Indonesian Journal of Computer Science
ISSN 2302-4364 (*print*) dan 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam, No. 1, Padang, Indonesia, Telp. (0751) 7056199, 7058325
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Semantic Information Search with Automatic Ontology Creation in Regulations National Standards for Higher Education in Indonesia

Nadila Wirdatul Hidayah[1], Ali Ridho Barakbah[2], Iwan Syarif[3]
[1]nadilawh0@gmail.com , [2]ridho@pens.ac.id , [3]iwanarif@pens.ac.id
Politeknik Elektronika Negeri Surabaya

Article Information

Submitted: 13 May 2023
Reviewed : 5 Jun 2023
Accepted : 27 Jun 2023

Keywords

Automatic Ontology, Search Engine, Regulation, Semantic Web

Abstract

In Indonesia, there are around ten types of legal products that contain higher education regulations. With a large number of articles, more effort is needed when users search for links between one article and another. Based on these problems, it is necessary to have an automatic article representation search system using an automatic ontology. Ontology refers to the hierarchical structure of entities and their relationships. In this paper, the results of the development of an information retrieval system with an automated ontology will be explained. This system describes a process begins with receiving input of higher education regulatory files which are used as data samples MoE Regulation No 3 of 2020. Then split the data into articles, paragraphs and contents which are then formed ontologies by building 3 detection functions (Definitive Creation, Compound Creation, and Reference Detection). System output has an accuracy of search results reaching an accuracy of 92.5%.

A. Introduction

Higher education regulation is a tool that contains a number of rules made to enforce order in the education level after education which includes diploma programs, bachelor programs, master programs, doctoral programs, and professional programs, as well as specialist programs, which are organized by tertiary institutions based on Indonesian culture. . There are various kinds of rules and legal products regarding higher education regulations in Indonesia and between regulations one with another has a relationship therein.

Higher Education Regulations refer to a set of rules that are made to enforce order in the after-school education level which includes diploma programs, bachelor programs, master programs, doctoral programs, and professional programs, as well as specialist programs, which are organized by universities based on national culture Indonesia. Various provisions and norms ranging from Laws, Government Regulations, and Ministerial Regulations as well as related provisions and guidelines have been stipulated to guide Higher Education institutions to implement quality and dignified education, so that it is hoped that they can become a reference for the academic community of a higher education institution. Regulations are also binding on individuals within a certain scope, meaning that if someone violates the regulations that have been in effect then certain sanctions or penalties will be imposed.

In this paper, we present an information retrieval system for Higher Education Regulations in Indonesia using Automatic Ontology. The purpose of this paper is to propose a system to automatically find and classify higher education regulations for information retrieval, related documents (search for similarities between one article and another), and the formation of ontology (arrangement or chart or diagrams of the contents of several products higher education regulatory law in Indonesia). Through the combination of the three will provide complete information search results and according to user needs.

B. Research Method

This research is divided into six main sections, namely Document input, Function Mapping, Document Metadata, Function Detection, and Visualization. Each part has a process which will be explained below. This system design shows the process of how the user gets the information needed and the formation of a complete ontology.

