Documentation Similarity System

Sutthikarn Bojukrapan* Khunathorn Thongnopphakhun Nininnate Kothutsadee
and Panawat Khongtanakunbawon
Department of Computer Science and Information Technology , Faculty of Science
Udon Thani Rajabhat University
*Corresponding author: 64 Thaharn Road, Muang, Udon Thani Province 41000 E-mail address:
sutthikarn@udru.ac.th, Tel.: 087-4909377

ABSTRACT

This research aims to develop and evaluate a similar document detection system in web application. The purposes of this research was 1) to study and develop an algorithm for cutting Thai word by using rule-based rule to be more effective and 2) to compare the similarities of text on web application. The tools in this research included with Visual Studio Code, Sublime text 3 and php language compare and analyze the similarity of text with Thai language abstracts based on student projects in Computer Science and Information Technology of Udon Thani Rajabhat University. The system was evaluated using Black Box testing.

The results of this research was found that the satisfaction of the experts are good ($\overline{X}$ = 4.80, SD. = 0.68) the satisfaction of the users is good level ($\overline{X}$ = 4.70, SD = 0.77). The accuracy of system is the highest satisfaction score.

Keywords: Similarity, Web Application, Cutting Word, Thai Abstract Similarity

INTRODUCTION

At present, the dissemination of academic information contributes to the advancement of science and technology. It is also a knowledge transfer to lead the development of innovation and development. Many academic articles are published by various media such as newspapers, magazines and journals or published in electronic format through the Internet. The students and researchers can find information over the Internet and access more resources the knowledge, concepts and theories in the research or in the articles from research to study and to develop their own research to increase efficiency. It is considered correct and is openness of information dissemination to the advancement of the academic and lead to the development of sustainable knowledge [1].

Words segmentation is important in processing for classification purposes. The Thai language category is very effective. The basic problem is the writing style. Thai is written as a string. No punctuation marks, such as English, use spaces between French words and other languages with spaces. The paragraph between the words clearly. This is a problem in the conversion of Thai language. The computer can recognize Thai as human. This is one of the barriers to division. Three main principles are rule based approach, dictionary approach and Corpus based approach [2].

Natural language processing is one of the fields in computer science that utilizes machine learning and artificial intelligence techniques to process various languages. It can be used to translate sentences from one language to another. It is an emerging area of computation which facilitates easier and refined communication in different languages [3].

N-gram model is one of the most widely used tools for language modeling in the field of Natural Language Processing. Basically, it works by processing the previous words in a sentence. N-gram acts as a window of “n consecutive words in a sentence”. According to standard N-gram model, probability of occurrence of a given word depends only upon the previous (N-1) words. Hence, maximum likelihood estimation for N-gram sequence probability can be given mathematically as described in [4]:

$$P(W_i|W_{i-n+1} \dots W_{i-1}) = \frac{C(W_{i-n+1} \dots W_i}{C(W_{i-n+1} \dots W_{i-1}}$$

Where C (.) represents the count of particular n-gram sequence in the training corpus.

In [5], author mentions three possible techniques for sentence generation namely rule based technique, corpus based technique, and template based technique. In corpus based technique, language rule and information is gathered from a given corpus (training data) and this information is used further in the actual system [6]. This forms the base of style categorization phase in our work.

Thai cutting word is developed in various forms. The main methods used are rule-based, dictionary, and use of the Corpus. Using the dictionary and Corpus are appropriate to apply to the rules because high speed correct post-cutting accuracy at high levels, but at the level of words is not good. Since the use of rules is the use of the pool principle consonants and canals can only be made in short words. Words cannot be cut with multiple words. How to use Dictionaries and the use of archives. Considering that the rules still apply, most of them are still in use. Manage unrecognized words and non-dictionary words. At present, there are more and more words from foreign languages. Thai language does not match the Thai grammar, according to the principle of transcription in Thai language. Therefore, the development of word rules is more effective. It supports word wrapping in other ways either using a dictionary or a news archive [7].

Therefore, the researcher developed the technique used to check the similarity of the text on web application development Calculate the similarity of the text and measure it as a percentage. The text used in the abstracts in Thai language is a way of measuring the similarity or relationship of a document. With word wrapping by using the THSplitLib [8] library, the array (Array) and the rule base come to the aid of the analysis. How to compare the similarities of the original file selection in the database, wrap the words and then store on the database, then select the abstract file to compare. The selected file from the computer will then cut the word and cut the link to check it, then compare it to the original file.

The objectives this research to study and develop an algorithm for cutting Thai word by using rule-based rule to be more effective and to develop and compare the similarities of text on web application.

