# Extraction of Keywords for Retrieval from Paper Documents and Drawings based on the Method of Determining the Importance of Knowledge by the Analytic Hierarchy Process: AHP

Kohei Arai
Saga University, Saga city
Japan

***Abstract*—Extraction method of keywords for retrieval from paper documents and drawings based on the method of determining the importance of knowledge by the Analytic Hierarchy Process: AHP method is proposed. The method allows distinguish the documents into three categories, letter, form and drawing types of documents, then the most appropriate knowledge about keyword for retrievals, font size, location, frequency of the words etc. are selected for each document type. Production rules are created with more than five of the knowledge on keywords for retrievals. Traditional production system employs isolated knowledge so that it is not easy to take overall suitability of the knowledge. In order to overcome this situation, AHP is employed in the proposed system. Through experiments with 100 documents and diagrams, 98% success rate is achieved, and it is found that appropriate candidates for keywords with likelihood or certainty factor can be extracted with the proposed system. The proposed production system shows 50% of improvement on success rate of the keywords extraction from documents and diagrams compared to the existing production system without AHP.**

***Keywords—AHP method; extraction keywords; production rule system; document/diagram recognitions; certainty factor***

## I. Introduction

Currently, iDC (internet Data Center) development is in progress, which manages a large amount of printed media such as paper and handwritten documents, drawings, etc. in a database so that they can be searched and published on electronic media etc. [1]. Generally, for paper media documents, the operator inserts the search keyword as a handwritten page, converts this page into an electronic medium with a scanner, and registers it.

Regarding the automatic generation of keywords from paper documents etc., the documents are limited to business documents, and are roughly classified into letters and form documents. In the former case, search keywords are used by using the knowledge of the position of the title character string in the layout. It has been proposed that the operator manually extract the data in the form of a table document, etc. [2]-[5]. On the other hand, the method of extracting the search keyword from the digital document after conversion to the electronic medium is as follows: (1) layout related to the position of the document title, (2) character font type, size, (3) appearance frequency, (4) character. There are methods [6]-[12] that use knowledge about the part of speech and importance of the words in the sequence. A method has also been proposed in which a semantic tag is added to a word to associate it with an ontology class and attribute items and to make it a keyword [13].

In this paper, the author classifies paper documents into letters, forms, and drawings, and propose a keyword extraction method for them. That is, a method of converting an image obtained by a scanner into HTML format, extracting layout information, decomposing words in a character string into parts of speech, and extracting a search keyword by using position, font size, and appearance frequency as knowledge. Is proposed. At this time, the importance of these knowledge varies depending on the document format, and in order to take this into account, the hierarchical analysis method (Analytic Hierarchy Process: AHP) [14] is used. Introducing a decision-making method with AHP to grasp a large amount of knowledge in a production system as a knowledge-based system [15] and constructed a system for extracting search keywords from paper documents, etc., and using real documents and drawings. As a result of testing, it was confirmed that the search keyword could be extracted efficiently, so the author reports here.

The next section describes the related research works on extraction method followed by the proposed method for extraction of keywords. After that some experiments are described followed by conclusion with some discussions together with future research works.

## II. Related Research Works

Method for real time text extraction of digital manga comic is proposed [17]. On the other hand, extraction of line features from multifidus muscle of CT scanned images with morphological filter together with wavelet multi resolution analysis is proposed [18]. Meanwhile, method for extraction product information from TV commercial is proposed [19] together with text extraction from TV commercial using blob extraction method [20].

Eye-based human-computer interaction allowing phoning, reading e-book/e-comic/e-learning, Internet browsing and TV information extraction is proposed [21]. Also, method for automatic e-comic scene frame extraction for reading comic on mobile devices is proposed [22] together with method for information extraction from Japan TV commercial [23]. Furthermore, automatic information extraction from on-air TV commercial contents is proposed and well reported [24].

## III. Proposed Method

### A. Basic Ideas of the Proposed Method

In the conventional method, first, an operator extracts a keyword for search based on knowledge and experience, writes the keyword on a separate sheet, attaches it to a target paper medium document, etc., and converts it into an image file by a scanner or the like. After that, character recognition is performed. The target paper media document etc. is registered in the database as an image file or in a text file format after recognition. At that time, a keyword is added as a result of recognizing the handwritten character entered on the keyword form. In addition, the attribute information for retrieval is made into a database as metadata. This method involves operator work, is inefficient, and has the problem that the validity and integrity of keywords based on the subjectivity of the operator cannot be maintained.