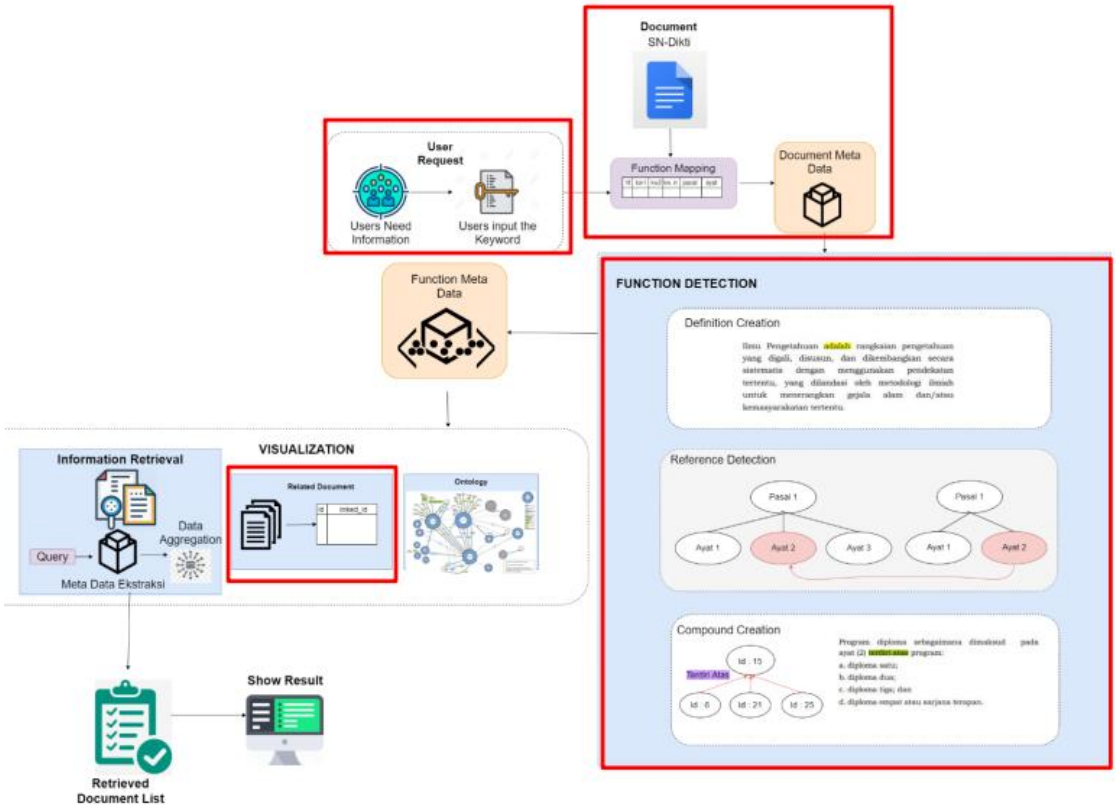


Figure1. Automatic Ontology Creation for Regulations National Standarts for Higher Education

a. User Request

In this first stage is the user request process and is initiated by the user or users who need some information, the information referred to here is information about higher education regulations in Indonesia. So start with the user entering this system to get what he wants. Information Retrieval System (Information Retrieval) is used to retrieve information that is relevant to user needs from a collection of information automatically, especially information regarding higher education regulations in Indonesia.

The information retrieval system aims to find relevant documents in an effort to fulfill information needs [3]. Information needs begin with a gap between the user's knowledge of a problem and what the user needs to know to solve the problem. After that, the user must enter certain queries according to the information needs to be obtained. So later when a user visits this search engine, he must enter a query in the form of keywords or keywords to find the data they want.

The query entered here is in the form of text, not an image or other format. Queries here are really needed because not all documents retrieved by the system are documents that are truly in accordance with what the user wants (relevant) [2]. This is what will be directly related to the user, providing the desired information search results. When a user visits a search engine and enters a search word usually with a few keywords, the search system will search for data from an index database,

matching data is then displayed, usually accompanied by a brief summary of the title of the document and sometimes a portion of the text.

b. Project Document Collection

The data used is one higher education regulation that was used in Indonesia, namely MoE Regulation No. 3 of 2020 concerning national standards for higher education. The data is taken from the regulation dataset. At the data collection stage it is divided into 4 main parts, namely Articles, Articles, paragraphs, and the contents of the paragraphs. The total data used in this study are 7 Articles, 70 articles, and 327 paragraph data, using previously stored data in csv format. From the regulatory data collection, as many as 327 data were obtained consisting of definitive creation, compound creation, and reference detection. The data will then be divided into 80:20 to be used in the training process and the testing process in classification. Furthermore, each data is labeled in which category it belongs, and from there an ontology or relationship of each element will be formed.

c. Function Mapping

Function Mapping can define a mapping between federated database functions or function templates and data source functions [7]. When defining a mapping, function mapping can associate federated or template database functions with functions in a particular data source that come from multiple data sources; for example, all data sources of a certain type and version. Function mapping can consist of some information, namely: function name, data-type, and version.

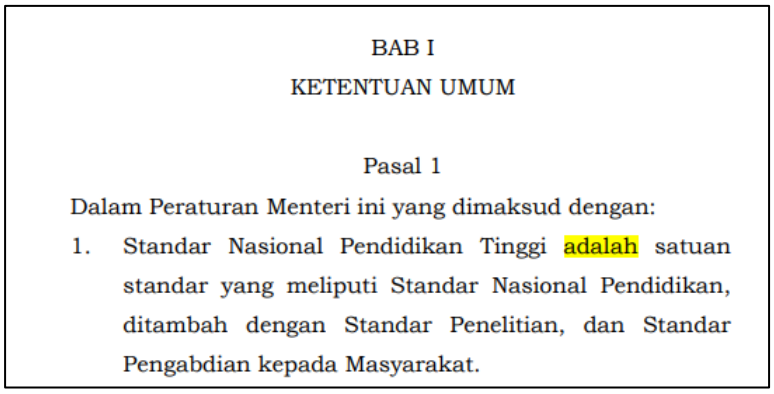
d. Function Detection

Function Detection consists of processes or basic functions to form an ontology. Some of the functions in function detection include Definitive Creation, Compound Creation, and Reference Creation. The following is an explanation of the functions of Function Detection:

