# Knowledge Organization

Article

Ontology-Based Semantic Search for Physics Subject: A Case Study at a Vietnamese High School

An Cong Tran[1],*, Binh Diep-Phuoc[2], Hai Thanh Nguyen[1]

[1]College of Information Communication and Technology, Can Tho University, 90000 Can Tho, Vietnam
[2]Nguyen Thi Minh Khai High School, 96000 Soc Trang, Vietnam
*Correspondence: tcan@ctu.edu.vn (An Cong Tran)
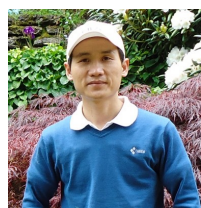
An Cong Tran is a senior lecturer at the College of Information and Communication Technology at Can Tho University, Vietnam. He received a Master's Degree (Hons) in Computer Science from the Asian Institute of Technology (AIT), Thailand, and a Ph.D. in Computer Science from Massey University, New Zealand. His current research interests include description logic learning, ontology learning, applications of blockchain in the public sector, and applications of deep learning methods.

Binh Diep-Phuoc is an Information Technology teacher at Nguyen Thi Minh Khai Gifted High School in Soc Trang Province, Vietnam. He holds a Master's degree in Information Systems from Can Tho University, Vietnam. He finds great interest and professional satisfaction in computing programs, contributing to his ongoing development. His current research interests also focus on ontology building and the Semantic Web.

Hai Thanh Nguyen is an Associate Professor at the College of Information and Communication Technology, Can Tho University, Vietnam. He received his Master's degree in Computer Science and Engineering from National Chiao Tung University, Taiwan, and his Ph.D. degree in Computer Science from Sorbonne University, France. His current research includes healthcare systems and applications using computer vision, machine learning, and data analysis.

Abstract

The implementation of the 2018 Vietnam General Education Curriculum has brought significant changes to teaching and learning methods in Vietnam. To adapt to these changes, teachers and students require information systems that more effectively support information search and promote effective teaching and learning. This study focuses on developing a physics ontology (PhyOntology) and a corresponding semantic search tool (PhyOwlSearch), tailored to textbooks and related educational materials. We leveraged Physics textbooks and the General Physics Dictionary of the Vietnamese Education Publishing House as the basis for proposing the new ontology. The first step was to manually develop an ontology, including the concepts and the fundamental relationships between these concepts (called BaseOntology). The second step was to build a corpus to expand the BaseOntology and then develop it further using the pattern-based approach (called PhyOntology). Subsequently, a semantic search engine (PhyOwlSearch) was developed based on PhyOntology and data collected from virtual sources. PhyOntology offers comprehensive coverage by capturing key concepts and their interrelationships from the General Education program, showing most of the physical connections between these concepts. The ontology-based semantic search engine enables students to access overlapping notions of knowledge (i.e., establishing knowledge relationships in search results). The system uses PhyOntology, creates indexes, and returns accurate search results that meet the demands of self-learning and teaching.

Keywords: high school physics; semantic search; ontology construction; pattern-based

1. Introduction

High school students often use search engines like Google to find learning materials. They enter keywords that pertain to the content they are studying, such as "Newton's 2nd Law", "Nuclear Physics", "Spectroscopy", etc. The results usually direct them to Wikipedia for definitions and search history related to their query. Google's search results can effectively support teaching and learning at high schools by providing quick access to extensive online resources for both teachers and students. However, using various techniques to achieve this, most search engines return results in the form of listings of content that most resemble the content being searched for. When applying this result to teaching and learning, teachers must manually read the found material to extract the desired information. Consequently, using this data source to develop information systems for teaching and learning in high schools is challenging.

In today's high school learning environment, there is a need for an intelligent information system that can support educational innovation-oriented development. This system should be compatible with the new high school curriculum, offering rich data sources, focusing on skill development, improved learning methods, self-study, and promoting students' sense of inquiry, research, and ability to apply knowledge to real-world situations. At the same time, it should be able to adapt to modern teaching methods such as project-based teaching and integrated teaching, which require students to actively find and learn from various documents. These challenges are real issues faced by high schools today.

Based on the analysis of the work of Antoniou and Van Harmelen (2004), building a semantics web-based on ontology can meet the learning and teaching needs described above. First, an ontology for educational domains must be constructed and then gradually completed. A teaching and learning service system based on the developed ontology can then be built. Physics is a subject that requires a close connection between chapters, lessons, and knowledge. Therefore, understanding the relevance of related knowledge from other chapters or lessons is often necessary when learning a concept or phenomenon. For instance, when studying the chapter on motor oscillations in the 12th-grade program, students need to understand solids, vibrations, mechanical motion, velocity, and acceleration taught in grades 10 and 11 as the entire chapter in the high school program. Therefore, searching for related information is more critical in physics than other subjects. Additionally, after completing grade 12, students typically participate in graduation and university entrance exams that often test their knowledge of topics covered in the entire high school program. For these reasons, this work focuses on 12th-grade physics knowledge to illustrate its functions.

This study proposes a semi-automatic approach for building an ontology based on the index of the Physics text-book. The resulting ontology is extended using a pattern-based technique from the "General Physics Dictionary" published by the Vietnamese Education Publishing House. The ontology, along with the reference materials for physics subjects in high schools (some of which were collected from the internet), is then applied to build a semantic search system that better supports the learning and teaching of physics in high schools.

The main contributions of this work can be highlighted as follows:

● First, we collected a high school physics learning material dataset from various sources, including physics textbooks, dictionaries, and online physics documents. Based on the collected dataset, we created a BaseOntology using the Ushold and King method, resulting in an ontology with 14 classes, 46 properties, and 859 instances.

● Second, we developed an extended version of the BaseOntology using the pattern-based method (a semi-automatic approach). There were eight patterns designed that helped to discover 211 concepts, four properties, and 721 instances with an accuracy of approximately 61.5%, outperforming a similar study that achieved about 48.5%.

● Third, we implemented a semantic search engine called PhyOWLSearch that uses PhyOntology to help high school students search for physics-related information. The system also provides essential functionalities to allow users to crawl new data, extend the PhyOntology with new crawled documents, and manage patterns (add, edit, remove), among others.

● Finally, we conducted several experiments to validate the ontology and evaluate the search engine. The results demonstrate the effectiveness of our approach and show potential applications in education and other domains. Additionally, we provide analysis and discussion to clarify our results further. Lastly, it's worth mentioning that these findings have significant implications for education and other fields.

2. Related Work

2.1 Ontology Development

The construction of ontology requires a significant amount of effort from experts in each field of study, and it was developed along with the growth of the semantic web. Therefore, various methods have been researched to approach ontology construction, including manual ontology construction, which involves numerous experts in the field study, as discussed in Elkan and Greiner (1993) and in Uschold (1996). In this construction process, supporting tools such as Protégé (Gennari et al, 2003), ISAVIZ (Kapoor and Sharma, 2010), and APOLLO (Lee et al, 2009) are usually used. Protégé is considered the most suitable tool for manual ontology construction as it provides many extensions and a good rendering engine (Alatrish, 2012).

The pattern-based is a semi-automatic approach to ontology construction that uses patterns analyzed from the

corpus to identify potential candidates for inclusion in the built ontology. Maynard et al (2009) developed a solution to apply the pattern-based method for ontology construction, focusing on two tasks: Ontology extension and knowledge extraction from general ontologies into domain ontologies. To use this method, it is necessary to develop suitable supporting tools to facilitate the construction process. This method has a correct identification rate of 48.50% for subclasses, 47.6% for entities, 48.0% for synonyms, and 22.4% for attributes, as reported in their SPRAT (Semantic Pattern Recognition and Annotation Tool) study.

Maass and Janzen (2009) proposed a method for building an ontology for the ambient environment based on patterns. The study demonstrates that combining systematic methods and various design patterns can be an effective approach to designing ontologies for ambient environments. The study focuses on the primary ontology development processes and facilitates explicit pattern-based modularization and abstraction of scopes. In our study, the pattern-based method is employed to extend the ontology in the Vietnamese language, specifically in general physics, to expand PhyOntology.

Ontology reuse involves using the existing ontologies in part or their entirety. This approach allows for the expansion and enrichment of ontologies by combining them. For instance, Maynard et al (2009) utilized patterns to reuse an ontology in their pattern-based ontology construction. In practical terms, ontology reuse can involve reusing the data or the structure of the ontology. These approaches often reference the structure of existing projects, demonstrating a form of partial reuse of the structure (Chen and Gu, 2018; Cook et al, 2011; Lamy, 2017; Sánchez and Pérez, 2018).

2.2 High School Physics Ontology Development

In physics ontology for high school, authors Chen and Gu (2021) have successfully developed a method for automatically generating a physics ontology for the Chinese language. This work emphasizes the importance of using semantic applications for learning, a growing education trend. The main contribution of this work is the development of two algorithms: T2KG and KG2Onto. These algorithms are responsible for automatically transforming text into a Knowledge Graph (KG) and then transforming the KG into an ontology, respectively. The successful application of these algorithms, based on the triple model $<C, R, A>$, has led to the development of a fully automatic method for creating ontologies. A significant achievement of this work is the representation of the definitions and laws of physics in an ontology. However, the method's drawback is that it can only precisely organize specific document sources into three components as described in the topic. Therefore, further evaluation of the ontology is necessary and should be compared with other methods. In addition, it would be beneficial to explore applications that use ontology to assess the quality of the ontology resulting from this creation process.

Another potential application of ontology in physics education is the development of an automatic question-answering system based on ontology (Abdi et al, 2018). The author of this study proposed a two-stage approach to processing a query (question), which involves first performing semantic analysis to describe the question's meaning fully and then transferring the analyzed content into the system. This work is a valuable contribution to the application of semantic technology in education. We can see the positive contribution of these approaches by examining works that apply semantic technology based on ontology to education at the high school level, especially to the subject of physics in secondary education. Although there have been several studies on ontology building and semantics in the Vietnamese language, more research still needs to be done on teaching and learning physics at the general education level.

