Multi-documents summarization based on clustering of learning object using hierarchical clustering

M Mustamiin¹, I Budi² and H B Santoso³

Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
m.mustamiin@gmail.com¹, indra@ui.ac.id², harrybs@cs.ui.ac.id³

Abstract. The Open Educational Resources (OER) is a portal of teaching, learning and research resources that is available in public domain and freely accessible. Learning contents or Learning Objects (LO) are granular and can be reused for constructing new learning materials. LO ontology-based searching techniques can be used to search for LO in the Indonesia OER. In this research, LO from search results are used as an ingredient to create new learning materials according to the topic searched by users. Summarizing-based grouping of LO use Hierarchical Agglomerative Clustering (HAC) with the dependency context to the user's query which has an average value F-Measure of 0.487, while summarizing by K-Means F-Measure only has an average value of 0.336.

1. Introduction
Open Educational Resources (OER) is a portal of teaching, learning, and research resources available in the public domain or have been released under an intellectual property license that allows a freely accessible [1]. The use of technology in education, including OER increases the number and variety of forms of learning content. The number of instructional content that causes the need for a convenient search system to obtain the desired learning content.

A previous study revealed that Learning Object (LO) search engine-based ontology has an average performance of 29.7%. It is better than using a search engine index-based document [2]. LO Ontology is designed to explain the concept and the relationship between learning content to facilitate the process of finding the learning content [2].

LO is granular more easily reused by other learning-related topics as well as the topic of learning itself [3]. Figure 1 shows that LO is a representation in the form of granules that can be reused as a material for the lesson topic that is still associated or used as components of the manufacture of new learning materials. LO in the form of granules and the search results on ontology-based LO search engines, in particular on the documents top of the search results, are often found materials that are related or even similar. It shows that between LO in the field of material related with LO in other materials, especially for materials that contain the same keywords, with these conditions it is possible to create an empowerment of the LO that is digital so they can be reused as material learning material a new one in accordance with the desired topic.

A study conducted by Paramartha [2] in relation to LO traceability techniques was limited to designing LO ontologies and search engines, while for LO use itself does not exist. The research also provides suggestions for creating a system that can create new materials based on LO ontologies that have been constructed by reusing sub-material that has been mapped.
LO utilization as a new learning materials that can be done by summarizing LO or sub-materials. Summarization is one of the topics in the acquisition of information retrieval to make information shorter and denser than the original document. Multi-documents summarization used to summarize more than one document, there are several methods in multi-documents summarization, ranging from methods based on graph-based and cluster-based [4]. Multi-documents summarization based graph using a specific query [5] and there is also multi-documents summarization using clustering sentence that produces good results [6].

To get the LO summarization diverse and do not overlap and can produce summarization accordance with the desires of user is required appropriate grouping method in the LO grouping of search results. According to Shepitsen [7] the context dependence on Hierarchical Agglomerative Clustering (HAC) can improve the result of recommendations in accordance with the user profile. In the context of search, user profiles can be compared with the wishes of the user in finding information. Therefore, reliance on HAC context that has been proven to improve on as the user desires can be applied to multi-documents summarization based on the results of the query entered by the user.

In this study, the authors conducted a development of the search engine LO-based ontologies created by Paramartha [2] by adding summarization features against LOs search results and representing it becomes a collection of information that can be a reference for the user to create learning materials related topics of interest. Document summarization process search results using multi-documents summarization based grouping HAC with dependency context of the user’s query.

2. Material and Method

2.1. Research Data

Data were taken from the research data tracking techniques ontology-based LOs by Paramartha [2]. The research data is in the form of ontology files, based on the information in the file, the file was created using OWL ontology API (version 3.4.2), where the ontology files is obtained from extraction of 115 lecture slides of five courses, including computer network (CN), web programming (WP), enterprise application integration (EAI), supply chain management (SCM), and e-commerce (EC) [2].

Figure 2 shows example in identifying titles, descriptions, and media of presentation slide. The format of presentation slides document used as test data is *.pptx. *.pptx is basically a collection of XML files compressed into a zip file. Each XML file represents a slide in a presentation slide in *.pptx format. XML files can be extracted to be identified further and can be mapped into ontology [2].
2.2. The Method
This study has several stages in the implementation include: (1) adding of summarization feature on search engines; (2) testing and evaluation of summarization experimental results; and (3) reporting research results. This part will focus on three research stages as shown on Figure 3.

![Additional Features on Search Engines](image.png)

**Figure 3.** Additional Features on Search Engines.

2.2.1. Adding of Summarization Feature on Search Engines
At this stage, the addition of summarization features on ontology-based search engine LO, the documents that are used as ingredients of summarization are the top 10 search results. Automatic summarization based on user-query the basis of summarizing, for the purpose of summarization done is seeking information related to the user's wishes are taken from the queries that users enter into search engines LO. Adding of Summarization Feature on Search Engines can be seen in Figure 4.