Definitive Creation

It is a function of forming the meaning of a boundary or meaning, it can also be interpreted as a word, phrase, or sentence that expresses the meaning, description, or main characteristics of a person, thing, process, or activity. In this case the definitive creation contains a description of an entity or object of higher education regulations. Example: education, science, technology, research, etc. Here is an example in the rules that I use as input data :

1. MoE Regulation number 3 of 2020 ARTICLE 1 Article 1 Paragraph 1



BAB I

KETENTUAN UMUM

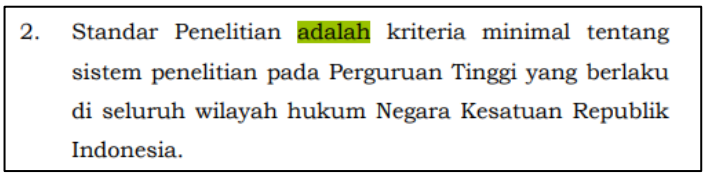
Pasal 1

Dalam Peraturan Menteri ini yang dimaksud dengan:

1. Standar Nasional Pendidikan Tinggi adalah satuan standar yang meliputi Standar Nasional Pendidikan, ditambah dengan Standar Penelitian, dan Standar Pengabdian kepada Masyarakat.

Figure 2. Example 1 for Definitive Creation in MoE Regulation Number 3 of 2020

2. MoE Regulation Number 3 Year 2020 Chapter 1 Article 1 Verses 2

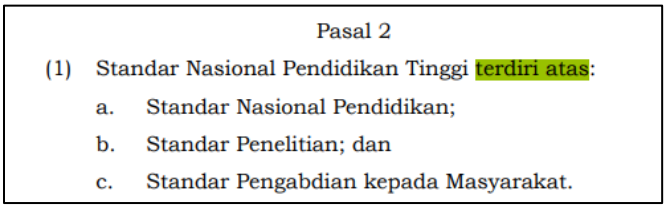


2. Standar Penelitian adalah kriteria minimal tentang sistem penelitian pada Perguruan Tinggi yang berlaku di seluruh wilayah hukum Negara Kesatuan Republik Indonesia.

Figure 3. Example 2 for Definitive Creation in MoE Regulation no 3 of 2020

Compound Creation

This function to form the meaning of "consists of" several objects. In this case the compound creation contains types or members of an entity or object of higher education regulations. Example: Cluster of knowledge, diploma program, academic degree, etc. Here is an example in the rules that I use as input data : MoE Regulation Number 3 Year 2020 Chapter 1 Article 2 Verses 1.



Pasal 2

(1) Standar Nasional Pendidikan Tinggi terdiri atas:

a. Standar Nasional Pendidikan;

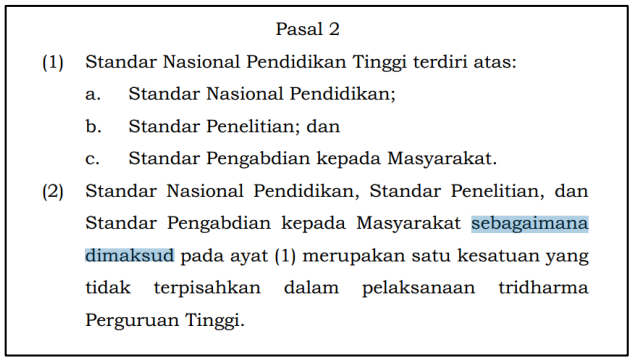
b. Standar Penelitian; dan

c. Standar Pengabdian kepada Masyarakat.

Figure 4. Example for Compound Creation in MoE Regulation no 3 of 2020

Reference Detection

It is a function to form the meaning "as intended" from one article to another. In this case Reference creation contains a relationship from an entity or object of higher education regulations. Example: Article 2 paragraph 2 have a relation with Article 2 paragraph 1. Here is an example in the rules that I use as input data :
MoE Regulation number 3 of 2020 Chapter 1 Article 2 Paragraph 2



Pasal 2

(1) Standar Nasional Pendidikan Tinggi terdiri atas:
   a. Standar Nasional Pendidikan;
   b. Standar Penelitian; dan
   c. Standar Pengabdian kepada Masyarakat.