In recent years, several studies have examined university students' understanding of key concepts in physics education, such as epistemology, ontology, and axiology (Bouchée et al, 2022; Suprapto, 2021). In addition, Chen and Gu (2018) developed an educational ontology to describe systematic knowledge of different subjects. Furthermore, recent evaluations of ontology-based approaches offer potential opportunities for developing an ontology design that can enhance the teaching and learning of physics (Rahayu et al, 2022). These studies provide valuable insights into the use of ontology in the context of physics education. Drawing on these studies can help us better understand the benefits and challenges of applying ontology in the field of physics education and identify opportunities for further research and development.

2.3 Semantic Search

Semantic search has been of significant interest and research in recent years, with various approaches and techniques being proposed. One popular approach to semantic search is ontology-based search, which uses ontologies to provide a structured representation of knowledge in a particular domain. For example, the studies by Rana and Tyagi (2012) and Parmeshwaran et al (2015) proposed ontology-based search approaches that leverage synonymy relationships of words and ontology-based inferences to improve the accuracy of search results. In addition, Parmeshwaran et al (2015) also proposed a semantic search model that expands the user query into related concepts to enhance search quality further.

Another essential consideration in semantic search is document ranking. The studies by Rana and Tyagi (2012) and Parmeshwaran et al (2015) address this issue, proposing ranking models based on the relationship between concepts in the document. In addition, the study by Faisal et al (2012) proposed a similarity-based ranking model that calculates the cosine similarity between the query and document vectors to determine the relevance of search results.
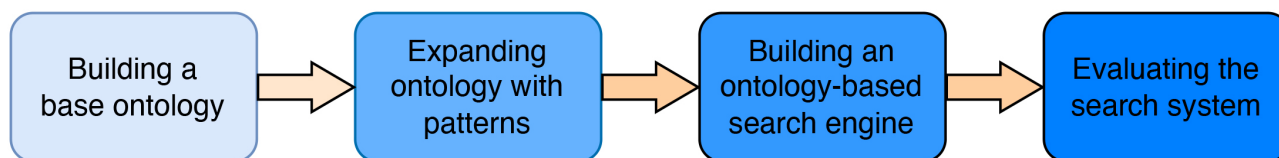
Fig. 1. The proposed workflow for searching terms in physics based on Ontology.

In addition to the above works, there have been several interesting studies on semantic search and annotations in educational activities. For example, Panchal et al (2012) proposed a semantic search system for educational resources that uses machine learning to classify and rank search results based on semantic similarity. Andrade-Hoz et al (2021) developed a semantic annotation tool to support collaborative learning activities. In contrast, Shapovalov et al (2021) proposed a semantic search engine for educational videos that leverages deep learning techniques to improve the accuracy of search results. The study by Bajpai et al (2021) also proposes a semantic search model for online learning resources that considers user preferences and semantic similarity. In contrast, the study by Usta et al (2021) proposed a semantic search system for educational e-books that utilizes natural language processing techniques to enhance search quality.

3. Ontology-Based Semantic Search for High School Physics Subject

Overall, these works demonstrate the diversity and potential of semantic search approaches, from ontology-based methods to deep learning techniques. They highlight the relevance of semantic search to a wide range of applications, including education.

This study aims to create applications for processing data, building and expanding ontology, and developing a semantic search system based on theoretical knowledge. Specifically, we focus on developing PhyOntology, an ontology for the field of general physics, and a semantic search system based on PhyOntology called PhyOwlSearch. Our approach involves four key steps, as illustrated in Fig. 1.

• Step 1: Manually create the basic physics ontology in the Vietnamese language following the method proposed by Uschold and King (1995), which includes creating the structure and core concepts of the ontology. This ontology is named BaseOntology.

• Step 2: Extend the BaseOntology using the pattern-based method, which adds concepts, attributes, and relations to the ontology. The result of this step is an ontology with better coverage, called PhyOntology.

• Step 3: Build a semantic search engine based on the extended ontology called PhyOwlOntology.

• Step 4: Collect data and test the system. The system was implemented using Python, based on the Django and Owlready2 libraries. According to Lamy (2021), Python is a modern language suitable for programming with ontology. On the other hand, Django is an ad hoc solution for building web-based applications based on Python.

Details of the above steps are presented in the following sections.

3.1 Ontology Construction for 12th-Grade Physics Program

In this work, the construction of the physics ontology involved two main steps: the manual creation of the base ontology and the subsequent extension of the base ontology using the pattern-based method. Unlike some other ontology construction approaches that employ ontology reuse, the PhyOntology in our work is created from scratch. This is because no general physics-related ontology could be reused at the time of the research. Therefore, manually creating the base ontology from scratch was necessary before the pattern-based method could be used for extension.

3.1.1 Building the BaseOntology From Vietnamese 12th Physics Textbook

The process for building the BaseOntology is depicted in Fig. 2, which is based on the Ushold and King method. Given that the data source for building the ontology is fixed content from a textbook with limited knowledge, we employ a manual approach to construct the ontology following the steps outlined in this section. First, we digitized the document, followed by word segmentation and term identification. Next, we built and revised the ontology with input from a high school physics teacher. Details of this process are illustrated in Fig. 3. While our approach is time-consuming, it allows for thoroughly examining the domain knowledge and provides opportunities for refining and improving the ontology. We anticipate that our approach will have implications for developing ontologies in other domains with similar characteristics. The following is a detailed description of the BaseOntology development process.

(1) Detecting and cataloging words: The chosen data source is the physics textbook, as the system aims to support high school learning, and the content should be textbook-based. In order to make the content more accessible to students, concepts are named after physics terms listed in the textbook. Through manual analysis of popular physics textbooks, a total of 1528 terms are discovered. Before utilizing these terms for various purposes in the system, they are reviewed and verified by a physics teacher at Nguyen Thi
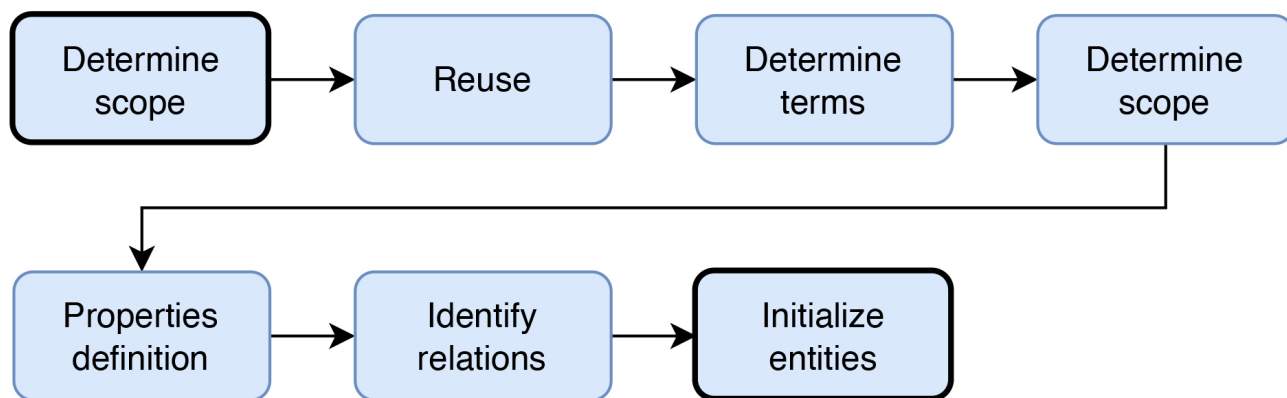
Fig. 2. A traditional ontology-building process (Uschold and King, 1995).
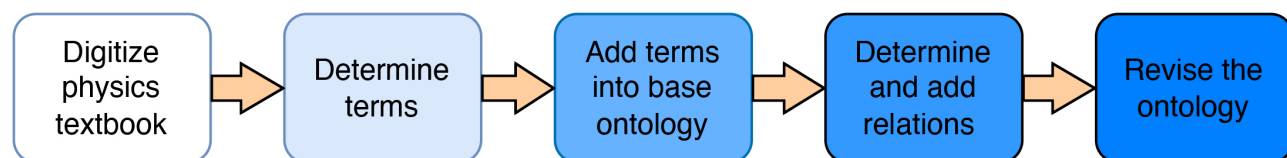

Fig. 3. BaseOntology initialization process diagram.

Minh Khai High School in Soc Trang Province, Vietnam. The core terms are then included in the BaseOntology to facilitate the process of data expansion and to support the word separation algorithm (since the conventional method cannot be used due to the different corpus). Some of the terms in the BaseOntology are listed in Table 1.

(2) Defining the BaseOntology classes and properties: The terms identified in Table 1 are used to define classes and properties in the BaseOntology, as shown in Table 2. Each term is mapped to a corresponding class or property based on its meaning and context in the physics domain. The classes and properties were defined according to the Uschold and King method, which involves identifying the kinds of things in the domain and their relationships.

The base ontology contains Grade, Chapter, Lesson, Concept, Document, and Index classes. In addition, subclasses of the above classes are also identified, such as Phenomenon, Equation, Device, Measurement Unit, Quantity, Law, etc., which are sub-classes of Concept.

Properties represent relationships between concepts in an ontology. They are essential components of an ontology as they represent the relationships between concepts and provide a structured representation of information. There are two types of properties: object property and data property. An object property represents relationships between objects (instances), while a data property represents relationships between objects and specific values. Some properties in the BaseOntology are described in Table 3.

