelligence*. 2019;33(7):643–655. Available from: https://dx.doi.org/10.1080/08839514.2019.1592344.
3) Garcia LF, Abel M, Perrin M, dos Santos Alvarenga R. The GeoCore ontology: A core ontology for general use in Geology. *Computers & Geosciences*. 2020;135. Available from: https://dx.doi.org/10.1016/j.cageo.2019.104387.
4) Calvo-Cidoncha E, Camacho-Hernando C, Feu F, Pastor-Duran X, Codina-Jané C, Lozano-Rubí R. OntoPharma: ontology based clinical decision support system to reduce medication prescribing errors. *BMC Medical Informatics and Decision Making*. 2022;22(1):1–12. Available from: https://dx.doi.org/10.1186/s12911-022-01979-3.
5) Shukla D, Azad HK, Abhishek K, Shitharth S. Disaster management ontology- an ontological approach to disaster management automation. *Scientific Reports*. 2023;13(1):1–15. Available from: https://doi.org/10.1038/s41598-023-34874-6.
6) Mahria BB, Chaker I, Zahi A. A novel approach for learning ontology from relational database: from the construction to the evaluation. *Journal of Big Data*. 2021;8(1):1–22. Available from: https://dx.doi.org/10.1186/s40537-021-00412-2.
7) Yun W, Zhang X, Li Z, Liu H, Han M. Knowledge modeling: A survey of processes and techniques. *International Journal of Intelligent Systems*. 2021;36(4):1686–1720. Available from: https://dx.doi.org/10.1002/int.22357.
8) Zulkipli ZZ, Maskat R, Teo NHI. A systematic literature review of automatic ontology construction. *Indonesian Journal of Electrical Engineering and Computer Science*. 2022;28(2):878–889. Available from: https://dx.doi.org/10.11591/ijeecs.v28.i2.pp878-889.
9) Guimarães NC, De Carvalho CL. A modular framework for ontology learning from text in Portuguese. *Multi-Science Journal*. 2020;3(3):37–42. Available from: https://dx.doi.org/10.33837/msj.v3i3.899.
10) Babič F, Bureš V, Čech P, Husáková M, Mikulecký P, Mls K, et al. Review of Tools for Semantics Extraction: Application in Tsunami Research Domain. *Information*. 2022;13(1):1–30. Available from: https://dx.doi.org/10.3390/info13010004.
11) Watróbski J. Ontology learning methods from text - an extensive knowledge-based approach. *Procedia Computer Science*. 2020;176:3356–3368. Available from: https://doi.org/10.1016/j.procs.2020.09.061.
12) Saeed MY, Awais M, Younas M, Shah MA, Khan A, Uddin MI, et al. An abstractive summarisation technique with variable length keywords as per document diversity. *Computers, Materials & Continua*. 2021;66(3):2409–2423. Available from: https://doi.org/10.32604/cmc.2021.014330.
13) Surana S, Pathak K, Gagnani M, Shrivastava V, Madhuri GS. Text Extraction and Detection from Images using Machine Learning Techniques: A Research Review. In: 2022 International Conference on Electronics and Renewable Systems (ICEARS). IEEE. 2022;p. 1201–1207. Available from: https://doi.org/10.1109/ICEARS53579.2022.9752274.
14) Zhang B, Wang Q. Outfit Helper: A Dialogue-Based System for Solving the Problem of Outfit Matching. *Journal of Computer and Communications*. 2019;7(12):50–65. Available from: https://dx.doi.org/10.4236/jcc.2019.712006.
15) Gomez MJ, Ruiperez-Valiente JA, Clemente FJG. Analyzing Trends and Patterns Across the Educational Technology Communities Using Fontana Framework. *IEEE Access*. 2022;10:35336–35351. Available from: https://dx.doi.org/10.1109/access.2022.3163253.
16) Javed MA, Muram FU, Kanwal S. Ontology-Based Natural Language Processing for Process Compliance Management. In: International Conference on Evaluation of Novel Approaches to Software Engineering;vol. 1556 of Communications in Computer and Information Science. Springer, Cham. 2022;p. 309–327. Available from: https://doi.org/10.1007/978-3-030-96648-5_14.
17) Fawei B, Pan JZ, Kollingbaum M, Wyner AZ. A Semi-automated Ontology Construction for Legal Question Answering. *New Generation Computing*. 2019;37(4):453–478. Available from: https://dx.doi.org/10.1007/s00354-019-00070-2.
18) Ayadi A, Samet A, de Bertrand de Beuvron F, Zanni-Merk C. Ontology population with deep learning-based NLP: a case study on the Biomolecular Network Ontology. *Procedia Computer Science*. 2019;159:572–581. Available from: https://dx.doi.org/10.1016/j.procs.2019.09.212.
19) Alemán Y, Somodevilla MJ, Vilariño D. Semi-Automatic Creation of Ontologies from Unstructured Pedagogical Texts to Assist in Significant Learning. *Computación y Sistemas*. 2022;26(1):245–260. Available from: https://dx.doi.org/10.13053/cys-26-1-4168.
