International Journal of Latest Trends in Engineering and Technology
Vol.(11)Issue(1), pp.074-080
DOI: http://dx.doi.org/10.21172/1.111.14
e-ISSN:2278-621X

EXTRACTION OF INDONESIAN FIELD-ASSOCIATED TERMS FROM JAPANESE FIELD KNOWLEDGE

Samuel Sangkon Lee[1]

Abstract- Field-associated terms refer to intuitive words which remind certain field of the document. These play roles as the clues for figuring out the field of the document quickly by using information of few words including common sense knowledge through extraction of field-associated terms within the document. Automatic document classification would be possible through document classification system while establishing a large field-associated term dictionary. A document field extraction system reads certain type of document randomly and extracts the most associated document field. This study has been conducted to classify document written in Indonesian language by using Japanese associate knowledge. The result from experiment on the field of politics showed 88.8% and 34.7% of precision and recall ratios respectively. The reason for low recall ratio may be that analysis has been carried out taking into account political environment of Indonesia. In this study, we will attempt to expand this field to classification of field-associated terms of various languages based on Japanese associated knowledge.
Keywords – Field-associated Term, Level of Field-associated Term, Theme Field, Indonesia Document Classification.

1. INTRODUCTION

In today, computer is utilized for various usages in numerous fields in the society including financial institution, SNS and companies' system as a result from miniaturization of computer hardware, its large capacity, low price and high-speed LAN. Computer has various merits that it can search large volume of information and delete large amount of unnecessary data. Accordingly, the number of documents digitalized by using computer is increasing a lot. Further, the need for computer software, which are for automatic document processing by using computer, is increasing and automatic classification technology of large-volume documents according to their fields is an example [1].

Document classification is a measurement for automatic classification of documents according to their contents [2]. In general, document classification is carried out by using statistical measurement such as machine learning [3]. For using the statistical measurement, however, large amount of documents as learning data is required and the content must be accurate. Field-associated Terms (FT), by which Tusji proposed, is a document classification measurement which does not use statistical measurement [4]. A field-associated term refers to an intuitive word [5-7] which reminds certain field of the document. This word helps figuring out the topic of the document by extracting field-associated term. The system for extracting theme field of the document by using dictionary, which is composed of large amount of field-associated terms, is one of the electronic document classification technologies.

The objective of this study is to automatic establishment of field-associated term dictionary, which is for classification of document written in Indonesian language, based on Japanese associated knowledge, which has been studied for a long time since 2002. The rest of the paper is organized as follows. In section 2, definition of field-associated term and document classification measurement will be examined as well as their problems will be discussed. In section 3, the measurement for establishment of field-associated term dictionary, which was proposed in this study, a described, and evaluation experiment will be conducted and the reliability of the measurement proposed from the experiment result will be verified. In section 4, conclusion of this study will be examined.

2. FIELD-ASSOCIATED TERM

2.1 Definition of Field-associated Term

In this section, the definition and overview of field-associated term are examined in details. Field-associated term refers to the words by which human can remind of certain field within the document. For example, field-associated terms of </Basketball> in basketball document include common nouns, such as “give and go (Play in which two players go forward exchanging ball. In basketball, a player passes ball and moves around to leave the defender, receive return pass and shoot the ball [Wikipedia].,” “travelling(Foul in which the player moves more than three steps holding ball. In basketball, traveling is a violation of the rules that occurs when a player holding the ball moves one or both of their feet illegally [Wikipedia].),” and names of personnel or organizations such as "Michael Jordan" and "LA Lakers (NBA basketball team)."

[1] Professor, Department of Computer Science and Engineering, Jeonju University, South Korea, 55069

Table 1 : Associated Field and Field-associated Term

Field	Field-Associated Term (Scores)
<Basketball>	Give and Go (80), Side Hand Pass (80), Japanese Basketball Association (60), FIBA (40), Travelling (30) etc.
<Fashion>	Best Dress Award (80), Fashion Show (60), Christian Dior (60), Japanese Jean Association (50) etc.
<Economy>	Economy of Japan (60), The Bank of Japan (40), Strong Yen (30), Yen Exchange (30), Prices (30), Financial Market (30), Bank Rate Policy (30) etc.
<Politics>	Diplomatic Policy (80), Congressman (40), Representative (40), Senator (40), Unicameral System (30), Standing Member of Committee (30), Finance (30), Nation (30) etc.