To ensure that the whole meaning of the terms from the textbook is conveyed, it is essential to identify classes and attributes that capture the essence of each term. Entities are grouped into classes and are related to other entities through attributes. The following examples illustrate entities from the class Lesson and Quantity. For instance, consider the lesson "Rotational motion of a solid around a fixed axis" ("Chuyển động quay của vật rắn quanh một trục cố định" in Vietnamese). The content of this lesson is represented using terms and properties defined in the BaseOntology as shown in Table 4, such as title, page number, main content, concepts involved, etc. Another example is the description of the concept (quantity) "angular acceleration" ("Gia tốc góc" in Vietnamese) shown in Table 5. This is a concept from the lesson "Rotational motion of a rigid body about a fixed axis"; its unit is radian/s^2, etc. The above examples show that the developed ontology can basically represent the concepts in physics.

(3) Building the BaseOntology using Protégé: The ontology was built using the software Protégé. First, the identified classes and properties are created (called the ontology TBox). The statistical results of the ontology are shown in Fig. 4. As shown in Fig. 4, the number of classes created is 14, the number of Object properties is 29, and the Data property is 17. Then, the instances are added to the ontology (ABox). The number of instances (individuals) is 292, and the number of relationships created between entities through properties is 567.

Overall, BaseOntology provides a framework for representing the concepts and relationships in the domain of high school physics education. The classes and properties defined in the ontology serve as a foundation for the system's semantic search and data expansion functionalities and future extensions and refinements of the ontology.

IMR Press

Table 1. Some physics terms are extracted from the 12th physics textbook.

No.	Term	No.	Term
1	âm sắc *(timbre)*	21	dao động tự do *(free oscillation)*
2	bảo toàn moment động lượng *(conservation of momentum)*	22	dao động tuần hoàn *(periodic oscillation)*
3	bức xạ hồng ngoại *(infrared radiation)*	23	độ cao âm *(amplitude)*
4	bức xạ sóng điện từ *(electromagnetic radiation)*	24	độ to của âm *(sound level)*
5	bức xạ tử ngoại *(ultraviolet radiation)*	25	gia tốc góc *(angular acceleration)*
6	cảm giác âm *(sound perception)*	26	hệ số công suất *(power factor)*
7	cảm giác nghe *(hearing sensation)*	27	hộp cộng hưởng *(resonance box)*
8	cảm ứng điện từ *(electromagnetic induction)*	28	moment *(moment)*
9	chu kỳ sóng *(wavelength)*	29	moment động lượng *(momentum)*
10	chuyển động quay biến đổi đều *(uniformly varying rotational motion)*	30	moment lực *(moment of force)*
11	chuyển động quay đều *(uniform circular motion)*	31	moment quán tính *(inertial moment)*
12	con lắc dây *(pendulum)*	32	nguồn âm *(sound source)*
13	con lắc đơn *(simple pendulum)*	33	nguồn nhạc âm *(audio source)*
14	con lắc lò xo *(spring oscillator)*	34	nhạc âm *(sound)*
15	con lắc vật lý *(physical pendulum)*	35	sóng âm *(sound wave)*
16	công suất *(power)*	36	tần số góc *(angular frequency)*
17	công suất tức thời *(instantaneous power)*	37	tần số góc riêng *(resonant frequency)*
18	cuộn cảm *(inductor)*	38	tần số sóng *(wave frequency)*
19	cường độ âm *(sound intensity)*	39	tọa độ góc *(polar coordinate)*
20	dao động riêng *(resonance)*	40	tốc độ góc *(angular velocity)*

Table 2. List of some classes implemented in Ontology.

No.	Class	Description
1	Khối *(Grade)*	The grade that the content/concept belongs to. The Vietnam education system has three grades at the high school level: grade 10, grade 11, and grade 12.
2	Chương *(Chapter)*	Chapters in the textbook.
3	Bài_học *(Lesson)*	Lessons in the textbook.
4	Khái_niệm *(Concept)*	The concepts in physics textbooks. They are cataloged into phenomena, Quantities, Fundamental concepts, equations, Measurement unit, Laws, and Device corresponding to sub-classes of this class.
5	Tài_liệu *(Document)*	Documents that are indexed.
6	Chỉ_mục *(Index)*	Indexes of each document that contains related concepts.

Table 3. Some properties in the BaseOntology.

No.	Class	Description
1	Khối *(Grade)*	The grade that the content/concept belongs to. The Vietnam education system has three grades at the high school level: grade 10, grade 11, and grade 12.
2	Chương *(Chapter)*	Chapters in the textbook.
3	Bài_học *(Lesson)*	Lessons in the textbook.
4	Khái_niệm *(Concept)*	The concepts in physics textbooks. They are cataloged into phenomena, Quantities, Fundamental concepts, equations, Measurement unit, Laws, and Device corresponding to sub-classes of this class.
5	Tài_liệu *(Document)*	Documents that are indexed.
6	Chỉ_mục *(Index)*	Indexes of each document that contains related concepts.

3.1.2 Ontology Extension Based on the Pattern

Typically, developing an ontology for a specific domain requires significant time and effort, necessitating the development of automatic or semi-automatic methods to expedite the extension process while ensuring comprehensive ontology coverage. This research employs the semi-automatic approach for ontology extension based on BaseOntology. Specifically, we use a pattern-based approach that emphasizes the relationships between terms and the semantic descriptions in the ontology. The following example illustrates how patterns can be used to identify and establish relationships for ontology expansion. In the sen-

Table 4. Content of the lesson "Rotational motion of a solid around a fixed axis" described in terms of the PhyOntology.

No.	Class	Description
1	Khối *(Grade)*	The grade that the content/concept belongs to. The Vietnam education system has three grades at the high school level: grade 10, grade 11, and grade 12.
2	Chương *(Chapter)*	Chapters in the textbook.
3	Bài_học *(Lesson)*	Lessons in the textbook.
4	Khái_niệm *(Concept)*	The concepts in physics textbooks. They are cataloged into phenomena, Quantities, Fundamental concepts, equations, Measurement unit, Laws, and Device corresponding to sub-classes of this class.
5	Tài_liệu *(Document)*	Documents that are indexed.
6	Chỉ_mục *(Index)*	Indexes of each document that contains related concepts.

Table 5. Description of the quantity "angular acceleration".

No.	Properties	Value
1	belongs_lesson	Rotational motion of a rigid body about a fixed axis
2	has_definition	The angular acceleration of a rigid body about an axis at time t is a quantity that characterizes the change in angular speed at that instant and is determined by the derivative of the angular speed concerning the time
3	has_unit	radian/s^2
4	has_equation	$\gamma = \frac{\Delta\omega}{\Delta t}$

Ontology header:

Ontology IRI http://127.0.0.1:8000/media/owldir/phyontology.owl

Ontology Version IRI http://127.0.0.1:8000/media/owldir/phyontology.owl

Annotations ➕

rdfs:comment
Ontology cơ bản về lĩnh vực Vật lí phổ thông được xây dựng dựa vào các thuật ngữ được trích xuất từ SGK

Ontology imports | Ontology Prefixes | General class axioms

Ontology prefixes:

Prefixes ➕

:
http://127.0.0.1:8000/media/owldir/phyontology.owl#

owl:
http://www.w3.org/2002/07/owl#

Ontology metrics:

Metrics	
Axiom	1,388
Logical axiom count	994
Declaration axioms count	351
Class count	14
Object property count	29
Data property count	17
Individual count	292
Annotation Property count	2

Fig. 4. BaseOntology stats (taken from Protégé's stats screen).

tence "The unit of work is Joule, denoted by J" ("Đơn vị của công là Jun, kí hiệu là J"), for example, it cannot be concluded that "work" is equivalent ("is") to joule. Rather, a relationship between "joule" and "unit" must be established in the context of a physical quantity, where joule is the unit symbol for work.

The initial step in extending the ontology is to select the appropriate data sources. The data source materials for ontology expansion are chosen from the high school teaching materials based on the following criteria: they are rich in knowledge, clearly demonstrate the relationships between terms, and have a uniform presentation style. The ontology expansion process follows the diagram in Fig. 5. Next, pattern analysis is conducted by examining the entire selected document to identify common features in the text presentation that can be used to form patterns in the identification process. These patterns are then tested to determine their effectiveness before being applied to extend the ontology.

Based on the criteria above for selecting data sources for ontology expansion, the terms from Hung and Khiet (2015) are selected as the corpus for the BaseOntology expansion in this study. As shown in Fig. 6, the content of each concept is presented within its respective category, and a consistent presentation method is used for all knowledge.

The content of each document has been digitized and organized into multiple documents, each representing a particular concept, theorem, law, or piece of knowledge.
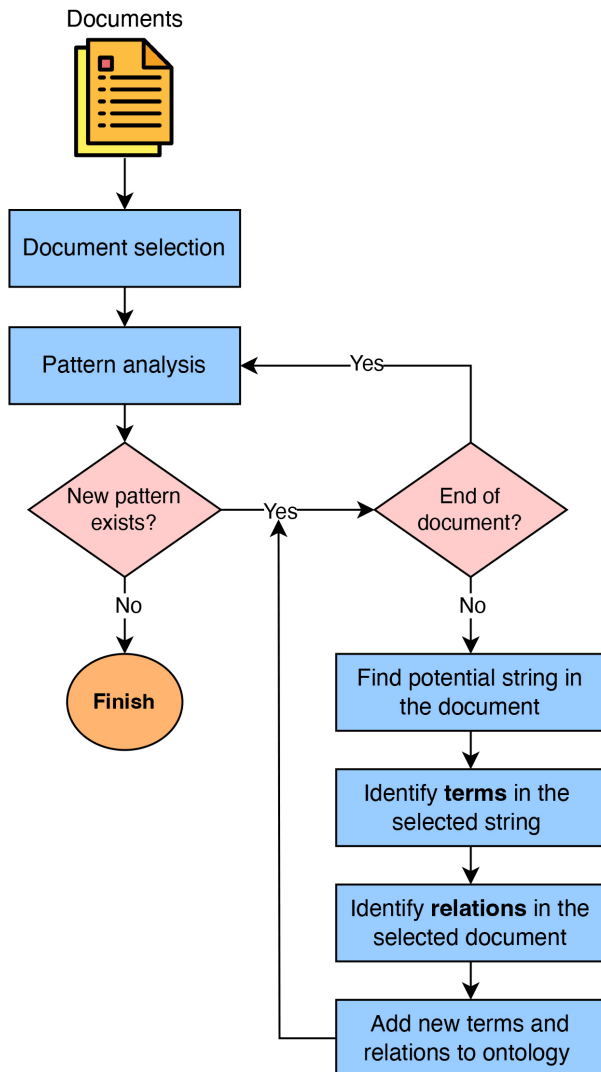

ánh sáng

Sóng điện từ có bước sóng khác nhau. Theo quan điểm hiện đại, chùm ánh sáng còn là chùm các phôtôn (lượng tử ánh sáng).