(2) Standar Nasional Pendidikan, Standar Penelitian, dan Standar Pengabdian kepada Masyarakat sebagaimana dimaksud pada ayat (1) merupakan satu kesatuan yang tidak terpisahkan dalam pelaksanaan tridharma Perguruan Tinggi.

Figure 5. Example for Reference Detection in MoE Regulation no 3 of 2020

So, formation of ontology from the above article relationship it can be interpreted that MoE Regulation Number 3 Year 2020 Chapter 1 Article 2 Verses 2 has a references or connection with MoE Regulation Number 3 Year 2020 Chapter 1 Article 2 Verses 1. So that later this article will be categorized as included in reference creation. So, when a user opens the article MoE Regulation Number 3 Year 2020 Chapter 1 Article 2 Verses 2, a related article or recommended article will appear that can be read.

e. Visualization

Visualization consists of a process for displaying the data or output of some features. In its application, data visualization often uses several visual elements such as graphs, charts to maps [9]. However, the visualization provided in this information system is in the form of text and ontology (graphics). In short, data visualization is used to present structured or unstructured data with graphs or charts to display information hidden in the data. The purpose of data visualization is of course to make it easier to see and understand trends, patterns and much more. Some of the features in Visualization include Information Retrieval, Related Documents, and Ontology. The following is an explanation of each of the features in it:

1. Information Retrieval

Information retrieval system is used to retrieve information that is relevant to user needs from a collection of information automatically [1]. One common application of an information retrieval system is a search-engine or search engine found on the internet network. Users can search for the Web pages they need through the machine. In Information Retrieval, getting the relevant documents is not enough. The goal that must be met is how to get relevant documents and not get irrelevant documents.

The application of information retrieval in my thesis is in the form of a search engine via a web search engine. Keywords are entered by the user to search for the desired information on the Search Engine, where the information obtained contains the expected relevance/relationship. So start with the user entering this system to get what he wants. Here the user must enter certain queries according to the information needs to be obtained. Queries here are really needed because not all

documents retrieved by the system are documents that are truly in accordance with what the user wants (relevant).

2. Related Document

It is a feature to search for similarities between one article and another related article. Making it easier for users to find links between articles. the related document feature is formed by utilizing the reference detection. The following is an overview of the related document mapping:

Table 1. Document Relation Mapping

ID	Linked_ID
15	1
15	6
22	1

From the table above it can be seen that the rules with ID 15 have a relationship or relationship with the rules ID = 1 and 6. The relationship between the two can vary, for example the relationship "consists of" or something else. So that later each article will have a relationship with other articles that have the same focus of discussion.

3. Ontology

Ontology is a model language that supports functions for integrating conceptually distributed domain knowledge and inferring relationships between concepts [4]. Ontology can be interpreted as the science of being or explaining the existence of something. According to Aristotle, ontology is the study of existence, that is, a system of categories or a system of classification. The ontology used in this application represents higher education regulations based on category context. Ontology itself has several constituent components, namely concepts, relations, functions, axioms, and instances based on target domain knowledge. The purpose of Automatic Ontology Creation is to process National Higher Education standard regulations to obtain information by paying attention to the meaning of the text to then form a diagram that illustrates some of the meanings in it.

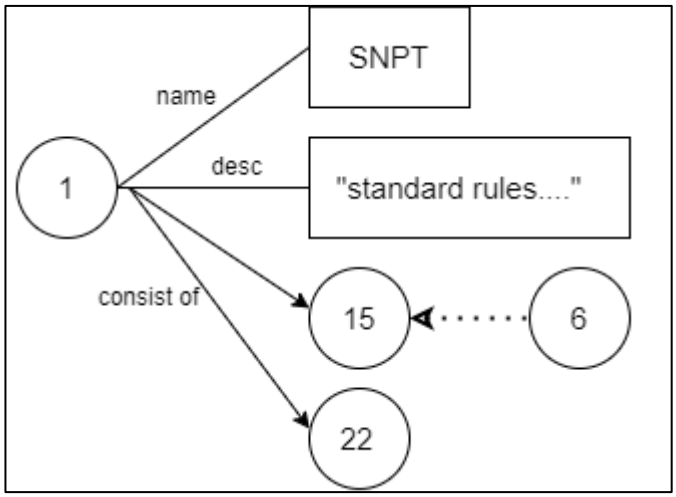


Figure 6. Illustrative Examples of Ontology Formation

C. Result and Discussion

The experiment was carried out from March to July 2022. Table 2 is a table of laptop specifications used for the experiment.