The proposed method minimizes the intervention of the operator and is designed so that the keywords for retrieval are automatically extracted from the documents, drawings, etc. existing on paper media, the database is constructed, and the retrieval is possible. First, convert a paper medium document into an image file with a scanner etc., classify the document into letters, forms and drawings according to a uniquely developed program, extract layout and font size information, and perform character recognition. Convert to HTML format by using the result (text format). After that, the text format sentences are converted into "divided" sentences using morphological analysis software [11], [12] ChaSen, and parts of speech are given to the words. ChaSen is a free Japanese morphological analyzer released from Nara Institute of Science and Technology Graduate School of Natural Language Processing on February 19, 1997. It analyzes the input Japanese, decomposes it into words, and returns the reading and part of speech of the words. ChaSen can also be used for full-text search, addition of phonetic alphabets, and extraction of specific parts of speech.

Next, the position of the character string, the font size, and the frequency of appearance of words are checked from the file after HTML conversion, and the most suitable keyword for the document etc. is extracted by the production system. At that time, in the knowledge base system, AHP is used in order to consider that the importance of these knowledge differs depending on the document format.

### B. Document Format

The author selected 100 types of target documents (letter format: 35, form format: 35, drawing format: 30) including paper media business documents and drawings. Drawings can be identified using the number of characters as an index, and the form format that includes many tables has many line segments such as ruled lines and many characters, and the letter format has few line segments and many characters, making the document 3 It can be easily classified into any of the types.

In the case of letter format documents, the document title is often used as a search keyword, and the position where it appears, and the font size are important. Next to them, the appearance frequency of the relevant keyword is important. Also, in form-based documents, titles often appear in the table, the position where the document title appears and the frequency of its appearance are important, and the font size is not so important. On the other hand, in the case of drawings, the position of the document title is the most important, the font size is not so important, and the appearance frequency is not so important because the number of characters is small. In other words, the importance of knowledge when extracting search keywords differs depending on the document format.

### C. Optimal Knowledge Importance Setting for Document Format

Based on the AHP, we examined the method of considering knowledge importance in advance. We optimized the setting of the importance level of the knowledge keyword based on the target document and 100 document formats described above and estimated the evaluation items required for this in advance. AHP is a problem-solving decision-making method that makes good use of subjective judgment and system approach in problem analysis. When one answer must be extracted from intricately entangled elements, there is a risk of overlooking an important element if it is divided too simply, and it is difficult to flexibly use it with an overly complicated method. Become. Therefore, AHP was born as a method for incorporating many elements in a well-balanced manner and making decisions.

AHP expresses the elements related to decisions in a hierarchical structure. Based on a certain criterion, the evaluation of options is judged hierarchically, and finally all layers are integrated to decide. This procedure is shown below.

*1) Extraction of elements of evaluation items and hierarchy of knowledge for decision making:* One element of the purpose of decision making, multiple elements of evaluation items for objective evaluation, and multiple alternatives for the purpose are prepared. In this paper, we set the importance of knowledge (weighting factor) as the purpose, the font size, the position of the character string and the appearance frequency as the evaluation item elements, and the size of these evaluation item elements as the alternatives. The top layer is the target element, the evaluation layer is the evaluation element for decision making, and the bottom layer is an alternative plan.

*2) Evaluation of the degree of influence of elements in the evaluation layer:* The target document is presented to 10 subjects, and the optimum word is selected as a search keyword, and the knowledge used when selecting the keyword, that is, the degree of influence of the evaluation item element (importance) I got a score from 0 to 1.

*3) One-to-one comparison between elements in each hierarchy:* Select a pair for each layer and perform a one-to-one comparison. If there are n comparison elements in the hierarchy, then n (n−1) / 2 one-to-one comparisons will be performed. The elements in the same hierarchy are compared using a one-to-one comparison table, and the one-to-one comparison matrix shown in Table 1 is created. After that, a one-to-one comparison of the evaluation item hierarchy is performed, and the relative importance between the elements is calculated.

*4) Calculating importance between elements in each hierarchy:* Appendix 1 shows the algorithm for obtaining the weighting factors between elements in each layer.

*5) Consistency check in one-to-one comparison:* In the one-to-one comparison, the inconsistency of the results occurs as the number of elements increases. Therefore, it is necessary to check this consistency. This method is shown in Appendix 2.

*6) Calculating relative importance of the entire hierarchy and determining alternatives:* A one-to-one comparison between the objective evaluation items and each alternative is made. Then, using the weighting factors found at each layer, the optimal alternative is found from the alternatives.