▶ **Ánh sáng nhìn thấy** (còn gọi là bức xạ khả kiến (mắt người phát hiện được), là các sóng điện từ có bước sóng từ 0,38 μm đến 0,76 μm.

▶ **Ánh sáng đơn sắc** là ánh sáng có bước sóng (hay tần số) xác định và có một màu nhất định phụ thuộc vào bước sóng của nó.

▶ **Ánh sáng trắng** là hỗn hợp của rất nhiều ánh sáng đơn sắc, có màu từ đỏ đến tím.

Ánh sáng có thể gây ra một số biến đổi nhất định ở thực vật.

Ánh sáng Mặt Trời kích thích quá trình hấp thụ canxi của xương, giúp trẻ chống bệnh còi xương.

Chùm ánh sáng trắng phát ra từ đèn chiếu vào lăng kính bị tách ra thành dải sáng liên tục nhiều màu, gồm bảy màu quy ước (đỏ, cam, vàng, lục, lam, chàm, tím).

Fig. 6. Part of the content of the term "light" ("ánh sáng" in Vietnamese) (Hung and Khiet, 2015).

Fig. 5. Ontology extension diagram.

Based on the criteria above for selecting data sources for ontology expansion, the terms from (Hung and Khiet, 2015) are selected as the corpus for the BaseOntology expansion in this study.

As shown in Fig. 6, the content of each concept is presented within its respective category, and a consistent presentation method is used for all knowledge. The content of each document has been digitized and organized into multiple documents, each representing a particular concept, theorem, law, or piece of knowledge.

A pattern in this study refers to a string of characters that conforms to the syntax of the Regular Expression language. Regular expression, also called Regex, is a powerful tool for advanced string processing, using expressions that follow rules. These rules include conventions for groups, wildcards, and other elements. For example, consider the sentence "The unit of work is Joule, denoted by J" (Đơn vị của công là Jun, kí hiệu là J). This sentence defines two terms: "energy" and "Joule", with the symbol value "J". Furthermore, we can establish two relationships between these terms and values using two attributes: "unit of" (đơn vị của) and "denoted as" (ký hiệu là). The regular expression that defines the pattern for the above relations is:

r'(unit of)(.+?)(is)(.+?)(denoted by)(.+?)'

Given an input sentence, the result of the above Regex is a list of 5 items with the following structure:

● result[3] (Concept) is a "unit of" (property) result[1] (Unit), and

● result[1] (Unit) is "denoted by" (property) result[5] (string literal).

For example, applying the above pattern to the sentence "The unit of work is Joule, denoted by J" yields the following result: ["unit of", "work", "is", "Joule", "denoted by", "J"]. Similar analyses are conducted to define other Regex patterns in this study to extend the ontology, as shown in Table 6.

3.1.3 Building the Word Segmentation PhyTokenizer

In Vietnamese, words are the same as in English and are not separated by spaces in other languages. Instead, words are combined into a string, making it difficult for computers to identify individual words. This can lead to problems in natural language processing, such as text classification, sentiment analysis, and information retrieval. Word segmentation or tokenization is the division of a string of characters into individual words or tokens. Many tools, such as Pivy and Underthesea, support word segmentation in Vietnamese. However, they can only segment common words or those from specific domains such as business and banking. Therefore, a special segmentation algorithm must be developed to meet the need for word segmentation in physics for indexing operations.

There are several methods for word segmentation in Vietnamese, including rule-based, dictionary-based, statistical, and machine learning-based approaches. This study proposes the PhyTokenizer segmentation algorithm based

8

Table 6. Patterns for BaseOntology extension.

No.	Type	Pattern	Parameters	Description
1	1	(the formula of) (.3,100?) (:)? ($.+?$)	D|2>has formula is>4	Recognizing concepts' formulas
2	1	(.10,40) (:) (.{100,}) (\n)	D|1>is stated as>3	Recognizing definitions of a concept
3	1	(measure) (.{10,100}) (by) (.+?) ([\,\.\n])	0|4>is used to measure>2	Recognizing measurement devices
4	1	(unit of. ?) (.+?) (is) (.+?) (\, denoted:) (.+)	0,D| 2>unit of >4, 4>denoted>6	Recognizing units and denotation of the physics concepts
5	1	(law.{10,100}?) (:) (.+) (\n) ([^\.;].{5,30})	D|1>is stated as>3	Recognizing the definition of laws of physics
6	2	(is called) (.+?[^\.]) ([();,\.\n])	0|3>called>1	Recognizing statements for a concept
7	2	(.{10,100}) (follow) (.{10,100})	0|S>follow>0	Recognizing concepts that follow a law of physics
8	2	(.{10,100}) (applied) (.{10,100})	0|S>applied>0	Recognizing the applications of concepts

on a dictionary-based segmentation method. Our corpus is too small and does not contain enough data (only 1.500 concepts) to support machine learning-based word segmentation. This segmentation algorithm is based on a dictionary-based segmentation method that tries all words that can occur in the dictionary (a list of physical concepts). To speed up the processing, a binary search method is applied to check the presence of potential concepts in the dictionary. Algorithm 1 presents the main idea of PhyTokenizer, where ViTokenizer is a Vietnamese general-purpose tokenizer (Nguyen and Tran, 2024).

3.2 A Method for Building an Ontology-Based Semantic Search Engine for Learning Materials of Physics Grade 12

To assess the effectiveness and relevance of PhyOntology in the context of physics education at the high school level, this study presents a method for constructing a semantic search engine based on the developed ontology. The efficiency of this search engine is subsequently analyzed to evaluate the quality of the ontology itself. Furthermore, a thorough discussion is undertaken to evaluate the utility of the semantic search engine. Overall, this investigation seeks to provide insight into the potential benefits of utilizing PhyOntology to enhance the teaching and learning of physics in high school settings.

In this work, the construction of a semantic search engine consists of the following tasks:

1. Collecting the documents for the semantic search engine.

2. Calculating the TF-IDF (Term Frequency – Inverse Document Frequency) values for the words in the documents.

3. Creating the index for the search data.

4. Designing the similarity calculation and ranking method.

5. Identifying the output structure of the search result.

Details of these tasks are described in the following sections.

3.2.1 Collecting the Documents

The first step in building the semantic search engine is to collect data (documents) from various sources. This study's resources were obtained mainly from high school physics teachers' materials. Additionally, various sources available on the internet were used to enrich the evaluation data. The collected data were used for both ontology development and search engine construction. The materials were saved in .doc or .pdf formats and sorted by topic, such as mechanics, electricity, etc., to facilitate indexing. Primary sources of the collected materials include the website of physics teachers at Nguyen Thi Minh Khai High School in Soc Trang province, Vietnam (http://thuvienvatly.com), as well as other reputable websites with physics content, such as https://vatlypt.com and https://vatly247.com. These websites are well-known in Vietnam for providing practical physics knowledge for teaching and learning.

Text data from the selected resources were collected and saved to the database using the BeautifulSoup library (https://pypi.org/project/beautifulsoup4/). The Hyper Text Markup Language (HTML) structure of the website was analyzed to identify the tags in which the article's main content is located, along with the associated ID or class. Furthermore, links to other articles on the website were collected to automatically expand the collection of documents

Algorithm 1 Vietnamese Physics Tokenizer Algorithm (PhyTokenizer).

Input: dic – physics terminology dictionary, text – text to be tokenized.
Output: a list of tokenized text.
textList = ViTokenizer(text);   *//ViTokenizer: a Vietnamese tokenizer*
tokens = $\varnothing$;
while $i <$ textList.length **do**
    for $n = 10$ **downto** 1 **do**
        token = combineWords(i, n);   *//combineWords(i, n): combines n continuous words from //word at the position i*
        if token $\in$ dic **then**
            tokens = tokens $\cup$ token;
            $i = n + i$;
        end if
    end for
    if $n = 1$ **then**
        tokens = tokens $\cup$ textList$[i]$;
        $i = i + 1$;
    end if
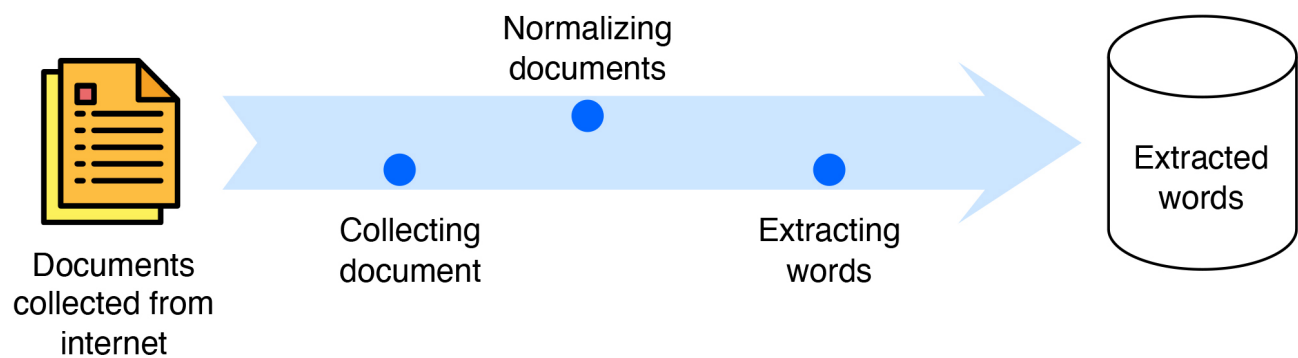   **return** tokens;
end while

Fig. 7. Document collection process.

to other pages within the same website and to ensure that the content is not duplicated.