Table 2. Device Specifications for Experiment

Component	Name
Laptop	Asus X441U
OS	Windows 10 Education 64-bit
Harddisk	1 TB HDD
Processor	4GB DDR4.
Memory	Intel Core i3-6006U Skylake gen 6th
GPU	Nvidia GEFORCE 840M 2GB
Cache	6 MB

Testing was performed on the Jupyter Notebook platform. The programming language used is the Python programming language and the Using D3.js Collapsible Tree library. The python libraries used are basic libraries such as pandas, numpy, re and literary. Pandas is a python library to facilitate data processing in the form of rows and columns such as csv, sql and no-sql data formats.Library D3.js is a JavaScript library used to manipulate documents based on data. D3.js helps transform data into visual forms such as bar charts, line charts and others that can be presented to readers via a web browser, in this case, we will transform it to the ontology form according to our needs.

Table 3. Results of grouping each verse into detection features

No	Data	Definition Creation	Compound Creation	Reference Detection	Name	Description
1	In this Ministerial Regulation what is meant by: 1. The National Higher Education Standards **are** a standard unit which includes the National Education Standards, coupled with Research Standards, and Community Service Standards.	1			Standard National higher education	Standard units which include National Standards Education, augmented by Standard Research, and Community Service Standards.
2.	National Higher Education Standards **consist of:** a. National Education Standards		1			
3.	National Education Standards, Research Standards, and Community Service Standards **as referred to** in paragraph (1) are an integral part of the implementation of the Tridharma of Higher Education.			1		

Type of detection or grouping of types of feature detection of each paragraph in the document data collection used. The following is a description of each detection obtained:

1) The first row gets a value of 1 in the definition creation column, meaning that the data or verse is included in the definition creation category. The data is detected as a definition creation through the arrangement of words in it. So, every verse that contains the meaning of the word definition, for example "adalah", the data is detected as a definition creation. Then, after the data is detected as definition creation, the data is split into two parts, namely as a name and description. The name is taken from the word in the sentence before the definition word " adalah ". An example of definition creation in MoE Regulation regulation No. 3 of 2020 is shown in table 4 below.

Table 4. Definition Creation Example

Verses
Standar Nasional Pendidikan Tinggi adalah satuan standar yang meliputi Standar Nasional Pendidikan

It can be seen that before the word “adalah” it acts as the “name” of a definition creation. And the sentence after the word definition " adalah " acts as a description.

2) The second row gets a value of 1 in the compound creation column, meaning that the data or verse is included in the compound creation category. The data is detected as a compound creation through the arrangement of words in it. So, every verse that contains the meaning of the word "consists of" then the data is detected as a compound creation. Then, after the data is detected as a compound creation, the data is split into two parts, namely the name and the branch. The name is taken from the word in the sentence before the definition word "consists of". An example of Compound creation in MoE Regulation regulation No. 3 of 2020 is shown in table 5 below.

Table 5. Compound Creation Example

Verses
Standar Nasional Pendidikan Tinggi terdiri atas: a.Standar Nasional Pendidikan

It can be seen that before the word "consisting of" it acts as the "name" of a compound creation. And the sentence after the word definition "consists of" acts as a branch, namely "Standar Nasional Pendidikan".

3) The third row gets a value of 1 in the reference detection column, meaning that the data or verse is included in the reference detection category. The data is detected as a reference detection through the arrangement of words in it. So, every verse that contains the meaning of the word "as intended" then the data is detected as a reference detection. Then, after the data is detected as reference detection, the data is split into two parts, namely as data and linked_data. An example of reference detection in MoE Regulation regulation No. 3 of 2020 is shown in table 6 below.

Table 6. Reference Detection Example

Verses
Standar Nasional Pendidikan, Standar Penelitian, dan Standar Pengabdian kepada Masyarakat **sebagaimana dimaksud** pada Verses (1) merupakan satu kesatuan yang tidak terpisahkan dalam pelaksanaan tridharma Perguruan Tinggi.

It can be seen from the example above that the paragraph is related to paragraph 1. If the editorial "as intended" only contains the paragraph, without the article included. That means that the verse is in the same Article. The ontology form that was successfully generated automatically using the 3 detection features above is as follows shown in figure 7 below.