After searching process with LO search engine is completed, there are four core steps: 1) indexing LO from top 10 search result documents, 2) selection of LO summary, 3) sorting LO result of summary process based on its original position in document, and 4) show the summary. The flow of the summary process can be seen in Figure 5.

Each LO have distance from the other based on the maximum value of similarity, i.e., 1 minus the value of the resemblance. LO can take the form of text, images, audio, video or animation, but in calculating the distance between LO based only on the content of the text is in it, for example if in one LO has an image of a computer and the components in it and a text image name "computer image" it is used as a determinant in the calculation of the distance with the other LO is the name of the picture only "computer image".

Figure 6 shows a branch of a tree cutting techniques (dendogram). Cutting a branch of a tree is determined by the division coefficient values as shown on the left. In this figure also shows an example of cutting dendogram based on context-dependent of users [7].
Figure 4. Additional Features on Search Engines.

Figure 5. Flow Diagram of Summarization.

Figure 6. Context-Dependent Hierarchical Agglomerative Clustering Technique [7].
2.2.2. Testing and Evaluation of Summarization Experimental Results

Summarization experimental evaluation process is done manually by comparing the experimental results with a summary of the summary made by the human user as a comparison. Summarization manually made by people who have competence in the field of computer science for the test data in this study were associated data learning materials in the field of computer science. The result of the manual summarization is a collection of LO which is a collection of slides contained in the powerpoint or document which will be summarized.

To assess the summarization experimental results based on comparison summarized results with a summary of the manual then calculate precision and recall value of each summarization experimental results. After that, a summarization experiment to compare one with the other used F-Measure value calculation. Based on consideration of the importance of the accuracy of the election LO summary F-Measure value calculation in this study involves assigning weights doubled in value precision.

2.2.3. Reporting Research Results

At the reporting stage, researchers put the entire process from the initial research to the final stage. Components that are reported in the form of the activities performed during the study, namely: adding of summarization feature on search engines, testing and evaluation of summarization results as well as conclusions.

3. Results and Discussion

This section explains about the results and discussions of the study.

3.1. Adding Summarization Feature on Search Engine

Documents which will be summarized in this study are the top 10 search results document. A summary consists of a collection of Learning Object (LO) found on top of the document by using extractive summarization method.

In this study, Learning Point (LP) results in this study was defined as LO and implemented into the code in the Java programming language into a node are then grouped and be part of the summary. The stages in the manufacture of a summary of this research, namely: LO indexing, determining the distance between LO, LO election summaries by several methods, sequencing LO based document id and LO position on the original document.

3.1.1. Random LO Selection

After all of the top LO then the indexed documents are drawn at random as the value of the compression rate summary specified. In this research, three summarizations with random LO selecting which compression rate of 10%, 15%, and 20% of the total number LO obtained from the search results.

3.1.2. HAC Method LO Selection

Selection is based on the grouping LO using methods Hierarchical Agglomerative Clustering (HAC) with cutting dendrogram only on the part corresponding to the user's query and the context of the electoral process is cutting dendogram LO performed on five coefficient distances between LO, namely 0.3, 0.4, 0.5, 0.6, and 0.7.

LO grouping process with HAC method using hierarchical clustering support library-java with some modifications and adjustments as well as additional functions for the determination group and cutting dendogram. In the process of classifying the HAC, the method used is the average linkage method, namely the determination of the group based on the average distance from members of the other group members.
3.1.3. K-Means Method LO Selection
In this study, summarization using LO election with K-Means method performed by three variations of the number of groups according to the degree of compression of the summary by 10%, 15%, and 20% of the total LO indexing results. Centroid determination early on K-Means clustering method is determined randomly and the maximum number of iterations in the process of classification is 10 iterations.

3.2. Summarization Experiment
Summarization results will appear on the user's query and LO resulting from the summarization (see Table 1). Each LO displays the title and the contents in the form of text or images in accordance with the contents of the LO.

![Table 1. Query List.](image)

| No. | Query                          | Symbol |
|-----|--------------------------------|--------|
| 1   | Supply Chain Management        | Q1     |
| 2   | Enterprise Application Integration | Q2     |
| 3   | E-Commerce                     | Q3     |
| 4   | Perancangan dan Pemrograman Web | Q4     |
| 5   | Computer Networks              | Q5     |

After the experimental results compared with the gold standard summarization then calculated the value of precision and recall summarization results. After that, an experiment to compare one with the other summarization used F-Measure value calculation. Based on consideration of the importance of precision in the manufacture of the summary calculation F-Measure value in this study involves assigning weights doubled in value precision. Figure 7 shows example of summarization results.