TABLE I. ONE-TO-ONE COMPARISON ON THE EVALUATION ITEM LAYER

| | Font Size | Y Position | X Position | Frequency | Relative Importance |
|---|---|---|---|---|---|
| Font Size | 1 | 2 | 3 | 4 | 0.47 |
| Y Position | 1/2 | 1 | 2 | 3 | 0.28 |
| X Position | 1/3 | 1/2 | 1 | 2 | 0.16 |
| Frequency | 1/4 | 1/3 | 1/2 | 1 | 0.09 |

### D. *Construction of Knowledge Base for Search Keyword Extraction*

Build a production system for extracting search keywords based on AHP in advance. In this section, we will take up the example of a target document in the “drawing format”, which has not been tried so far, and show how to determine the importance and certainty of knowledge based on AHP. Fig. 1 shows an example of the target document. As illustrated, the characteristic of drawing-format documents is that the number of line segments is large and the number of characters is small. The top layer of the AHP is the purpose, the importance (weighting coefficient), and the evaluation items of the middle layer are the font size, vertical and horizontal position, and appearance frequency. Also, the alternatives of the bottom layer are of high and low importance.

*1) Evaluation items (knowledge about keyword candidates):* In order to extract keyword candidates from the extracted multiple character areas, the following knowledge is effective. (1) Keywords are larger than other characters (Fontsize). (2) The positions with keywords are often left, middle, right (X Position) and vertical (Y Position) in the drawing. (3) Keywords appear frequently in the drawing (Frequency). The evaluation item values (Fontsize, Y Position, X Position, Frequency) of these knowledges are extracted as follows.

*a) Regarding font size (Fontsize) and position information (Y Position, X Position),* the target document was read by a scanner, and converted into HTML language as layout information and extracted by a unique program.

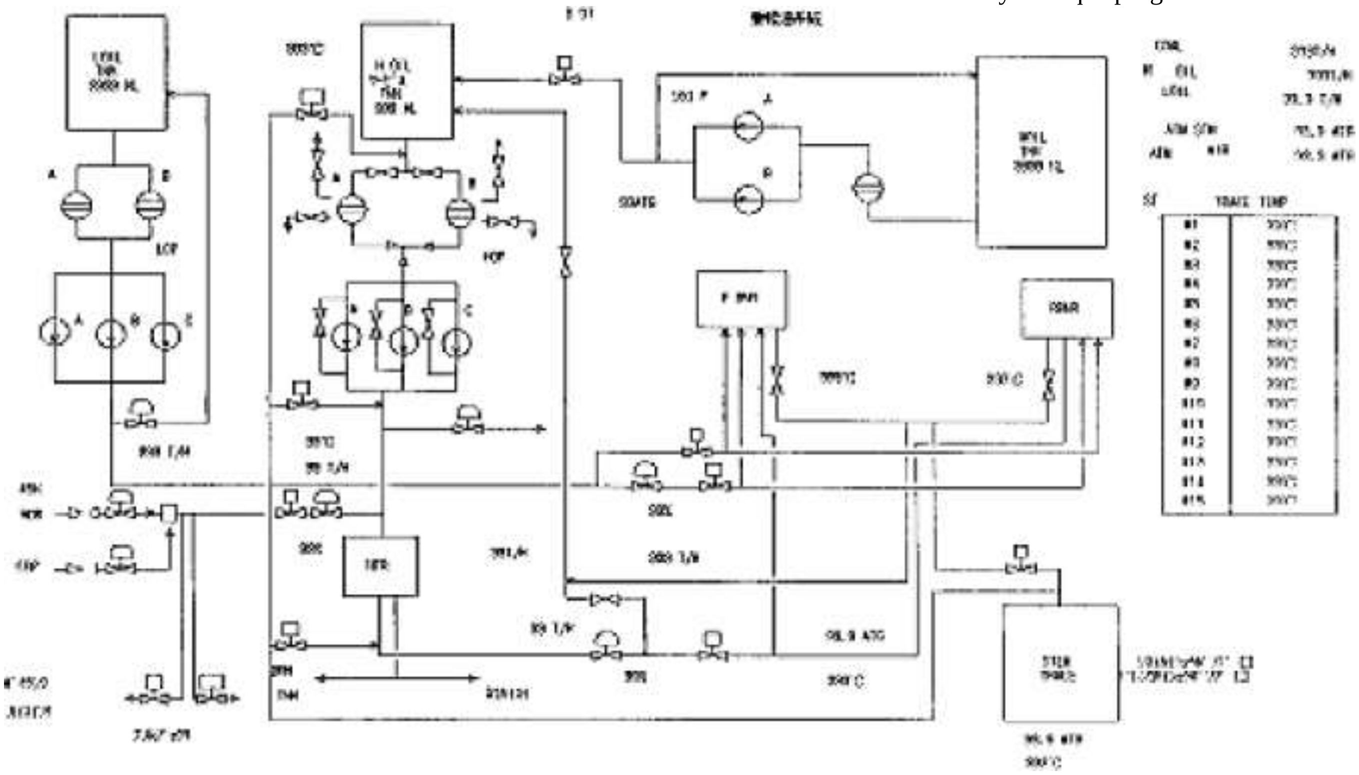


Fig. 1. An Example of the Drawing Diagram Documents.