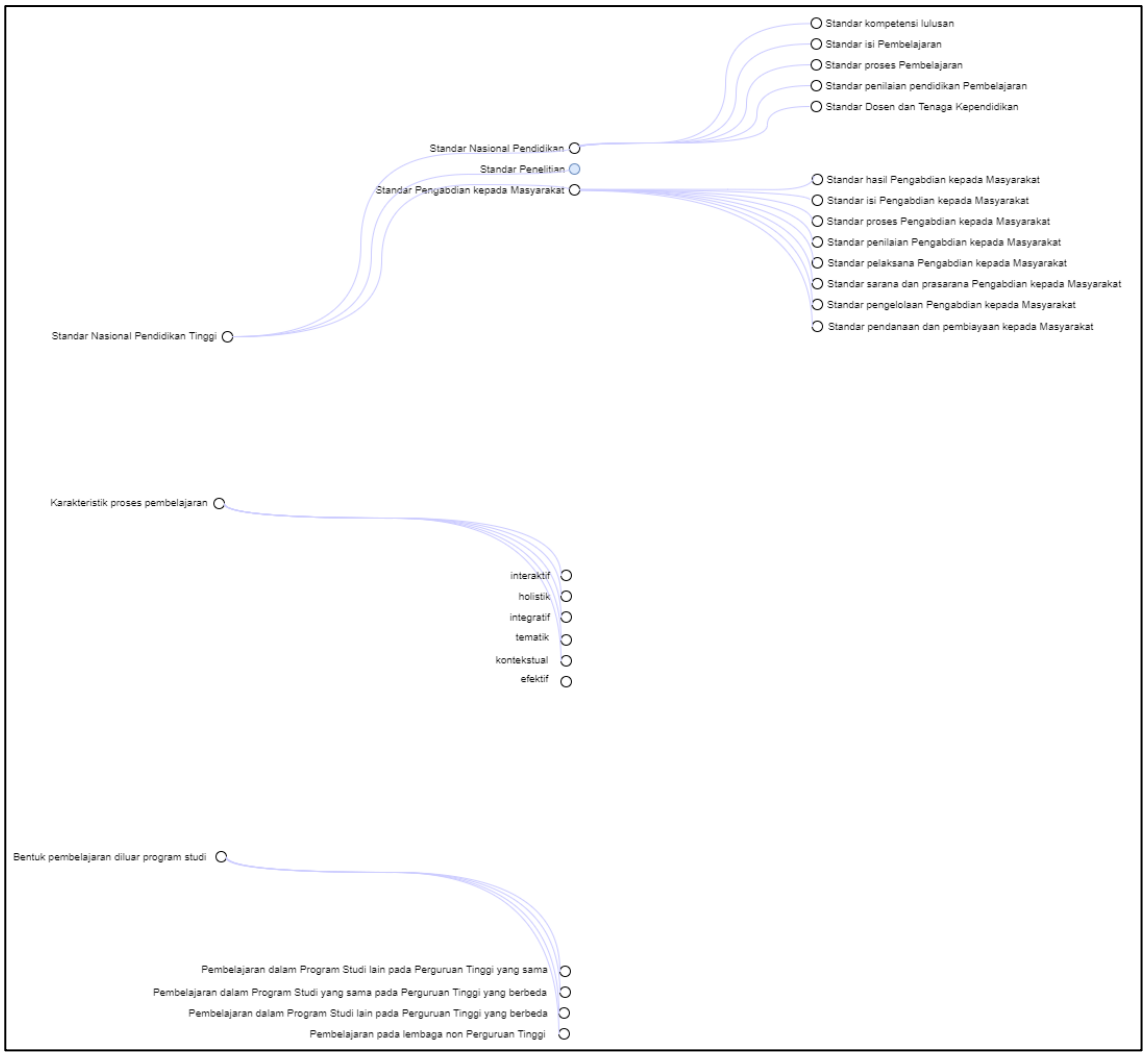


Figure 7. Ontology mapping result

Web Visualization

An information retrieval system is said to be successful when it can provide a response according to the user's input query and display several related articles or have discussions related to the input query.

Figure 8. Landing Page of the information retrieval system

Figure 8 is the initial display of the information retrieval application dashboard. On the start page, the user can enter a query that is used to find information according to user input. Then, the selected article will appear according to the information the user needs.