Fig. 7 shows the document collection process.

● Step 1. Collect documents from the Internet: The BeautifulSoup library selects and extracts text from HTML files downloaded from the Internet. It is necessary to ensure that each document is associated with a unique path in this step.

● Step 2. Normalization of data: This step involves the removal of documents that do not contain the expected content, special characters, and numbers and the standardization of the Vietnamese accent to ensure homogeneity.

● Step 3. Tokenization of the collected documents: The PhyTokenizer algorithm is used to tokenize the selected documents, and the extracted words are then stored in a database.

3.2.2 Calculating the TF-IDF

The cosine similarity calculation based on TF-IDF is implemented using the scikit-learn library to facilitate the document vectorization and the similarity calculation be-

tween the query and documents of the search engine. First, the TfidfVectorizer function from the scikit-learn library is used to vectorize the documents and calculate the TF and IDF matrix. Then, the cosine similarity function is used to evaluate the similarity between the query and the indexed documents. Finally, the similarity is used to rank the search results. The appropriate parameters for these functions are determined based on thorough experiments.

3.2.3 Creating the Index for the Search Data

The primary goal of this step is to create the index for the search data (documents). The input of this step is the PhyOntology and the list of the list words tokenized by the PhyTokenizer, and the output is the indexed terms of the documents. The index creation involved two steps. First, the terms are sorted and traversed to count their frequency of occurrence. Then, each term and its frequency are added to the result list. Then, the result list is added to the PhyOntology using the workflow described in Fig. 8. Two classes, "Document" and "Index" store the index in the PhyOntology.
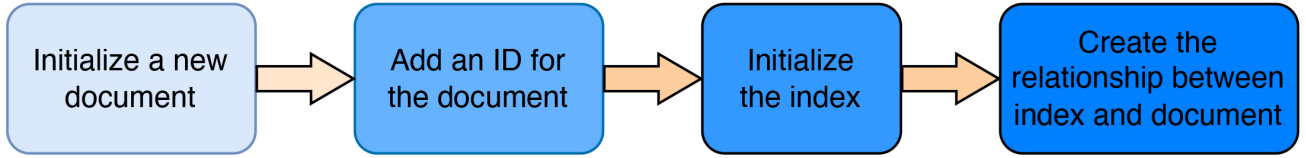
Fig. 8. OwlSearch index creation diagram.

3.2.4 Designing the Similarity Calculation and Ranking Method

Based on the standard cosine similarity score calculation (Salton, 1989), which is commonly used to evaluate the similarity between two documents' vectors with the same representation, this study introduces an extension of this calculation to accommodate the use of ontology. The cosine similarity between two documents based on TD-IDF is given in Eqn. 1.

$$\text{CosineSimilarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

Where:

- A, B: TF-IDF vectors of the two documents for which we want to calculate the similarity.

- $A \cdot B$: the dot product of the two vectors.

- A_i, B_i: the TF-IDF values for term i in documents A and B.

- $\|A\|, \|B\|$: the magnitudes (norms) of vectors A and B.

To compute the similarity between the input query and the documents with semantic information, we first extend the concepts in the input using PhyOntology. Then, the similarity score between a query Q and a document d is calculated using Eqn. 2 as follows:

$$\text{Score}(Q, d) = \frac{\sum_i W_{Q_i} W_{d_i}}{\sqrt{\sum_i (W_{Q_i})^2} \sqrt{\sum_i (W_{d_i})^2}} \tag{2}$$

This calculation is based on the idea of the cosine similarity. However, instead of computing the cosine similarity using the TF-IDF of the original query and the document, we add the semantics for the terms in the query and the document before calculating. In addition, the algorithm also includes a parameter that allows the adjustment of the importance level of a concept. Typically, the concepts in the query are considered more important than the extended concepts to ensure that documents related to the concepts in the query are ranked higher than those related to the extended concepts. Therefore, W_{Q_i} and W_{d_i} are defined in Eqn. 3 as follows:

$$W_{D_i} = tf(Q_i) \cdot idf_{\text{owl}}(Q_i)$$
$$W_{d_i} = tf(d_i) \cdot idf_{\text{owl}}(d_i) \tag{3}$$

In which,

$$idf_{owl}(q) = \frac{\sum_{i=1}^{len(Q)} m_i idf(Q_i)}{\sum_{i=1}^{len(Q)} m_i} \tag{4}$$

Parameter m in Eqn. 4 is the important level of the term Q_i mentioned above.

The proposed method for measuring the similarity between a query and a document is presented in Algorithm 2.

In Algorithm 2, the document's score value obtained represents the degree of relevance of the document to the search query, with higher scores indicating higher relevance. Documents with a score greater than 0 are ranked and returned as the result of the query.

3.2.5 Identifying the Output Structure of the Search Result

Expanding related concepts based on PhyOntology, as mentioned in the previous sections, helps learners save time in searching for documents related to a specific knowledge domain. It is crucial to provide learners with a comprehensive overview of the structure of the knowledge block they are studying, which helps them to systematize their knowledge. Therefore, in addition to expanding the search results with related keywords as presented, this study classified the keywords based on PhyOntology. The search results are expanded based on keyword categories. Based on the structure of PhyOntology, a search keyword is classified into one of the following three categories:

1. Type 1 — The keyword is the chapter's name: The search results are supplemented with the names of the lessons that belong to this chapter (using the attribute "belongs to chapter").

2. Type 2 — The keyword is the name of a lesson: The search results are supplemented with the chapter's name containing this lesson and other lessons of the same chapter. Also, the concepts in the lesson are added to the search results.

3. Type 3 — The keyword is a concept, law, theorem, or physical quantity: Depending on the case's specifics, the corresponding knowledge is added to the search results. For

Algorithm 2 Similarity Scoring Algorithm.

Input: Q – a query, Ont – the PhyOntology, Docs – set of documents.
Output: matched documents.

$Q' = Q$;
for each $q \in Q'$ **do**
　　$Q = Q \cup$ Ont.getRelatedConcepts(q);
　　/ Ont.getRelatedConcepts(q): get the related concepts of a concept q in Ont */*
end for
/ calculate the score for the tokens in Q */*
/ m_i: important level of a concept m_i, will be dicussed in the text */*
for each $q \in Q$ **do**
　　$idf_{owl}(q) = \frac{m_i idf(q)}{\sum_{i=1}^{len(q)} m_i}$
end for
/ calculate the similarity score between Q and documents in Docs */*
for each $d \in$ Docs **do**
　　$\text{Score}(d, Q) = \frac{\sum_i W_{d_i} W_{Q_i}}{\sqrt{\sum_i \left(W_{d_i}\right)^2} \sqrt{\sum_i \left(W_{Q_i}\right)^2}}$
　　/ i: no of tokens in d/Q; d_i: token number i in d; Q_i: token number i in Q */*
　　/ $W_{Q_i} = tf\left(Q_i\right) * idf_{\text{owl}}\left(Q_i\right)$ */*
　　/ $W_{d_i} = tf\left(d_i\right) * idf_{\text{owl}}\left(d_i\right)$ */*
end for
Docs = sort(Docs); */* sort Docs in descending order based on their scrore over Q */*
return Docs

example, if it is physical quantities, the name of the quantity, the unit of measurement, the symbol of the quantity, etc., will be added to the search results.

4. Implementation and Evaluation

A semantic ontology-based search engine was implemented based on the proposed method, and some experiments were conducted with the developed system to evaluate the method.

4.1 System Implementation

A semantic search system based on PhyOntology, called PhyOwlSearch, has been implemented using Python programming language and Owlready2 (Lamy, 2021) to provide a variety of features, some basic of which are listed below:

● "Crawler" function: This function allows administrators to collect data from the internet. The administrator specifies a link and a description of the HTML file structure to be crawled. The collected data is stripped of HTML tags, special characters, and numbers and then processed using the PhyTokenizer algorithm to extract words stored in the database.

● "Upload documents" function: This function allows users to upload files in doc, docx, and pdf formats to the search engine. The users must provide a title, description, path to the document file, cover image, and source document. In addition, this function will verify the documents before uploading them to the system.

● "Index" function: This function indexes all documents uploaded to the system. If the term list in the docu-

ment changes, the user must synchronize terms between the ontology and the dictionary (the index is stored in the ontology) by calling the "Synchronize Dictionary" sub-function of this function and then calling the index function again.

● "Search" function: This function allows users to enter a search keyword and returns a list of documents ranked by relevance, with the highest-ranked documents displayed first.

● "Pattern management" function: This function allows the system administrator to manage patterns such as creating new patterns, modifying an existing pattern, setting the pattern parameters, etc.

The system architecture is described in Fig. 9, which consists of three primary functional components: a database and an ontology.