MATERIALS AND METHODS

Web application development tools include programming using Visual Studio Code, Sublime text 3, and php language. The system development cycle (SDLC) has been studied for its effective development. Analyze and design web applications with architectural structure as shown in Figure 1.

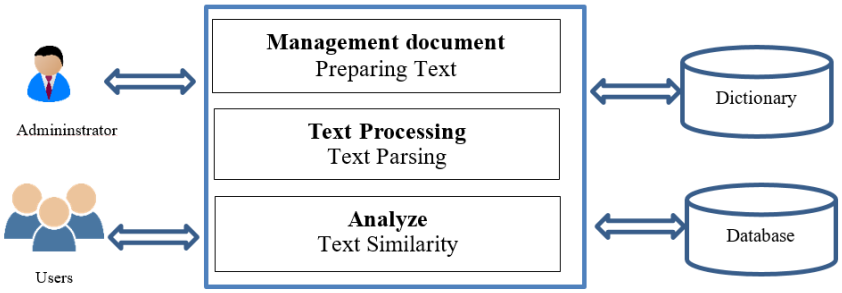

Fig 1. The process of cut word

1. Document management was prepared the abstract text ready for processing. The abstract is

available in Microsoft Word (docx) and Portable Document Formats (pdf).

2. Text processing was text parsing is part of word wrapping. To wrap words in sentences using

THSplitLib.

3. Analyze the text in the database to compare the text for similarity. Analyze and calculate the

similarity of the text. Using the formula as equation 1.

$$Result = \frac{numx100}{X} \quad (1)$$

When *Result* is Number of words in the file copy and *x* is number of words in the source file.

The system is a web site for checking the similarities of the text. The information used in the test is Thai Abstract. The system can check the copy of the message. The user can see how much of this abstracts correspond to the selected abstracts. Users can check the similarities of the document without having to subscribe and download the document in pdf format. If users want to upload information, they should register before using. This research uses the THSplitlib program, which is an open source program. The Thai word wrapping principle is used dictionary compare word wrap with words stored in dictionary. The work is shown in Fig 2.

.

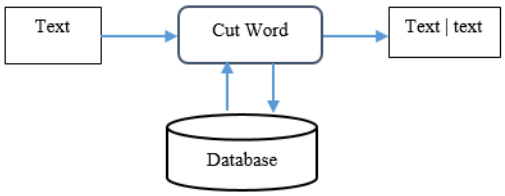

Fig 2. The process of cut word

From Figure 2. word wrapping is used to identify key words of the sentence analysis.

The operation of the system is as follows.

1. Upload the original abstracts into the system and then cut out the words and store them in the database.
2. Run the file you want to compare to the original.
3. Bring the original abstracted text files to the database and the files to compare, then compare the data.
4. Display the word data, cut off and resemble the similarity.

The overview of word wrapping and similarity of text.

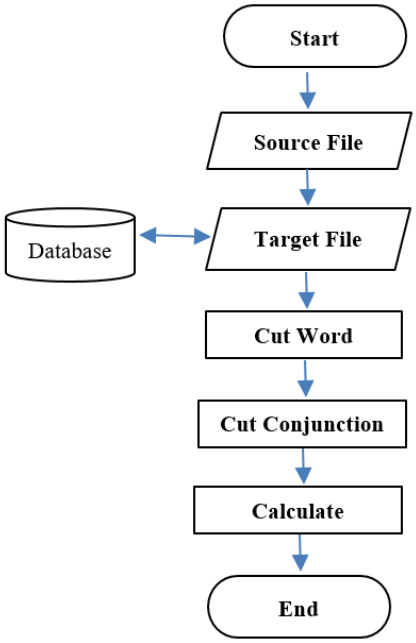

Fig 3. The system process

In the Figure 3. the operation of the system is as follows.

Upload the source file and the target file want to compare to the original. Bring the original abstracted text files to the database and the files to compare, then compare the data. Cut the word by using the THSplitLib program. To calculate the similarity of the source file and the file to check similarity.

RESULTS

To compare the loosening of a message. Can be added to the compare screen will be added to select the text file in the database. When selecting a file to measure loosening. Then hit the plagiarism Checking button to compare the words as in Figure 4.