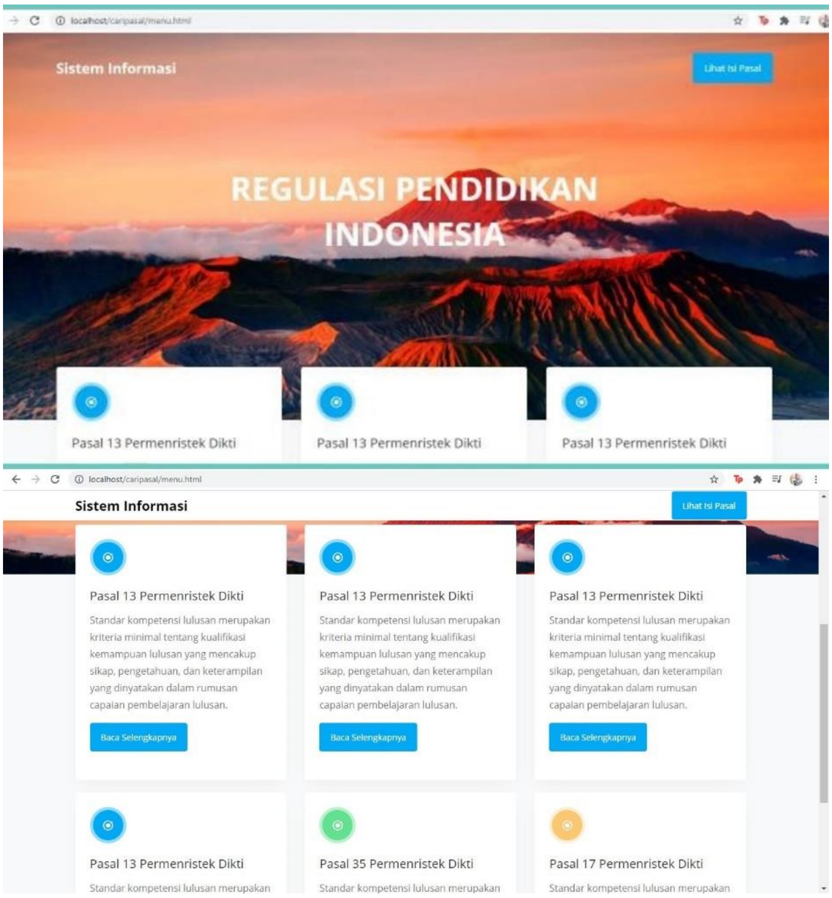


Figure 9. The article results page is related to the query

In addition, users can also view the complete contents of the selected article.