*b) The frequency was extracted by character recognition*, converting it to text, and then using the morphological analysis software ChaSen described above to examine the frequency of the "divided words".

*2) Impact of evaluation items:* The target document in the drawing format was presented to 10 subjects, and the importance or impact of the evaluation items shown in 1) was specified in a 10-point scale from 0 to 1, and the font size : 0.98, frequency of occurrence: 0.98, vertical position: 0.694, horizontal position: 0.23 were found to be the average degree of influence.

*3) Knowledge base design:* Knowledge of the proposed method is expressed in the form of the production rule “IF is then THEN”. As for the knowledge of the proposed method, the evaluation item state of the evaluation item such as "the size of Fontsize" was described in the condition part, and the keyword part was described in the consequent part. That is, “IF Fontsize is Big THEN is a keyword in confidence CF”. At this time, the evaluation item to be used is to obtain the certainty factor in consideration of the relative importance determined in advance according to each document format. Also, in the knowledge of the proposed method, only the form of logical sum is used in which multiple knowledges with unequal condition parts derive the same consequent part.

*4) Certainty factor:* One of the features of the production system is that it makes it possible to handle uncertain knowledge by imposing a weighting factor called certainty factor on each knowledge. In this paper, the certainty factor is defined as an index that expresses the degree to which the consequent part can be derived by the condition part in a certain knowledge rule. The range of this value is ± 1, and when 0 means that the conditional part is not taken into consideration when deriving the consequent part. Negative means a degree of negative consequent derivation, and positive means a degree to support consequent derivation. At this time, the certainty factor was obtained based on the relative importance obtained by AHP. Fig. 2 shows the process steps from knowledge representation and input of the target document to determination of certainty factor. The procedure is shown below.

*a) To determine the certainty factor,* the purpose of the uppermost layer of the AHP hierarchical structure was to calculate the importance of knowledge, and the evaluation items were Fontsize, X Position, Y Position, and Frequency. In the alternative layer, which is the lowest layer, Big is used when the Fontsize is large, Y Position and X Position are close to the place where importance is considered, and when there is a lot of Frequency in each evaluation item, and Small is used.

*Regarding X Position and Y Position*, to determine which position in the layout of the character string is important, the document is divided into 5 parts vertically and horizontally, and the positions are A (the end) and B. It is represented by a fuzzy set with (somewhat end), C (middle), D (somewhat end), and E (most end). Fig. 3 shows the membership function. Also, by using the knowledge that the position of a keyword in the target document is often left, right, upper, or lower in the drawing, A or C or E> for each of the position importance evaluation items, X Position, and Y Position. B or D.

### E. Example of Calculation Result of Relative Importance by AHP of Target Document in Letter Format

Table 1 shows the weighting factors (relative importance) of the evaluation items calculated by one-to-one comparison based on AHP. In the case of letter format, 10 subjects were judged to be important for keyword selection in the order of font size, vertical position, horizontal position, and appearance frequency. Therefore, the one-to-one comparison table (matrix) of the evaluation item hierarchy is arranged in the order of Fontsize, Y Position, X Position, and Frequency, and their order numbers are the matrix elements in the first row. Since the element (i, j) of this matrix is the reciprocal of the element (j, i), all elements are obtained as shown in Table 1. This matrix can be used to determine A in Eq. (1), and the eigenvalues and eigenvectors for this can be found to calculate the relative importance in Table 1.

Tables 2 to 5 show the one-to-one comparison results of each alternative for each evaluation item.

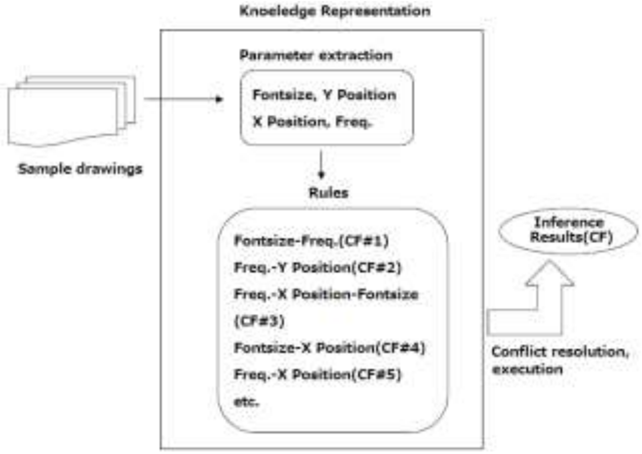


Fig. 2. Process Flow of the Proposed Method and the Relation between Confidential Factors and Knowledge Representations.