However, words including "case" or "usage" do not remind certain field therefore they cannot be used as field-associated terms. FTs are established according to pre-defined field system therefore they are registered in FT dictionary. The registered field-associated terms are assigned with scores for the field and the values can be set according to the level of association. Table 1 examines examples of scores for field-associated terms. In the Table 1, "Economy of Japan" and "Financial Market" are field-associated terms of <Finance>. However, "Economy of Japan" only reminds <Economy>. Therefore, the score of "Financial Market" is low at 30 points. Further, FTs consist of single (simple word) or compound words. The simple word here means the word registered in morpheme dictionary [9] while compound words means the words which are composed of more than two single words. Let us call the FTs composed of single word and compound words as single and compound field-associated terms respectively. For example, the aforementioned "Travelling" is a single FT term while "Japan Basketball Association" is a compound one composed of "Japan", "Basketball" and "Association"[10-11]. In this study, field names including basketball are written in < > for example <Basketball> and field-associated terms including travelling are marked in “” for example “travelling."

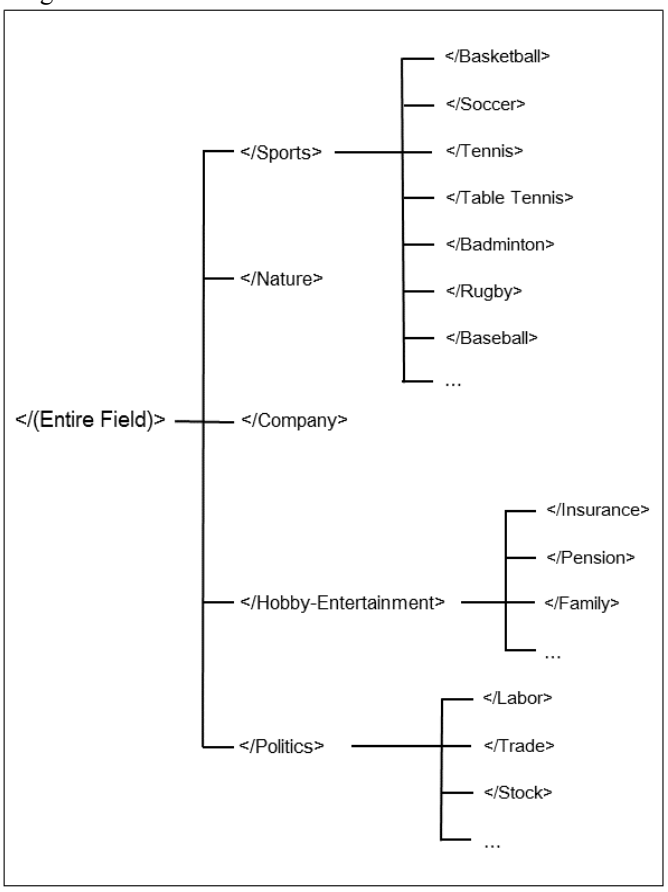

Figure 1: Example of Field Scheme

2.2. Field Scheme

Field scheme refers to the collection of a certain field. This system expresses upper and lower relation as a tree structure. This tree structure is called field tree. The theme field, which is the leaf of field tree, is called terminal field. In addition, the other nodes are all called intermediate field [6]. Moreover, the upper field of certain field, which has direct upper and lower relation, is called parent field while the lower one is called child field. An example of field tree is examined in the Figure 1. For instance, </Sports>, </Life> and </Economy> are called intermediate field while </Baseball>, </Insurance> and

</Stock> are called terminal one. Field designation is described as field path <S> and is omitted from </Entire Field> which is the root (/ or /Entire Field). Further, if there is no certain contradiction, entire path designation is omitted and the field of the document is described as terminal field. For example, this is described as field path <S> = </Sports/Rugby> is described as </Rugby> which is the childe terminal field of </Sports>.

2.3. Level Setting

There are differences in the range of FTs. Field-associated terms of certain field has terms associated with single terminal or intermediate fields. In contrast, the other FTs have terms for compound terminal or intermediate fields. Therefore there can be difference in range of association of field-associated terms. Accordingly, the levels of association can be classified in to five ones according to the range. Level 1 FT means the term which is associated with terminal field </Sumo> at once as with "Yokozuna." Level 2 FT means the words which are associated with two or more terminal fields including </Tennis>, </Table Tennis> and </Badminton> as with "single" or "double." The level 3 intermediate field-associated term means the words which are not associated with terminal field, such as "match," but one intermediate field. Level 4 multiple FT means the words which are associated with two or more terminal field </Hobby-Entertainment/Japanese Chess>, such as "victory or defeat" or intermediate field </Sports>. Level 5 none FT means the words which are not limited to certain field including "case" or "usage." Table 1 indicates major field-associated terms and their levels.