1. Data collector: This component provides functions to collect data for building the ontology and search engine, including (1) the Crawler provides functions to collect data from websites that provide general physical materials, and (2) the Document Uploader allows users to upload documents to the system. In addition, this component also provides some pre-processing data functions, such as removing special characters, tokenizing words (using the PhyTokenizer algorithm), etc.

2. Indexer: This component provides two primary functions: (1) calculating the TF-IDF for concepts, and (2) creating indexes for documents in the search engine. The TF-IDF values are stored in the system database, while the index is stored in PhyOntology.

3. Search engine: This component performs three main functions. First, the Query Parser receives user
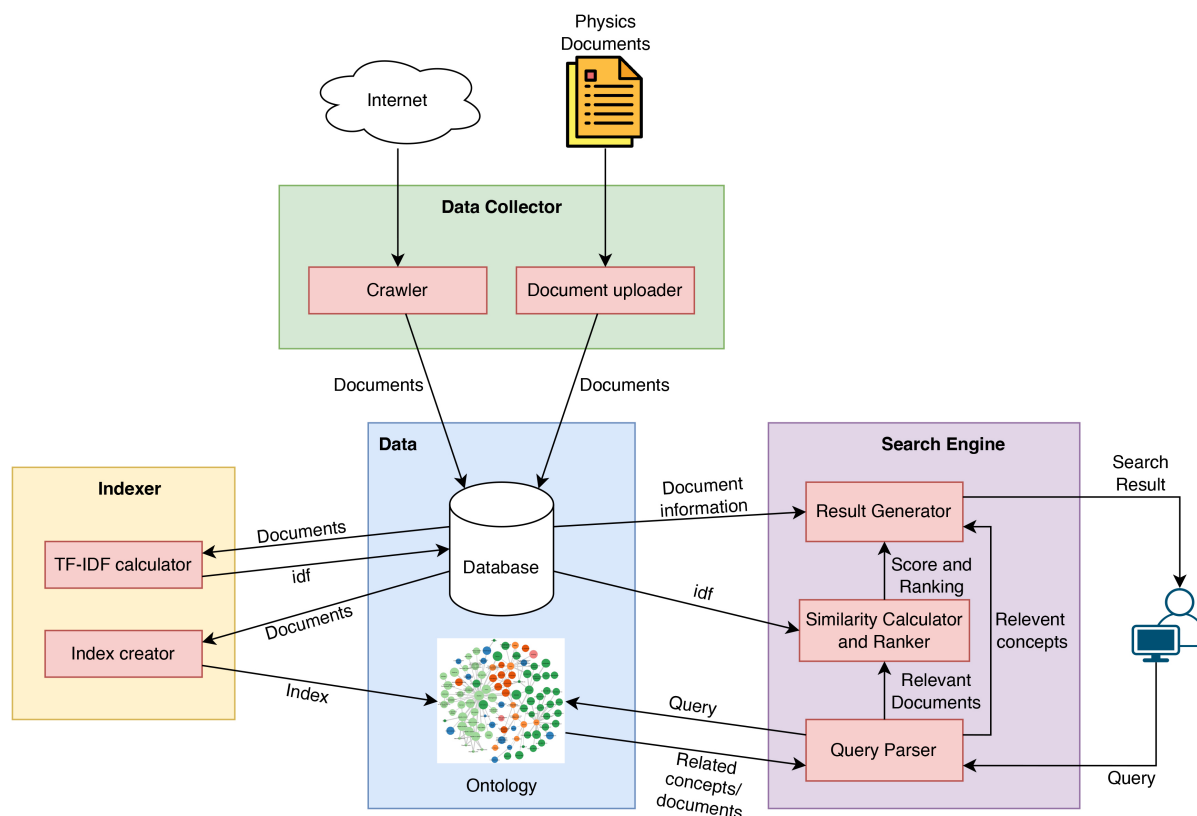
Fig. 9. The PhyOwlSearch system architecture. TF-IDF, Term Frequency - Inverse Document Frequency.

queries, tokenizes them into terms, queries the ontology to obtain a list of related terms, calculates the coefficient for concepts, analyzes the query to identify the type of concept that occurs in the query, and is used to expand the query result. Second, the Similarity Calculator and Ranker component calculates the similarity between the expanded query and the document stored in the search engine. Third, the index stored in the ontology speeds up the search process. Finally, the Result Generator generates the search result by combining the results generated by Query Parser, and Similarity Calculator and Ranker.

Fig. 10 illustrates the most crucial feature of the system, the semantic search function. This function allows users to enter keywords to find the desired information and choose between semantic and keyword-based searches (using the TF-IDF and cosine similarity). In the search result, the user can see the content of the terms or open the documents in the list of related documents.

More particular, Fig. 10 shows the result of the query "conservation of mechanical energy" ("Bảo toàn cơ năng" in Vietnamese), which consists of two parts. Part 1 presents the knowledge content associated with the search terms "conservation of mechanical energy" using the properties "constant", "have", and "is the lesson's content" (textbook lesson). Part 2 contains documents matching the search terms, which are sorted in descending order according to the relevance of the documents to the search terms.

4.2 Data Collection and Index Creation

To evaluate the system, we first collected data using the Crawler components. Next, this data was used to investigate the system's performance. When using this function, it is necessary to focus on the content of the articles to find out which HTML tags correspond to the desired values and how to select them. For example, Fig. 11 shows the result of crawling data from 6 physics content provider websites, which yielded 5596 crawled documents and the total size of the dataset is about 300 MB.

The crawled data was utilized to extend the ontology and generate the search engine index. The indexing process for the crawled documents took 115,644 milliseconds. The computational complexity of this task is O($n + n \log n$), where n represents the number of words in the indexed documents. Additionally, ontology expansion required approximately 20 seconds per pattern on the crawled dataset. When a new document is crawled, the time required for ontology expansion and index creation can be estimated based on these results, as the process is expected to scale linearly with the data size.

4.3 Search Result Analysis and Evaluation

To evaluate the implemented system, three test cases were created and tested. In addition, a comparison between the semantic search results and the TF-IDF method was also provided.

Enter search keywords   **Submit**

☑ OWL Search   ☐ TF-IDF Search

Search result for the keywords: conservation of mechanical energy

Kết quả tìm kiếm của nội dung: bảo toàn cơ năng

Thời gian tìm kiếm: 378.9ms - Số lượng kết quả trả về: 170   *Search time: 378.9 ms – Number of results: 170*

không đổi   cơ năng   *unchanging: mechanical energy*

có   biểu thức thế năng   *has: potential energy expression*

có   biểu thức động năng   *has: kinetic energy expression*

là nội dung của bài   dao động cơ   *is the content of the lesson: mechanical oscillation*

là nội dung của bài   năng lượng dao động điều hòa   *is the content of the lesson: energy in simple harmonic motion*

topic on mechanical work, power, kinetic energy, potential energy, general physics (Score: 0.9919)

chuyên đề công cơ học, công suất, động năng, thế năng, vật lý phổ thông | vật lý phổ thông *(Score: 0.9919)*

Kiến thức liên quan trong tài liệu: bảo toàn cơ năng, cơ năng, biểu thức thế năng, biểu thức động năng

related documents: conservation of mechanical energy, mechanical energy,...

Xem chi tiết nội dung... | cache file...   *View detail content... | cache file...*

bài tập vật lý lớp 10 bảo toàn năng lượng, vật lý phổ thông | vật lý phổ thông *(Score: 0.9538)*

Kiến thức liên quan trong tài liệu: bảo toàn cơ năng, cơ năng

Xem chi tiết nội dung... | cache file...

bài tập vật lý lớp 10 va chạm, đạn nổ, vật lý phổ thông | vật lý phổ thông *(Score: 0.9483)*

Kiến thức liên quan trong tài liệu: bảo toàn cơ năng, cơ năng

Xem chi tiết nội dung... | cache file...

bài tập dao động tắt dần, các loại dao động cơ | vật lý phổ thông *(Score: 0.9104)*

Kiến thức liên quan trong tài liệu: bảo toàn cơ năng, cơ năng

Xem chi tiết nội dung... | cache file...

Fig. 10. Search results with the semantic search engine with the term "conservation of mechanical energy" ("Bảo toàn cơ năng" in Vietnamese).

links	*description* mô tả	*no of returned documents* số tài liệu
http://360.thuvienvatly.com	Website chứa nhiều thông tin vật lý như bài báo, thông tin thú vị, các ứng dụng thực tế,...	4509
https://vatlypt.com	Tài liệu vật lý THPT	608
http://www.vatlyphothong.net/vat-ly-thpt	Có bài viết chi tiết của chương trình vật lý 12	38
https://hoc247.net/ chuong-trinh/ lop-12/vat-ly	https://hoc247.net/ chuong-trinh/ lop-12/vat-ly	50
https://dinhluat.com/vat-ly	Dữ liệu về các định luật vật lý	90
https://vatly247.com	Vật lý 8 đến 12	301

Fig. 11. Data collected by Crawler.

4.3.1 Test Case 1 (Chapter Name)

In test case 1, the keyword "sóng cơ" ("mechanical waves" in English) was entered. This is the title of a chapter in the Year 12 General Physics program, and the search engine correctly recognized it. Therefore, as described in Section 3.2, the "Chapter: Mechanical Waves" and its summary were given priority to be displayed first in the search results, as shown in Fig. 12.

Furthermore, the search result also included additional information about this keyword obtained by expanding the search keyword using PhyOntology. This information included the level of knowledge (year 12), a list of lessons within the chapter (five lessons) associated with the lesson number, and its page number. In addition, expanding the search keyword, its hypernyms (e.g., "wave phenomenon", "physical phenomenon") and related concepts (e.g., "amplitude", "wavelength", "frequency", "wave propagation velocity", etc.) were also displayed in the search results.