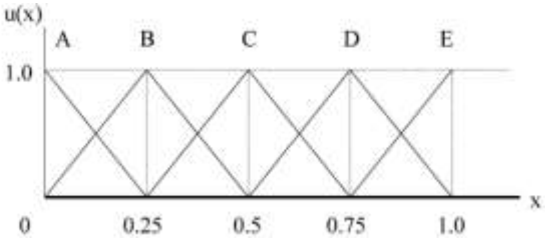


Fig. 3. Membership Function of the Fuzziness on the Location at which the Keywords for Document Retrievals Appear.

TABLE II. ONE-TO-ONE COMPARISON ON THE ALTERNATIVES ON “FONT SIZE”

| | Big | Small | Relative Importance |
|---|---|---|---|
| Big | 1 | 9 | 0.9 |
| Small | 1/9 | 1 | 0.1 |

TABLE III. ONE-TO-ONE COMPARISON ON THE ALTERNATIVES ON "Y POSITION"

| | Big | Small | Relative Importance |
|---|---|---|---|
| Big | 1 | 6 | 0.85 |
| Small | 1/6 | 1 | 0.15 |

TABLE IV. ONE-TO-ONE COMPARISON ON THE ALTERNATIVES ON "X POSITION"

| | Big | Small | Relative Importance |
|---|---|---|---|
| Big | 1 | 4 | 0.8 |
| Small | 1/4 | 1 | 0.2 |

TABLE V. ONE-TO-ONE COMPARISON ON THE ALTERNATIVES ON "FREQUENCY"

| | Big | Small | Relative Importance |
|---|---|---|---|
| Big | 1 | 2 | 0.67 |
| Small | 1/2 | 1 | 0.33 |

The maximum relative importance of the evaluation items is 0.47, and the number of subjects is 10. Therefore, if you multiply the relative importance by 20 and make it an integer, the integer values for the evaluation items of Fontsize, Y Position, X Position, and Frequency are It becomes 9, 6, 4, and 2. The one-to-one comparison matrix for the alternative is a 2x2 matrix because there are two alternatives (the evaluation items are large or small). Therefore, the font size of the maximum relative importance is 1: 9 for "large" compared to "small", and the results shown in Table 2 are obtained. The results shown in Tables 3 to 5 are obtained for other evaluation items as well. The relative importance of Tables 2 to 5 can be obtained by the same method as the relative importance of Table 1.

In addition, by checking the consistency of the one-to-one comparison result (calculating the Consistency Index: CI) by the method shown in Appendix 2, it was confirmed that the consistency was found in all cases from Tables 1 to 5 below 0.1. it can. Therefore, although the number of subjects is as small as 10, the results of these one-to-one comparisons are reliable.

Table 6 shows the result obtained by multiplying the weighting factors of the evaluation items in Table 1 by each alternative, adding their values, and integrating.

This shows that the relative importance is 0.85 when all the evaluation items are the Big, and the relative importance is 0.15 when all the evaluation items are the Smallest. Based on Tables 1 to 5, the confidence level of all knowledge is determined as shown in Fig. 4. This is a calculation example of the certainty factor when the font size is large, the horizontal position is large, the vertical position is small, and the appearance frequency is large.

Based on the confidence factor synthesis method [16], the certainty factor was synthesized by the logical sum of two pieces of knowledge. The specific method is shown in Appendix 3.

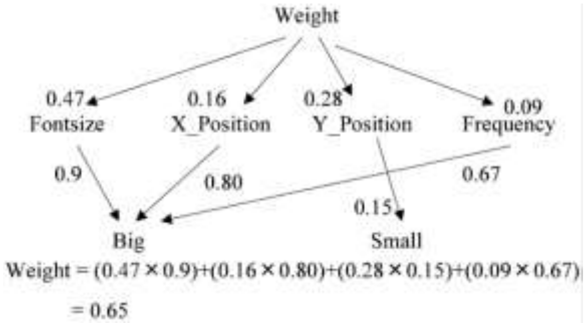


Fig. 4. An Example of the Confidential Factor Determination (In the case that Big: Font Size, Small: Y Position, Big: X Position, Big:. Frequency).

TABLE VI. THE ESTIMATED PRIORITIES OF THE ALTERNATIVES

| | Font Size | Y Position | X Position | Frequency | Relative Importance |
|---|---|---|---|---|---|
| Big | 0.9x0.47= 0.42 | 0.85x0.28= 0.24 | 0.8x0.16= 0.13 | 0.67x0.09= 0.06 | 0.85 |
| Sm all | 0.1x0.47= 0.05 | 0.15x0.28= 0.04 | 0.2x0.16= 0.03 | 0.33x0.09= 0.03 | 0.15 |

In Fig. 5, if there are multiple matching knowledge rules in the knowledge base, knowledge conflicts are avoided by selecting and executing the one with the most detailed knowledge condition part. Then, the certainty factor of the keyword and the keyword candidate are output.