[Definition 1] Levels of FTs
(Level 1) Complete Field-associated Terms: words which are associated with only one terminal field
(Level 2) Half Complete Field-associated Terms: words which are associated with terminal field with the same parent field
(Level 3) Intermediate Field-associated Terms: words which are associated with only one intermediate field
(Level 4) Multiple Field-associated Terms: words which are associated with multiple intermediate or terminal fields
(Level 5) Non Field-associated Term: words which are not associated with any field as one field

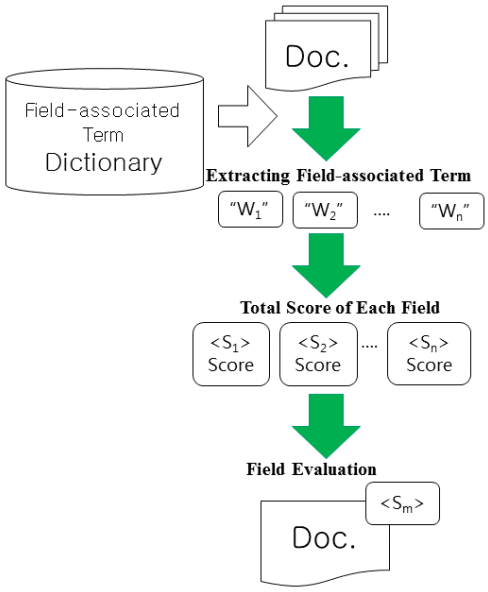

Figure 2 : Control Flow of Classification

2.4. Classification by using Field-associated Terms

Document classification by using field-associated terms can be explained as follows. The control flow of document classification is indicated in Fig. 2. In this figure, the processes can be classified into three steps as follows: [Stage 1] Extraction of FT, [Stage 2] Total Score of Each Field, and [Stage 3] Field Evaluation. The following is the details of the aforementioned three classification processes [4].

[Stage 1] Extraction of Field-associated Terms

Extract all field-associated terms included in the document by using field-associated term dictionary. For example, the field-associated terms indicated in document. Further, the fields of extracted field-associated terms and scores are indicated in the Table 2.

[Stage 2] Scoring for Each Field

As shown in [Stage 1], all of the FT extracted from the document cannot be associated with one field. Accordingly, score assigned to FTs need to be aggregated for each field to evaluate which field is the most appropriate. If field-associated term for association of a field $< S >$ is extracted as $n\,(n \geq 1)$ the score of $< S >$ field is calculated from the following formula (1)

by setting field-associated term as $w_i(i=1,2,\ldots,n)$. For simplifying explanation, we do not consider the times field-associated terms appear.

$$<S>'s\ Score = \sum_{i=1}^{n} (Score\ assigned\ to\ w_i) \quad (1)$$

[Document A]
Labor unions of steel and iron, ship building and marine transport and nonferrous metal, which joined Japanese key industry labor unit alliance (key labor unit), the industry unit of each major industry, submitted requests for wage increase or labor condition improvement on 12th and initiated conflict mediation in earnest in Spring in 2016
In the head quarter of ISIJ, which is the largest steel and iron company, president ○○ Ichiban announced that 「we request for support for improvement of labor conditions of all employees work in the group or related companies for improvement of productivity and reinforcement of international competitiveness」 on 10 in the morning. Further, he sent a request for wage improvement (total 80 million JPY) within 2 years to director ○○ Ichiban.
The performance of steel and iron industry is quickly decreasing due to economic shrinkage of China. The management team of ISIJ announced that 「the situation is getting worse and we cannot accept the policy for constant increase in wage」 and an important idea about low wage increase.
After that, labor units of automobile and electricity companies submitted request as well and are still negotiating for conclusion in at the meeting in the middle of March.

(Underlined Words : Field-associated Terms)

Figure 3 : FTs Recognized in document (A)

Table 2 : Total Scores

Field	Score
</Business>	350
</Economy>	60
</Ship>	30
</Steel Raw Material>	30
</Automobile>	30

[Stage 3] Field Evaluation
The field with the highest score is set to be the document's field from document A. Further, in Table 3, the field of a document becomes </Business>. In case of multiple fields with the highest score, multiple fields are set to be the document's field. Further, the case in which field-associated term is not extracted from the document at all is set as </Field Neutral> [7].