4.3.2 Test Case 2 (Lesson Name)

In this test case, we entered the keyword "phương trình sóng" ("wave equation" in English), which is the name of a lesson. As in Test case 1, the keyword was correctly classified as a lesson name, and the lesson's number, name, and page number were displayed first. Then, the concepts contained in the lesson were listed, which are the main content of the lesson. In addition, based on the PhyOntology,

Kết quả tìm kiếm của nội dung: **sóng cơ**   *Search result for the keywords: mechanical waves*

Thời gian tìm kiếm: 730.5 ms - Số lượng kết quả trả về: 1534   *Search time: 730.5 ms – Number of results: 1534*

Chương Sóng cơ   *Chapter Mechanical wave*

Thuộc khối K12   *Grade K12*

Gồm có các bài sau: *Include lessons:*

- Bài 14. Phương trình sóng (trang 70)
- Bài 17. Sóng âm và nguồn nhạc âm (trang 90)
- Bài 16. Giao thoa sóng (trang 84)
- Bài 15. Phản xạ sóng và sóng dừng (trang 79)
- Bài 18. Hiệu ứng đốp-ple (trang 96)

Lesson 14: Wave equation
Lesson 17: Sound waves and musical sound sources
Lesson 16: Wave interference
Lesson 15: Wave reflection and standing waves
Lesson 18: Doppler effect

là hiện tượng sóng   *is wave phenomenon*
là hiện tượng vật lý   *is physical phenomenon*

có nội dung chính là hay còn gọi là sóng cơ học. Sóng cơ học là loại sóng lan truyền dao động cơ học của các phần tử môi trường vật chất. Trong khi sóng có thể di chuyển và truyền năng lượng trên quãng đường dài, thì các phần tử mô trường chỉ dao động quanh vị trí cân bằng của nó. [1] Sóng mặt ở mặt nước. Môi trường vật chất có tính đàn hồi và quán tính nên một tác động dịch chuyển sẽ gây rung động và tạo ra sóng. Cũng vì thế mà sóng cơ học khác với sóng điện từ là không lan truyền được qua môi trường không vật chất như chân không. Năng lượng được truyền đi cùng hướng với sóng. Khi dao động nhỏ không gây biến dạng môi trường thì sóng cơ học còn được gọi là sóng đàn hồi (elastic wave), như sóng âm thanh. Các dao động quá lớn thì gây phá hủy liên kết của môi trường, không còn đàn hồi nữa, như sóng động đất, sóng bạc đầu ở mặt nước.

được định nghĩa là dao động cơ lan truyền trong một môi trường. Khi các phần tử của môi trường dao động theo phương vuông góc với phương truyền sóng, ta gọi là sóng ngang. Ví dụ như sóng mặt nước

Các kiến thức liên quan: biên độ, bước sóng, chu kỳ, dao động, tần số, sóng dọc, tốc độ truyền sóng, sóng ngang , tâm dao động, đỉnh sóng   *Related knowledge: amplitude, wave length, cycle, mechanical work, frequency, …*

Fig. 12. Search results with the keyword "Sóng cơ" ("Mechanical Waves" in English).

the hypernym (Chapter "mechanical waves") and the hyponyms (a list of related concepts, e.g., "amplitude") were also included in the result. The result of this test case is shown in Fig. 13.

4.3.3 Test Case 3 (Physical Quantity)

In this test case, we entered the keyword "chu kỳ" ("cycle" in English). This is a physical quantity, and as designed, the search engine is expected to return the related concepts and related knowledge of the cycle. Fig. 14 shows the result of the search engine. First, the search term was recognized as a physical quantity. Therefore, the related information about this quantity is displayed, such as the notation, unit, formula, and definition. In addition, the related knowledge to cycle was also returned, such as "mechanical waves", "motor oscillations", and "oscillation systems". With each piece of knowledge returned in the search result, a link to that knowledge was also provided so that the users could click on the link to navigate to the desired content.

Search result for the keywords: wave equation
Kết quả tìm kiếm của nội dung: **phương trình sóng**
Thời gian tìm kiếm: 521.8 ms - Số lượng kết quả trả về: 68
Lesson: Lesson 14. Wave equation (page 70)
Bài Bài 14. Phương trình sóng (trang 70)

Thuộc chương Sóng cơ   *Belong to chapter Mechanical*

Gồm có các nội dung: *Content included:*

- Sóng cơ
- Bước sóng
- Chu kỳ sóng
- Tần số sóng
- Sóng sọc
- Sóng ngang
- Tâm dao động
- Đỉnh sóng

- *Mechanical wave*
- *Wavelength*
- *Wave cycle*
- *Wave frequency*
- *...*

có nội dung chính là tìm hiểu các phương trình của sóng cơ

Các kiến thức liên quan: li độ   *Related knowledge: displacement*

Fig. 13. Search result with the keyword "Phương trình sóng" ("Wave equation" in English).

Table 7. The precision @10 result of two search methods.

No.	Keyword	Precision @10	
		Semantic	TF-IDF
1	Chu kỳ *(cycle)*	1.0	0.2
2	Dao động *(oscillation)*	1.0	0.9
3	Chuyển động quay *(rotational motion)*	0.9	0.7
4	Biến thiên *(variation)*	1.0	0.7
5	Tán xạ *(scattering)*	0.8	0.6
6	Tán sắc ánh sáng *(dispersion of light)*	1.0	0.8
7	Hiện tượng quang điện *(photoelectric effect)*	1.0	1.0
8	Máy biến áp *(transformer)*	0.8	0.7
9	Hiệu suất *(efficiency)*	0.9	0.8
10	Dòng điện *(electric curent)*	1.0	0.9
	Average	94.0%	73.0%

Search result for the keywords: cycle

Kết quả tìm kiếm của nội dung: chu kỳ

Thời gian tìm kiếm: 257.3 ms - Số lượng kết quả trả về: 1024

Ký hiệu là T *Notation T*

có đơn vị là giây *has unit second*

có công thức $T = \frac{2\pi}{\omega}$ (Chu kỳ T của dao động)

được định nghĩa là thời gian thực hiện 1 chu trình

Các kiến thức liên quan: sóng cơ, dao động cơ, hệ dao động
Related knowledge: mechanical waves, motor oscillations, oscillating systems

Fig. 14. Search result with the keyword "chu kỳ" ("cycle" in English).

4.3.4 Comparison between Semantic and TF-IDF Search Results

Test cases 1, 2, and 3 have shown different semantic search results and provided discussions to explain the search results and their benefits. This evaluation uses the precision @10 metric to compare the semantic search based on ontology and the typical search method using TF-IDF combined with cosine similarity (TF-IDF method). Table 7 compares these two methods.

Based on the precision @10 results for the ten keywords in Table 7, it can be seen that the proposed semantic search method outperformed the traditional TF-IDF method using the cosine similarity method. The precision @10 for the semantic search ranges from 0.8 to 1.0 with an average of 94%, while the precision and recall @10 for the TF-IDF method ranges from 0.6 to 1.0 with an average of 77%.

The results indicate that the semantic search method better understands the meaning of the query and the documents in the physics domain, which leads to more relevant results. In contrast, the TF-IDF method relies on the frequency of occurrence of the search terms in the documents and may not capture the subtle relationship between

the terms. Overall, the proposed semantic search method shows promise in improving the accuracy of search results in the physics domain. Further studies with larger datasets and diverse queries may provide more insight into the method's effectiveness.

Table 8 shows the ten documents with the highest similarity score found by the semantic search method for the keyword "chu kỳ" (cycle). These documents were evaluated by the high school physics teachers at Nguyen Thi Minh Khai High School, Soc Trang Province, Vietnam, and all of them are relevant to the search term. Table 9 shows the search result for the same search term using the TF-IDF method, which yielded only 2 of 10 relevant documents. Of the remaining documents, one was related to astrophysics, and the remaining seven were related to chemistry and physics.

5. Conclusion

This work focuses on building an ontology for the field of general physics that essentially meets the requirements of supporting information search engines and visually represents learning content through search results. To achieve this, a physics ontology, called PhyOntology, was created that contains fundamental concepts, properties, and instances of general physics, emphasizing data representation. The PhyOntology was created using the pattern-based method, a semi-automatic approach. First, the base ontology was created manually using a dictionary of physics. Then, 13 patterns were defined to extend the base ontology. The result is an ontology with a more extensive scope that better supports information retrieval in semantic search engines. Finally, a semantic search engine for teaching and learning general physics was developed, which involves searching, ranking, and presenting results based on semantics. It also provides an effective way to access and understand general physics information.

The creation and testing of the PhyOwlSearch application showed that the system's search results have several advantages. The returned results show the relationship be-

Table 8. Top 10 documents returned from the semantic search engine PhyOwlSearch containing the keyword "chu kỳ" ("cycle" in English).