If this certainty factor exceeds a certain threshold (0.96 in this paper), keyword candidates are automatically adopted. Otherwise, the certainty factor and keyword candidates are presented to the operator for selection. And register the determined keywords in the database.

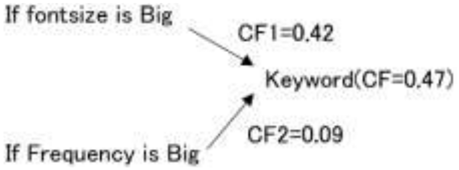


Fig. 5. An Example of Synthesizing Two Confidential Factor into One.

## IV. PROPOSED METHOD

### A. Evaluation of the Proposed Method and the Conventional Method

The proposed system was evaluated by comparison with the subjective evaluation. The evaluation was performed using 100 types of target documents according to the following procedure for 10 subjects.

*1)* The subject is presented with the target document and asked to select what seems to be a keyword. In addition, for all keyword candidates, the importance (selectivity) was evaluated from 0 to 1 in 1/8 steps.

*2)* Select a keyword based on the proposed method. In addition, the importance (selectivity) was evaluated for all of the keyword candidates.

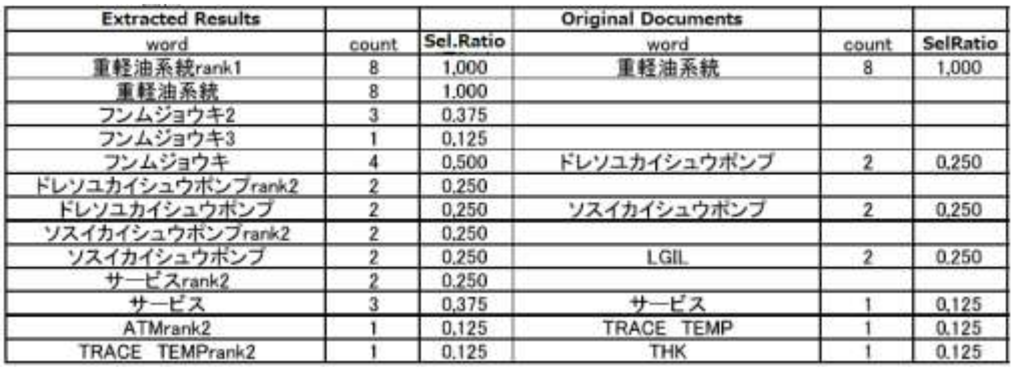
*3)* The keywords extracted in steps (1) and (2) are compared and the matching rate of the keywords is obtained.

Meanwhile, evaluation of conventional methods is also performed. Similarly, the author selected keywords from 100 target documents, if the importance of all knowledge (evaluation items) was the same and compared them with the keywords selected based on the subjectivity of 10 subject and evaluated the concordance rate.

## *B. Evaluation Results*

Fig.6 shows an example of the keyword matching rate. The results selected by the proposed system as keyword candidates are shown in the left column, and the results selected by the subjects are shown in the right column. Furthermore, at this time, the importance of the keyword candidates selected based on the font size, vertical position, horizontal position, and appearance frequency selected as the evaluation items (parameters), the keyword candidates and the certainty factors determined based on AHP, and the final importance.

Fig.7 shows the degrees. The word "heavy diesel fuel system" is found to be the most important keyword in both the proposed system and the subject. In the case of 98% of 100 types of target documents, we confirmed that the keywords that the subjects considered most important were included in the keyword group selected by the proposed method. In the case of the symmetric document in the illustrated drawing format, since there was only one keyword candidate with a certainty factor of 100%, the keyword could be automatically extracted as a search keyword. However, according to the conventional method that considers that the importance of the knowledge to be used is all equal, when there are multiple keyword candidates with the same certainty factor, a keyword different from the subjectivity of the subject may be selected. The method has a higher concordance rate because the learning result by the teacher set as pre-learning is reflected when determining the importance of the knowledge to be used.