2.5. Automatic Establishment of FT Dictionary
For establishing field-associated term, large-scale documents collected from each terminal field of a field scheme are used as learning data. Frequency of appearance of each noun is calculated by morphological analysis regarding the learning data. After that on which each noun appears intensively is calculated to set single field-associated term and the level. The procedure for setting single field-associated term is indicated. In single field-associated term decision algorithm, the frequency of appearance of the words collected from learning data is used. However, it is very difficult to collect learning data equally for each terminal field. Accordingly, for frequency of the word of terminal field <S> is indicated as F(w, <S>) by defining the frequency of all words appeared in terminal field <S> as T(<S>). Further, N(w, <S>) which is the normalized value as the formula (2) is used.

$$N(w,<S>) = \frac{F(w,<S>)}{T(<S>)} \times \gamma \quad (2)$$

$\frac{F(w,<S>)}{T(<S>)}$ is very small value therefore $N(w,<S>)$ is adjusted as a whole number through appropriate a constant number. In addition, which is the frequency of word $N(w,<S'>)$, for intermediate field $<S'>$, is calculated as sum of frequency $N(w,<S>)$ of all terminal field <S> which exist under. If field $<S'>$ is set as the parent field of field $<S>$, the concentration of words for field $<S>$ is defined as in formula (3).

$$C(w,<S>) = \frac{N(w,<S>)}{N(<S'>)} \quad (3)$$

2.6. Level Decision Algorithm of Single FT
Single FT Decision Algorithm
Input: Word w, N(w, <S>) for each field <S>, field tree.

Output: The field and level of field association when w becomes the field-associated term.
Stage 1: Deciding Complete FT (Level 1)
Whether the word w is concentrated on certain field or not regarding parent field <P/C> of field scheme's root <S> = </Entire Field> is decided by using formula (4) by selecting threshold.

$$P(w,<P/S>)\geq\alpha \tag{4}$$

If this condition is satisfied, change <P/C> to <P> and repeat the evaluation with the same measurement in the parent field. If <P/C> becomes the terminal field by repeating this procedure, w is decided to be the complete field-associated term of field <P/C>. If there is no parent field <P/C> of <P>, which satisfies the conditional formula through this process, [Stage 2] is used.

Stage 2: Deciding Half-complete FT (Level 2) and Intermediate FT (Level 3)
Extract <P/C>, which is equal to

$$P(w,<P/C>)=\frac{N(w,<S>)}{m} \tag{5}$$

from $m\geq 2$ parent field <P/C>. Then the values are accumulated and added in a descending order from $P(w,<S>)$ and $k\,(1<k<m)$ ones are added. If the first total value exceeds α and if k child field <P/C> is all terminal one, w is decided as half-complete field-associated term of field <P/C>. If all of them are not terminal ones, the following procedure is used. If, however, the accumulated value added does not exceed α, field w is decided as intermediate field-associated term of field <P>.

Stage 3: Deciding Multiple FT (Level 4)

Extract terminal field <P/C> from k child field <P/C> and set w as multiple field-associated term of field <P/C>. Change child field <P/C> apart from terminal field to root <P> of field and set w as multiple field-associated term for the field of complete, half-complete and intermediate field-associated terms decided in stage 1 and 2.
This indicates example of deciding field-associated term by using single FT algorithm. Frequencies of each field of “Yokozuna,” “singles,” “matching,” and “victory and defeat” are indicated in parenthesis. Further, the number of child field of </Entire Field> while the number of </Sports>, </Hobby-Entertainment>, </Politics> are 19, 13 and 14 respectively and threshold set. The process in which this algorithm is applied in actual system is examined by using the following four examples: [Example 1] Process for deciding field of "Yokozuna" which is a candidate of FT, [Example 2] Process for deciding field of "singles" which is a candidate of FT, [Example 3] Process for deciding "match" which is a Candidate of FT, and [Example 4] Process for deciding "victory and defeat" which is candidate of FT. These four examples will be explained at my presentation date.
Bing Search is a search engine by which Microsoft provides. In the past, Bing Search is a service which was opened to the public as the name of Live Search, Windows Live Search or MSN Search. Bing Search was opened through introduction by CEO of Microsoft on 28th of May in 2009 and is providing Bing Search API. In the present study, sentence collection program was produced by using this service. In the present study, FT is set as search keyword and entered in Bing Search API and the first page of the search result are stored. Before document collection, FT candidates are collected to prepare FT candidate lists. The reason for choosing the 1st page is that association level increases with decrease in the number of searched pages when entering certain search keyword in search engine.