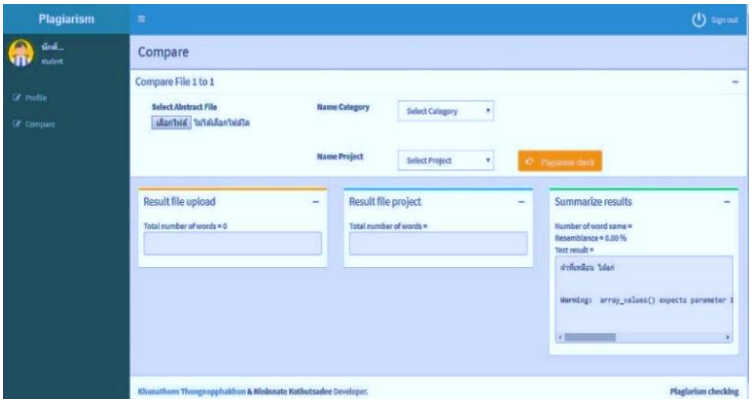

Fig 4. Screen of select the file you want to compare.

Compare the words and then display the results and the number of words and it will show up as a percentage as shown in Figure 5.

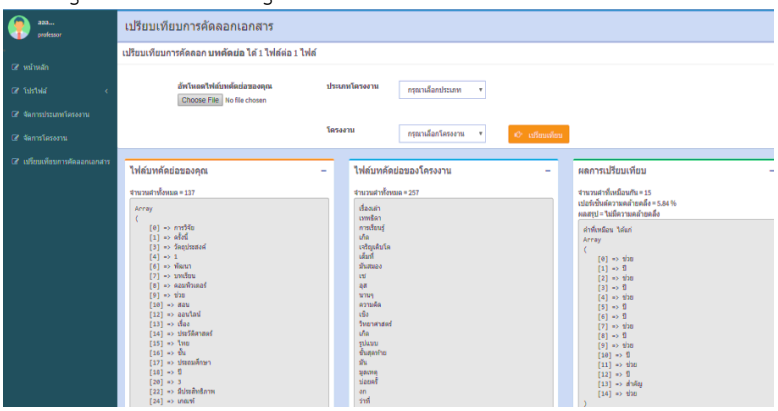

Fig 5. Comparison screen

The experiment to comparison of web text similarities with the user group involved in the system and those who want to use to evaluate the satisfaction of use. The results of the evaluation are summarized in Table 1.

Table 1. Summary of the performance evaluation of the system from the user.

Evaluation Items	Users		Experts	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1. Evaluation of the system capabilities of the system directly to the user requirements.	4.10	0.66	4.00	0.63
2. Evaluation of the system for the accuracy of the work system.	4.70	0.60	4.80	0.52
3. Evaluation of the system's difficulty in using the system.	4.10	0.70	4.20	0.75
4. Evaluation of the system of processing the system.	4.00	0.87	4.20	0.98
5. Security systems assessment system security	4.30	0.63	4.40	0.52
Total	4.70	0.77	4.80	0.68

CONCLUSION AND DISCUSSION

The information that will be useful to use the technology to assist in the operation. Since the collection and verification of data processing into information and maintenance of information for use. The text comparison system prepared to collect student project of Computer Science and Information Technology Udon Thani Rajabhat University. This working on the web using a database management system makes it possible to control duplication. They also have problems sharing information and copying of others into their own or copyright infringement. The idea is to develop a technique to monitor documents by means of calculating the similarity or similarity of documents in the information system or document relationship.

This system was developed to verify the similarity of the document. It works to find the frequency of words and duplicate words to check the similarity of the project. So the next step is to be able to check the similarity of the sentence and the source of the sentence.

REFERENCES

[1] SutinAhingsaro. (2011). **Detecting Plagiarism Using Document Similarity Method**. Information Technology King Mongkut's University of Technology North Bangkok.

[2] ChattiyaSaksomboon. (2011). **Automatic Use Case Diagram Creation from Scenarios in Thai Language Using Document Frequent and Vector Space**. Computer Sciences King Mongkut's University of Technology North Bangkok.

[3] Ashwini I Gadag and B M Sagar. (2016). **N-Gram Based Paraphrase Generator from Large text Document**. International Conference on Computational Systems and Information Systems for Sustainable Solutions. P92-94.

[4] F. Peng and X. Huang. (2007). **Machine learning for Asian language text classification**. Journal of Documentation, vol. 63, no. 3, pp. 378-397.

[5] N. Ito and M. Hagiwara. (2011). **Natural language generation using automatically constructed lexical resources**," in The 2011 International Joint Conference on Neural Networks (IJCNN), pp. 980 - 987.

[6] A. K. Yadav and S. K. Borgohain. (2014). **Sentence generation from a bag of words using N-gram Model**. International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), May 2014, pp. 1771 - 1776

[7] KandaSaikaew and PayothornUrathummakun. (2006). **Thai word wrap by updating rules and new dictionaries**. Department of Computer Engineering Faculty of Engineering: Khon Kaen University.

[8] SuwichaPeangim. (2513). THSplitLib Program. April 1, 2015, from http://www.alogik.com/thsplitlib/