![Figure 7. Interface Example of Summarization Results.](image)

Figures 8, 9, 10 shows F-Measure value from summarization with Random LO Selection, HAC LO Selection and K-Means LO Selection. Fig. 11 shows that summarization with HAC method decisively superior to every query number. When viewed from their mean, the average value of the experiment HAC is 0.487 with the highest value at query experiment number 1 which has a value of F-Measure reached 0.768.

Based on analysis of queries and documents related to the search results that are used as ingredients in summarization can be seen that the search results heavily influence result of summarization. Good search results will get results LO considerable and are related to the search context.
In the search results Q3 and Q4, LO coverage results are still lacking so much so that the summarization result is not optimal. Based on the analysis of the outcome document number from search results in Q3 and Q4, the number of search results that are slightly due to the material "e-commerce" and "design and web programming" has a lot of LO that do not contain the word search query.

Meanwhile, LO from Q1 and Q5 search result despite quite a lot but the LO many are outside the context of search. This occurs because the query that is used, the words "enterprise application integration" and "computer networks" contains a common word in which the word is also found in many other LO course material.

4. Conclusion
Learning content or LO are granular and can be reused for constructing new learning materials. LO from ontology-based search results can be used as an ingredient to create new learning materials according to the topic searched by users. The experimental results of this study indicate that summarization method using context dependency with HAC is better than summarization using the K-
Means. This is indicated by the F-Measure mean value of the summarization using HAC with context dependency better than a mean value of F-Measure with K-Means method.

References

[1] Hewlett Foundation. OER. [Online]. Available: http://www.hewlett.org/Programs/Education/OER

[2] Paramartha A G 2014 Teknik Penelusuran Learning Object Berbasis Ontologi (Depok: Tesis, Universitas Indonesia).

[3] Noor S F, Yusof N and Hashim S Z 2011 Creating granular learning object towards reusability of learning object in e-learning context. Bandung: International Conference Electrical Engineering and Informatics (ICEEI) pp. 1-5.

[4] Meena Y K, Jain A and Gopalani D 2014 Survey on graph and cluster based approaches in multi-document text summarization. Jaipur: International Conference Recent Advances and Innovations in Engineering (ICRAIE) pp. 1-5.

[5] Pandit S R and Potey M A 2013 A query specific graph based approach to multi-document text summarization: simultaneous cluster and sentence ranking. Katraj: Machine Intelligence and Research Advancement (ICMIRA) pp. 213-217.

[6] Gupta V K and Siddiqui T J 2012 Multi-document summarization using sentence clustering. Katharagpur: Intelligent Human Computer Interaction (IHI) pp. 1-5.

[7] Shepitsen A, Gemmell J, Mobasher B and Burke R 2008 Personalized recommendation in social tagging systems using hierarchical clustering. New York: ACM pp. 259-266.

[8] Ooi J M 2015 A survey of query expansion, query suggestion and query refinement techniques. Kuantan: Software Engineering and Computer Systems (ICSECS) pp. 112-117.

[9] Chauhan R, Goudar R, Sharma R and Chauhan A 2013 Domain ontology based semantic search for efficient information retrieval through automatic query expansion. Gujarat: Intelligent Systems and Signal Processing (ISSP) International Conference pp. 397-402.

[10] Grimm S 2010 Knowledge representation and ontologies. Berlin: Scientific Data Mining and Knowledge Discovery pp. 111-137.

[11] Hasibuan Z A and Cenka B A 2012 Toward indonesian open educational resources. Bali: International Symposium on Open, Distance, and E-Learning (ISODEL)

[12] Learning Technology Standards Committee. Draft Standard for Learning Object Metadata. New York: Institute of Electrical and Electronics Engineers (IEEE), 2002.

[13] Manning C D, Raghavan P and Schütze H 2009. An Introduction to Information Retrieval. (Cambridge: Cambridge University Press)

[14] PadmaLahari E, Kumar D V and Prasad S 2014 Automatic text Summarization with statistical and linguistic features using successive thresholds. Ramanathapuram: IEEE pp 1519-24.

[15] Princeton University. 2010. About WordNet. (Princeton University)

[16] Rehak D R and Mason R 2003 Engaging with the Learning Object Economy, in Littlejohn, Alison, Reusing Online Resources: A Sustainable Approach to E-Learning ( London: Kogan Page) pp. 22-30.

[17] Wang S, Li W, Wang F and Deng H 2010 A Survey on Automatic Summarization. Kunming: IEEE pp. 193-196.

[18] Wiley D A 2000 Connecting Learning Objects to Instructional Design Theory: A Definition, A Metaphor, and A Taxonomy.

[19] Santos H B, Cenka B A, Priyogi B, Paramartha Y and Hasibuan Z A 2015 Platform development and design evaluation of the indonesian open educational resources. Hangzhou: The 23rd International Conference on Computers in Education (ICCE)