3. EXPERIMENT AND RESULT

3.1. Preparation for Experiment

In this experiment, titles of each page were all collected in database of Indonesian Wikipedia [8] as candidates of FTs. The total number of this FT candidate was 2,564. Keywords are entered in Bing Search API and searched through Bing Search API and the first abstract is stored as text file. The result from Bing Search API is entered in Microsoft Translator API then is translated into Japanese. Finally, the sentences translated into Japanese are matched to Japanese FT dictionary to evaluate the theme field of document. Currently, the Japanese field-associated dictionary about </Politics> include 5,124 FTs. The output result of our proposed method were compared to the result from evaluation by people then precision and recall ratios were measured.

3.2. Evaluation Measurement

The reliability of our idea is evaluated by measuring precision and recall ratios through comparison on the results from system evaluation and manual evaluation realized according to the measurement for proposal of FT candidate. The criteria for judging field-associated term candidate by human is that the answers belong to one of politician (a personal noun), name of political party of a country, policy of a country and field of politics were regarded as correct answers.
C. Experimental Results
The precision and recall ratios of the experiment result were measured by comparing the results from evaluation as political field by the system and human. Precision and recall ratios were calculated through the following formula:

$$Precision.= \frac{Right\ Answers}{\#No.of\ Right\ Answers\ in\ That\ the\ System\ Judges\ the\ Word\ as\ </Poiltics>}, \quad Recall = \frac{Right\ Answers}{\#No.of\ Answers\ in\ That\ Human\ Judges\ the\ Word\ as\ </Poiltics>}$$

The result shows that precision ratio is high but recall ratio is low. In the next section, the reason for low recall ratio will be explained. The result shows that precision ratio is high but recall ratio is low. In the next, the reason for low recall ratio will be explained.

検索キーワード：Partai Kebangkitan Bangsa（インドネシアにある政党団体）

国民覚醒党（PKB）選挙 INDONESIAPartai の目覚め（PKB）は kiai kiai Nahdlatul ウラマーによって…提案された名前は国民覚醒党、党と党 Nahdlatul ウンマ…さらに、党は国イスラム教徒家に警告するつもり… イマーム Nahrawi 会計…GOLPUT「政治分割. に対するファトゥワ恐怖…インドネシアの国イスラム教イスラム教の独立の布告の国民覚醒党（PKB）

Search keyword : Partai Kebangkitan Bangsa(the Name of Political Party of Indonesia)
Partai Kebangkitan Bangsa(PKB)-Deliberation (PKB) of election Indonesia Partai is kiai kiai Nahdlatul. The name by which Ulama2 proposed ... is PKB and party Nahdlatul Umma3 ... Moreover, the party will warn the houses of Islamic believers ...
Imam4 Nahdlatul accounting ... Golput Part5 for Golput political division and fear Partai Kebangkitan Bangsa (PKB) of report of independence of independence of Islam of Indonesia

Figure : 4 Example of Translation to Japanese

The result from evaluation on the proposal of this study indicated high precision and low recall ratios. The reason for the low recall ratio is described as follows: Political party or group of Indonesia are closely related to certain religion (Islam). Therefore, words related to the religion appear a lot in the abstract on the web page rather than the contents about political party. As a result, many documents are judged as </Religion>.
A politician performs other tasks apart from politics. For example, the politician is a soldier as well as an actor. He is not registered as politician in abstract therefore it is evaluated as the other field. There is a problem that politics of overseas or events of the other countries are evaluated as </Overseas-International> in case of evaluation by using Japanese field-associated dictionary. Further, the reason for low recall ratio is indicated in the Fig. 5. An example of translation of "Partai Kebangkitan Bangsa (the name of a political party of Indonesia)" is indicated in Fig. 4. Further, the example of putting the same keyword on Japanese field-associated dictionary according to the proposal is indicated in Fig. 5.