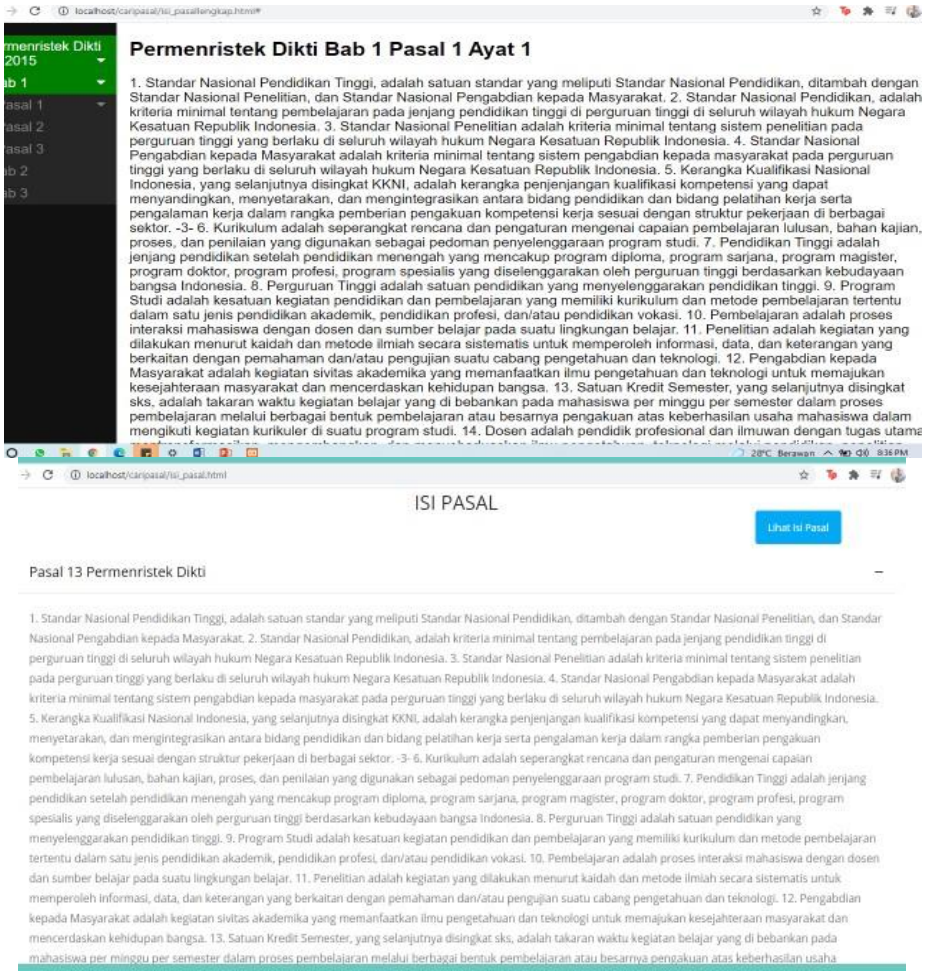


Figure 10. Information system for the entire contents of the article (details of the selected articles)

D. Conclusion

This research makes a search engine application for Indonesian Higher Education Regulations using Automatic Clustering. We get data on Higher Education Regulations from legal products that have been passed and officially apply in Indonesia. The system implementation consists of some stages, namely User Request, Project Document Collection, Function Mapping, Function Mapping, and then the automatic ontology is created. The data used in this study is MoE Regulation number 3 of 2020. The amount of data obtained from these products is approximately 327 data.

Based on the experiments in the formation of the automatic ontology, 76 data are included in the definitive creation category, 52 data are in the compound creation category (there are 121 branches), and 19 data are included in reference detection. Preliminary experimental results show that the purposed ontologies improve the precision and recall (92,5%) of the documents retrieval. As conclusion of this work we would like to highlight that semantic retrieval approach can provide better search capabilities, thus achieving an improvement over keyword-based retrieval by means of the introduction and exploitation of ontologies. Future research works include further experiments by considering large number of documents and improve number and coverage of the queries. It is also interesting

to have a generic ontology and document processing which can be used for various other event related documents.

E. Acknowledgment

Acknowledgment to God Almighty, because of His blessings and grace, I can finish this scientific writing. I realized that without the help and guidance of various parties, especially friends involved in this research and the support of facilities from the Surabaya State Electronics Polytechnic (PENS) would be quite difficult for me to complete this scientific paper.

F. References

[1] Harrow, I., Balakrishnan, R., Jimenez-Ruiz, E., Jupp, S., Lomax, J., Reed, J., … Woollard, P, "Ontology mapping for semantically enabled applications", 2019.

[2] Sayed, A., & Al Muqrishi, A. (2017). IBRI-CASONTO: Ontology-based semantic search engine. Egyptian Informatics Journal, 18(3), 181–192.

[3] Kambau, R. A., & Hasibuan, Z. A. (2017). Concept-based Multimedia Information Retrieval System using ontology search in Cultural heritage. 2017 Second International Conference on Informatics and Computing (ICIC).

[4] P. Escobar, R. -García, M. D. M., Peral, J., Candela, G., and GarcíaNieto, J. (2020). An Ontology-Based Framework for Publishing and Exploiting Linked Open Data: A Use Case on Water Resources Management. Applied Sciences, 10(3), 779.

[5] Bazeer, A., & Razwan, N. M. (2020). "Enhancement Framework of Semantic Query Expansion Using Mapped Ontology."

[6] Chi, X. (2019). "The Information Retrieval Model for Computer Crime Evidence Based on Domain."

[7] Beirade, F.; Azzoune, H.; Zegour, D.E.: Semantic query for quranic ontology. J. King Saud Univ. Comput. Inf. Sci. (2019).

[8] Satyaveer Singh and Mahendra Singh. 2019. ONTOLOGY LEARNING PROCEDURES BASED ON WEB MINING TECHNIQUES. ResearchGate 7 (2019)

[9] Girija Attigeri, Manohara Pai M M, Radhika M Pai, and Rahul Kulkarni. 2018. Knowledge Base Ontology Building For Fraud Detection Using Topic Modeling. ELSEVIER 8 (2018).

[10] I. Harrow, R. Balakrishnan, E. Jimenez-Ruiz, S. Jupp, J. Lomax, J. Reed, M. Romacker, C. Senger, A. Splendiani, J. Wilson, et al., Ontology mapping for semantically enabled applications, Drug discovery today (2019).

[11] Zouaoui, S., & Rezeg, K. (2021). A Novel Quranic Search Engine Using an Ontology-Based Semantic Indexing. Arabian Journal for Science and Engineering, 46(4), 3653–3674.