No.	Title of document	Expanded Terms	Similarity Score
1	Sơ lược kiến thức trọng tâm vật lý 12 *(An overview of the core knowledge of physics 12)*	giây, dao động cơ, hệ giao động *(second, mechanical oscillation, coupled oscillation)*	0.958
2	Bài tập dao động tắt dần, các loại dao động cơ *(Exercise on Damped Oscillations and Various Types of Oscillating Systems)*	giây, dao động cơ, hệ giao động *(second, mechanical oscillation, coupled oscillation)*	0.947
3	660 câu bài tập trắc nghiệm lý thuyết vật lý 12 *(660 multiple-choice theoretical exercises in Physics 12)*	dao động cơ, hệ giao động *(mechanical oscillation, coupled oscillation)*	0.941
4	Bài giảng dao động và sóng (phần 5) *(Lecture on oscillation and waves (Part 5))*	giây, dao động cơ, hệ giao động *(second, mechanical oscillation, coupled oscillation)*	0.934
5	Lecture on oscillation and waves (Part 2) *(Bài giảng dao động và sóng (Phần 2))*	giây, dao động cơ, hệ giao động *(second, mechanical oscillation, coupled oscillation)*	0.934
6	Bài giảng dao động và sóng (Phần 1) *(Lecture on oscillation and waves (Part 1))*	giây, dao động cơ, hệ giao động *(second, mechanical oscillation, coupled oscillation)*	0.934
7	Đề thi thử vật lý lớp 12 thi quốc gia, đề thi số 11 *(Grade 12 Physics Mock Exam for National Exam Preparation, Exam No. 11)*	sóng cơ, dao động cơ *(mechanical waves, mechanical oscillation)*	0.928
8	Các dạng bài tập về sóng cơ *(Various exercises on mechanical waves)*	giây, sóng cơ *(second, mechanical waves,)*	0.923
9	Dao động cưỡng bức - cộng hưởng (Forced oscillation and resonance)	hệ giao động *(coupled oscillation)*	0.922
10	Bài tập vật lý lớp 12 xác định tần số góc, tần số, chu kỳ dao động điều hòa của lắc lò xo *(The 12th-grade physics homework on the angular frequency, frequency, and period of harmonic oscillations of a spring pendulum)*	dao động cơ, hệ giao động *(mechanical oscillation, coupled oscillation)*	0.921

Table 9. Top 10 documents returned from the search engine using TF-IDF and cosine similarity containing the keyword "chu kỳ" ("cycle" in English).

No.	Title of document	Similarity Score
1	Quỹ đạo dao động của con lắc *(Oscillation trajectory of a pendulum)*	0.234
2	Lịch Maya không hề tiên tri ngày tận thế 2020 *(The Maya calendar did not predict the end of the world in 2020)*	0.229
3	Nguyên tố Rutherfordium *(Rutherfordium element)*	0.192
4	Bảng tuần hoàn hóa học tốc hành (Số 29) *(Express chemical periodic table (No. 29))*	0.191
5	Nguyên tố Nnobellium *(Nnobellium element)*	0.187
6	Bảng tuần hoàn hóa học tốc hành (Số 95) *(Express chemical periodic table (No. 95))*	0.185
7	Trái tim thuộc của trái đất làm thay đổi độ dày ngày đêm *(The Earth's restless heart causes changes in daily thickness)*	0.183
8	100 câu trắc nghiệm con lắc đơn có đáp án *(100 multiple choice questions on simple pendulum with answers)*	0.181
9	Nguyên tố Roentgenium *(Roentgenium element)*	0.177
10	Nguyên tố Hassium *(Hassium element)*	0.169

tween knowledge by chapters, lessons, and lesson content, represented visually to show their interconnections. In addition, the system enables users to navigate through terms with links to the relevant content, recognizing the entered query as the name of the chapter, article, or term and offering different representations according to the content. The system has introduced an algorithm to calculate the similarity and rank of documents based on the semantic relationship between the query and other related concepts. Based on this, the system can display comprehensive information so that users can quickly and conveniently access relevant documents. The proposed method was compared with a traditional search approach that combines TF-IDF and cosine similarity. The evaluation result shows that the PhyOwlSearch system yielded a better result on the precision @10 metric.

Besides the results achieved, this work has some limitations. First, the documents were collected from real sources used by teachers and students. However, the various authors' varying writing styles and presentation methods have led to inconsistencies that make extracting information to create ontology difficult. In addition, the presen-

tation of the content in the form of a list of documents needs to be developed more intuitively, such as the presentation of the relationship between concepts through diagrams.

Availability of Data and Materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

ACT: Conceptualization, Formal Analysis, Supervision, Methodology, Validation, Writing–original draft, Writing– review & editing. BD-P: Conceptualization, Data curation, Validation, Software, Formal Analysis, Writing–review. HTN: Validation, Writing–original draft. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Acknowledgment

Not applicable.

Conflict of Interest

The authors declare no conflict of interest.

References

Abdi A, Idris N, Ahmad Z. QAPD: An ontology-based question answering system in the physics domain. Soft Computing. 2018; 22: 213–230. https://doi.org/10.1007/s00500-016-2328-2

Alatrish ES. Comparison of ontology editors. eRAF Journal on Computing. 2012; 4: 23–38.

Andrade-Hoz J, Vega-Gorgojo G, Ruano I, Bote-Lorenzo ML, Asensio-Pérez JI, Bravo F, et al. Educawood: A socio-semantic annotation system for environmental education. In European Conference on Technology Enhanced Learning. Springer International Publishing: Cham. 2021. https://doi.org/10.1007/978-3-030-86436-1_37

Antoniou G, Van Harmelen F. A semantic web primer. MIT Press: Cambridge, MA. 2004.

Bajpai AK, Pandey R, Tripathi SP. A semantic annotation scheme for mapping provenance in educational ontology. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT) (pp. 872–875). IEEE. 2021. https://doi.org/10.1109/CSNT51715.2021.9509610

Bouchée T, de Putter-Smits L, Thurlings M, Pepin B. Towards a better understanding of conceptual difficulties in introductory quantum physics courses. Studies in Science Education. 2022; 58: 183–202. https://doi.org/10.1080/03057267.2021.1963579

Chen J, Gu J. ADOL: a novel framework for automatic domain ontology learning. The Journal of Supercomputing 2021; 77: 152–169. https://doi.org/10.1007/s11227-020-03261-7

Chen J, Gu J. Developing educational ontology. In Proceedings of the 10th International Conference on Education Technology and Computers. ACM Press. 2018. https://doi.org/10.1145/3290511.3290546

Cook DL, Bookstein FL, Gennari JH. Physical properties of biological entities: An introduction to the ontology of physics for biology. PLoS One. 2011; 6: e28708. https://doi.org/10.1371/journal.pone.0028708

Elkan C, Greiner R. Building large knowledge-based systems: Representation and inference in the CYC project: D.B. Lenat and R.V. Guha. Artificial Intelligence. 1993; 41–52. https://doi.org/10.1016/0004-3702(93)90092-p

Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubézy M, Eriksson H, et al. The evolution of Protégé: an environment for knowledge-based systems development. International Journal of Human-computer Studies. 2003; 58: 89–123. https://doi.org/10.1016/S1071-5819(02)00127-1

Hung VV, Khiet VT. Physics Dictionary. Vietnam Education Publishing House: Ha Noi, VN. 2015. (In Vietnamese)

Kapoor B, Sharma S. A comparative study ontology building tools for semantic web applications. International Journal of Web & Semantic Technology. 2010; 1: 1–13. https://doi.org/10.5121/ijwest.2010.1301

Lamy JB. Ontologies with Python. Apress: Berkeley, CA. 2021.

Lamy JB. OwlReady: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artificial Intelligence in Medicine. 2017; 80: 11–28. https://doi.org/10.1016/j.artmed.2017.07.002.

Lee E, Harris N, Gibson M, Chetty R, Lewis S. Apollo: a community resource for genome annotation editing. Bioinformatics. 2009; 25: 1836–1837. https://doi.org/10.1093/bioinformatics/btp314

Maass W, Janzen S. A pattern-based ontology building method for ambient environments. In CEUR Workshop Proceedings. 2009.

Maynard D, Funk A, Peters W. SPRAT: A tool for automatic semantic pattern-based ontology population. In International Conference for Digital Libraries and the Semantic Web. 2009.

Nguyen VH, Tran VT. ViTokenizer. 2024. Available at: https://github.com/trungtv/pyvi (Accessed: 10 April 2024)

Panchal R, Swaminarayan P, Tiwari S, Ortiz-Rodriguez F. AISHE-onto: A semantic model for public higher education universities. In DG.O2021: The 22nd Annual International Conference on Digital Government Research. ACM. 2012. https://doi.org/10.1145/3463677.3463750

Parmeshwaran P, Rege J, Nair S. The use of ontology in semantic search techniques. International Journal of Computer Applications. 2015; 127: 21–24.

Faisal R, Kitasuka T, Aritsugi M. Semantic cosine similarity. In The 7th International Student Conference on Advanced

IMR Press

Science and Technology ICAST. South Korea: University of Seoul. 2012.

Rahayu NW, Ferdiana R, Kusumawardani SS. A systematic review of ontology use in e-learning recommender systems. Computers and Education: Artificial Intelligence. 2022; 3: 100047. https://doi.org/10.1016/j.caeai.2022.100047

Rana RK, Tyagi N. A novel architecture of ontology-based semantic web crawler. International Journal of Computer Applications. 2012; 975: 8887.

Salton G. Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley. 1989.

Sánchez JAP, Pérez TS. A Conceptual model for an OWL ontology to represent the knowledge of transmedia storytelling. In Challenges and Opportunities for Knowledge Organization in the Digital Age (pp. 511–520). Ergon-Verlag. 2018.

Shapovalov YB, Shapovalov VB, Tarasenko RA, Usenko SA, Paschke A. A semantic structuring of educational research using ontologies. In CTE Workshop Proceedings (pp. 105–123). 2021.

Suprapto N. Physics education students' understanding of the concept of epistemology, ontology, and axiology. Journal of Physics: Conference Series (pp. 012015). IOP Publishing. 2021. https://doi.org/10.1088/1742-6596/1747/1/012015

Uschold M, King M. Towards a Methodology for Building Ontologies. Artificial Intelligence Applications Institute, University of Edinburgh: Edinburgh. 1995.

Uschold M. Building ontologies: Towards a unified methodology. In Proceedings of 16th annual conference of the British Computer Society Specialists group on expert systems. Citeseer. 1996.

Usta A, Altingovde IS, Ozcan R, Ulusoy Ö. Learning to rank for educational search engines. IEEE Transactions on Learning Technologies. 2021; 14: 211–225.