20) Gómez-Suta M, Echeverry-Correa JD, Soto-Mejía JA. Semi-automatic extraction and validation of concepts in ontology learning from texts in Spanish. In: Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics;vol. 1625. New York, NY, United States. Association for Computing Machinery. 2020;p. 7–16. Available from: https://doi.org/10.1145/3405962.3405977.
21) Korel L, Yorsh U, Behr AS, Kockmann N, Holeňa M. Text-to-Ontology Mapping via Natural Language Processing with Application to Search for Relevant Ontologies in Catalysis. *Computers*. 2023;12(1):1–25. Available from: https://dx.doi.org/10.3390/computers12010014.
22) Alexander NS, Yusof A. A Study of Ontology Engineering in Malay Unstructured Document Using Entity Relationship Model. *Journal of Emerging Technologies and Industrial Applications*. 2022;1(2):1–8. Available from: http://jetia.mbot.org.my/index.php/jetia/article/view/17.
23) Fu Y, Lin N, Yang Z, Jiang S. An Open-Source Dataset and A Multi-Task Model for Malay Named Entity Recognition. *arXiv*. 2021;p. 1–12. Available from: https://doi.org/10.48550/arXiv.2109.01293.
24) Awatramani P, Daware R, Chouhan H, Vaswani A, Khedkar S. Sentiment Analysis of Mixed-Case Language using Natural Language Processing. In: 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE. 2021;p. 651–658. Available from: https://doi.org/10.1109/ICIRCA51532.2021.9544554.
25) Husein Z. Natural-Language-Toolkit library for Bahasa Malaysia. 2023. Available from: https://github.com/huseinzol05/malaya.
26) Gui X, Deng L, Jiang G. The Use of Corpus in English Writing: Scholarly Development and Implications. *Journal of Education and Educational Research*. 2022;1(1):64–71. Available from: https://dx.doi.org/10.54097/jeer.v1i1.2476.
27) Chiang TA, Che ZH, Huang YL, Tsai CY. Using an Ontology-Based Neural Network and DEA to Discover Deficiencies of Hotel Services. *International Journal on Semantic Web and Information Systems*. 2022;18(1):1–19. Available from: https://dx.doi.org/10.4018/ijswis.306748.
28) HaCohen-Kerner Y, Miller D, Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE*. 2020;15(5):1–22. Available from: https://dx.doi.org/10.1371/journal.pone.0232525.
29) Sarode V, Joshi B, Savakare T, Warule H. A Real Time Chatbot Using Python. *International Journal for Research in Applied Science and Engineering Technology*. 2023;11(5):7385–7389. Available from: https://dx.doi.org/10.22214/ijraset.2023.53453.
30) Hickman L, Thapa S, Tay L, Cao M, Srinivasan P. Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*. 2022;25(1):114–146. Available from: https://doi.org/10.1177/1094428120971683.

31) Guo B, Zhang C, Liu J, Ma X. Improving text classification with weighted word embeddings via a multi-channel TextCNN model. *Neurocomputing*. 2019;363:366–374. Available from: https://dx.doi.org/10.1016/j.neucom.2019.07.052.
32) Albalawi R, Yeap TH, Benyoucef M. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*. 2020;3:1–14. Available from: https://dx.doi.org/10.3389/frai.2020.00042.
33) Yuan H, Tang Y, Sun W, Liu L. A detection method for android application security based on TF-IDF and machine learning. *PLOS ONE*. 2020;15(9):1–19. Available from: https://dx.doi.org/10.1371/journal.pone.0238694.
34) Patil R, Boit S, Gudivada V, Nandigam J. A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access*. 2023;11:36120–36146. Available from: https://dx.doi.org/10.1109/access.2023.3266377.
35) Sulistiani H, Muludi K, Syarif A. Implementation of Dynamic Mutual Information and Support Vector Machine for Customer Loyalty Classification. In: The 2nd International Conference on Applied Sciences Mathematics and Informatics ;vol. 1338 of Journal of Physics: Conference Series. IOP Publishing. 2019;p. 1–8. Available from: https://dx.doi.org/10.1088/1742-6596/1338/1/012050.
36) Liu Y, Shi M, Li C. Domain ontology concept extraction method based on text. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). IEEE. 2016. Available from: https://doi.org/10.1109/ICIS.2016.7550933.
37) Nguyen TMH, Webb S. Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*. 2017;21(3):298–320. Available from: https://dx.doi.org/10.1177/1362168816639619.
38) Prasetiyowati MI, Maulidevi NU, Surendro K. Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. *Journal of Big Data*. 2021;8(1):1–22. Available from: https://dx.doi.org/10.1186/s40537-021-00472-4.
39) Ramli F, Noah SAM, Kurniawan TB. Using Ontology-Based Approach to Improved Information Retrieval Semantically for Historical Domain. *International Journal on Advanced Science, Engineering and Information Technology*. 2020;10(3):1130–1136. Available from: https://doi.org/10.18517/ijaseit.10.3.10180.
40) Strömert P, Hunold J, Castro A, Neumann S, Koepler O. Ontologies4Chem: the landscape of ontologies in chemistry. *Pure and Applied Chemistry*. 2022;94(6):605–622. Available from: https://doi.org/10.1515/pac-2021-2007.