Search Keyword : Partai Kebangkitan Bangsa

1. </Religion/Islam>, (300)
   Ulama(4, 60), Umma(4, 60), Imam(2, 60), Fatma(1, 60), Islam(2, 30), Islam believer(3, 30)
2. </Overseas-International>, (200)
   banjangra(2, 60), national party(2, 40), Islam party(3, 40), presidential election(1, 30), secretary general(1, 30)
3. </Politics/Politics>, (190)
   Representative(1, 40), politics(7, 30), nation(6, 30), Political power(1, 30), democracy(1, 30), the number of seats(1, 30)
4. </Politics/Election>, (130)
   Election commission(1, 40), election(17, 30), presidential election(1, 30), presidential candidate(1, 30)
5. </Book/Dictionary>, (120)
   ウィキペディア(4, 60), Wikipedia(1, 60)
==========
Field(output): Islam

Figure 5 : Example of Registration in Japanese FT Dictionary

[2] Ulama: The spiritual leader with authority
[3] Umma: Soeur
[4] Imam: Title of spiritual leader of Islam or the person who has high education level in the Islamic society
[5] Fatwa: Opinion or religious authority of jurist Interpretation and application of Islamic law,

4. CONCLUSION

In the present study, the measurement for automatic establishment of field-associated dictionary of Indonesian language by using the measurement for extraction of Japanese field-associated terms was examined. Further, field-associated terms of Japanese language were defined. In addition, various types of field-associated terms were applied to document classification to discuss about problems which can occur during extraction of field-associated terms. Further, related studies were listed as well as the similarity and difference from the present study were examined.

The result from evaluation on the proposal of this study indicated high precision and low recall ratios. The reason for the low recall ratio is described as follows: Political party or group of Indonesia are closely related to certain religion (Islam). Therefore, words related to the religion appear a lot in the abstract on the web page rather than the contents about political party [9]. As a result, many documents are judged as </Religion>.

A politician performs other tasks apart from politics. For example, the politician is a soldier as well as an actor. He is not registered as politician in abstract therefore it is evaluated as the other field. There is a problem that politics of overseas or events of the other countries are evaluated as </Overseas-International> in case of evaluation by using Japanese field-associated dictionary. An example of translation of "Partai Kebangkitan Bangsa (the name of a political party of Indonesia)" is indicated in Fig. 5. Further, the example of putting the same keyword on Japanese field-associated dictionary according to the proposal will be published in Japan.

Further, the reason for low recall ratio is indicated in the Fig. 5. An example of translation of "Partai Kebangkitan Bangsa (the name of a political party of Indonesia)" is indicated in Fig. 4. Further, the example of putting the same keyword on Japanese field-associated dictionary according to the proposal is indicated in Fig. 5.

5. REFERENCES

[1] Yuma Fujita, Yoshiaki Ichihashi, Shunsike Kanda, Kazuhiro Morita, and Masao Fuketa (2016) "Full-Text Search Using Double-Array CDAWG," International Journal of Future Computer and Communication, vol. 5(6), pages 237-240

[2] Samule Sangkon Lee, Masami Shishibori, Totu Sumitomo, and Junichi Aoe (2002), "Extraction of Field-coherent Passages," An International Journal of Information Processing and Management, vol. 38(2), pages 173-207

[3] Fabrizio Sebastiani (2002) "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34(1), pages 01-47

[4] Takako Tsuji, Masami Fuketa, Kazuhiro Morita, and, Junichi Aoe (2000) "An Efficient Method of Determining Field Association Terms of Compound Words. Journal of Natural Language Processing, vol. 7(2), pages. 03-26 (In Japanese)

[5] Masao Fuketa, Sangkon Lee, Toru Sumitomo, and Junichi Aoe (2000) "Determining Text Fields Using Association Words," World Multiconference on Systemics, Cybernetics, and Informatics, Vol. 6(3), pages 79-84

[6] Masao Fuketa, Sangkon Lee, Takako Tsuji, Makoto Okada, and Junichi Aoe (2000) "A Document Classification Method by Using Field Association Words," An International Journal of Information Sciences, vol. 126(4), pages 57-70

[7] Sangkon Lee and M. Shishibori (2002) "Passage Segmentation based on Topic Matter," International Journal of Computer Processing of Oriental Languages, vol. 15(3), pages 305-339

[8] https://id.wikipedia.org/ wiki/ Halaman_Utama by Indonesian Wikipedia

[9] Arifin and Ketut Eddy Purnama (2012) "Classification of Emotions in Indonesian Texts Using K-NN Method," International Journal of Information and Electronics Engineering, vol. 2(6), pages 899-903

[10] Samuel Sangkon Lee, Masami Shishibori, and Chia Y. Han (2013) "An Improvement Video Search Method of VP-Tree by Using Trigonometric Inequality," Journal of Information Processing Systems, vol. 6(2), pp. 315-332