| Extracted Results | | | Original Documents | | |
|---|---|---|---|---|---|
| word | count | Sel.Ratio | word | count | SelRatio |
| 重軽油系統rank1 | 8 | 1.000 | 重軽油系統 | 8 | 1.000 |
| 重軽油系統 | 8 | 1.000 | | | |
| フンムジョウキ2 | 3 | 0.375 | | | |
| フンムジョウキ3 | 1 | 0.125 | | | |
| フンムジョウキ | 4 | 0.500 | ドレソユカイシュウポンプ | 2 | 0.250 |
| ドレソユカイシュウポンプrank2 | 2 | 0.250 | | | |
| ドレソユカイシュウポンプ | 2 | 0.250 | ソスイカイシュウポンプ | 2 | 0.250 |
| ソスイカイシュウポンプrank2 | 2 | 0.250 | | | |
| ソスイカイシュウポンプ | 2 | 0.250 | LGIL | 2 | 0.250 |
| サービスrank2 | 2 | 0.250 | | | |
| サービス | 3 | 0.375 | サービス | 1 | 0.125 |
| ATMrank2 | 1 | 0.125 | TRACE TEMP | 1 | 0.125 |
| TRACE TEMPrank2 | 1 | 0.125 | THK | 1 | 0.125 |

Fig. 6. Shows an Example of the Keyword Matching Rate.

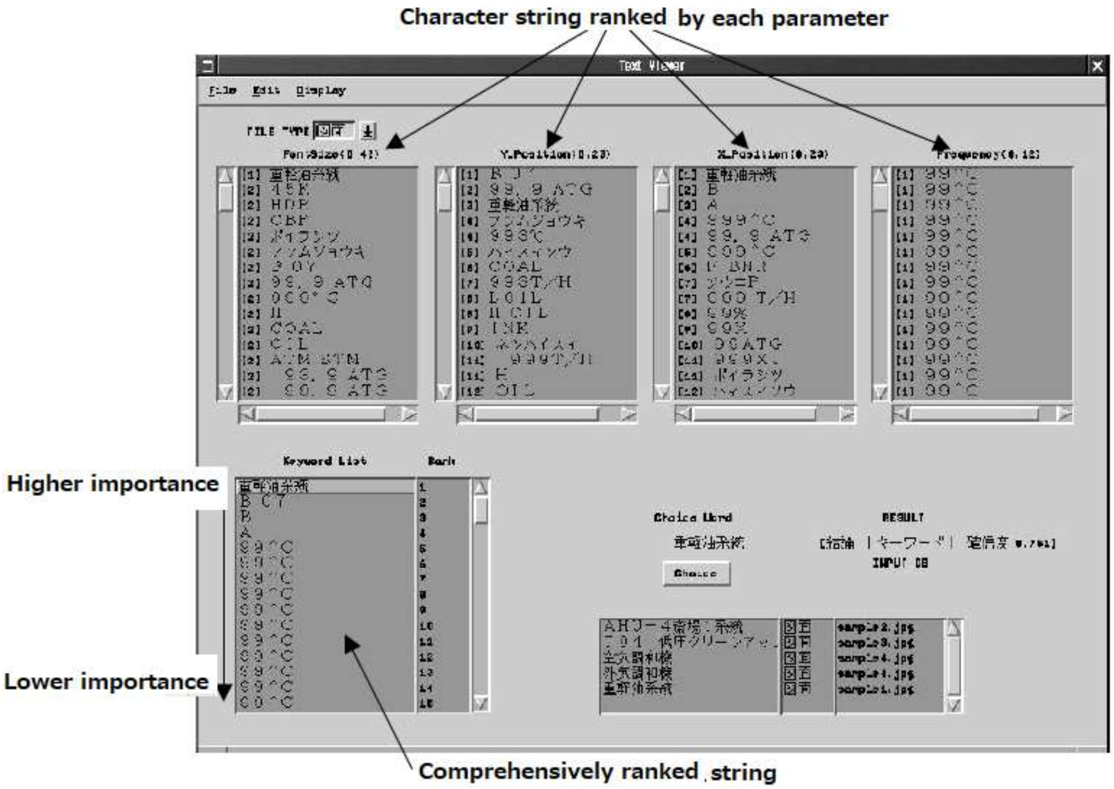


Fig. 7. An Example of the Knowledge Priority Determination (Objective Document is the Drawing Diagram Document of Fig. 1).

## V. Conclusion

In the case of a document (2%) in which the keyword size, the position of the keyword candidate in the document, and the frequency of occurrence are exactly the same among the 100 types of target documents (2%), the operator selects the keyword candidate and the importance level. (Confidence factor) had to be presented to make a decision, but it was confirmed that the keywords of the remaining majority (98%) of the target documents could be automatically extracted.

When the importance of knowledge evaluation items (Fontsize, Y Position, X Position, Frequency) is all equal without estimating the certainty factor by AHP proposed in this paper, the font size and appearance frequency of specific keyword candidates are considered. If it is different from other candidates, the keyword can be extracted correctly, but, knowledge about the appearance position of the keyword does not work effectively, and as a result, only 75% can automatically extract the keyword. In the end, it was found that the effect of the confidence evaluation by AHP was about 1.5 times higher in success rate.

This has a great effect of automatically identifying the target document in advance in the letter format, the form format, and the drawing format, and using the importance level of the knowledge relating to the keyword candidate selection suitable for each format. This is because it is possible to grasp the whole thing and make a comprehensive judgment.

## VI. Future Research Works

Further experiments are required for the validity check of the proposed method for automatic letter document keyword extractions.

## Acknowledgment

The author, also, would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

## References

[1] iDC home page: http://www.idcinit.com/.

[2] iOffice home page: http://software.fujitsu.com/jp/ioffice/index.html.

[3] AIST home page: http://www.carc.aist.go.jp/nlwww/~y.matsuo /keywaord-extraction/DDD/.

[4] Gengokk home page: http://www.gengokk.co.jp/jyuuyou.htm.

[5] http://H. Sakai, K. Ohtake and S. Masuyama, On retrieval support system by suggesting terms to a user, Proc. of NTCIR WS-2, pp.222-226, 2001.

[6] Fujii Home page: http://software.ssri.co.jp/fuji/ts proinfo.html.

[7] H. Fujii and B. Croft: "A comparison of indexing techniques for Japanese text retrieval", Proc. of SIGIR'93, pp.237–246 (1993).

[8] S. Ananiadou: "A methodology for automatic term recognition", Proc. of COLING'94, pp.1034–1038 (1994).

[9] K. Kageura and B. Umino: "Methods of automatic term recognition: a review", Terminology, Vol.3, No.2, pp.259–289 (1996).

[10] H. Nakagawa and T. Mori: "Nested collocation and compound noum for term recognition", Proc. Of COMPTERM'98, pp.64–70 (1998).

[11] Nara Institute of Science and Technology Homepage: http://chasen.aist-nara.ac.jp/.

[12] SJIT version dictionary source file http://chasen.aist-nara.ac.jp/stable/ ipadic.win/.

[13] Masaki Matsudaira, Toshio Ueda, Hiroyuki Onuma, Masamune Fuchigami, Yukihiki Morita: "Keyword Extraction from Documents and Collection of Related Information", AI Conference, SIG-SWO-A303-02, pp.02-01-02-06 (2002).

[14] Thomas L. Saaty: Decision making for leaders: The analytic hierarchy process for decisions in a complex world 1999/2000 Edition (2000).

[15] Yuichiro Anzai: "Production Systems and Artificial Intelligence Research —Toward the Problem of Knowledge Acquisition", Computer Software, Vol.1, No.3, pp.2–12 (1984).

[16] Buchanan, B.G., and E.H. Shortliffe: Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project, Reading, MA, Addison- Wesley (1984).

[17] Kohei Arai, Tolle Herman, Method for real time text extraction of digital manga comic, International Journal of Image Processing, 4, 6, 669-676, 2011.

[18] Kohei Arai, Yuichiro Eguchi and Yoichiro Kitajima, Extraction of line features from multifidus muscle of CT scanned images with morphological filter together with wavelet multi resolution analysis, International Journal of Advanced Computer Science and Applications, 2, 8, 60-66, 2011.

[19] Kohei Arai and Tolle Herman, Method for extraction product information from TV commercial, International Journal of Advanced Computer Science and Applications, 2, 8, 125-131, 2011.

[20] Kohei Arai and Tolle Herman, Text extraction from TV commercial using blob extraction method. International Journal of Research and Review of Computer Science, 2, 3, 895-899, 2011.

[21] Kohei Arai, Ronny Mardiyanto, Eye-based human-computer interaction allowing phoning, reading e-book/e-comic/e-learning, Internet browsing and TV information extraction, International Journal of Advanced Computer Science and Applications, 2, 12, 26-32, 2011.

[22] Kohei Arai and Tolle Herman, Method for automatic e-comic scene frame extraction for reading comic on mobile devices, Proceedings of the Information Technology for Next Generation (ITNG2010) 2010.

[23] Kohei Arai, T.Herman, Method for information extraction from Japan TV commercial, Proceedings of the 260th conference in Saga of Image and Electronics Engineering Society of Japan, 19-25, 2012.

[24] Kohei Arai, T.Herman, Automatic information extraction from on-air TV commercial contents, Proceedings of the International Conference on Convergence Content 2012, 171-172, 2012.

[25] T.L.Satty:"The Analytic Hierarchy Process," McGrawHill, 1980.

## Appendix

*1)* Algorithm for obtaining weighting factors between elements in each layer: Let A = aij be the one-to-one comparison matrix of the elements A1, A2, …,, An. If the weighting factor w to be obtained is given by w1, w2, ⋯, wn when A is known, A becomes as shown in Eq. (1).

$$A = \left[a_{ij}\right] = \begin{bmatrix} w_1/w_1 & - & w_1/w_n \\ | & | & | \\ w_n/w_1 & - & w_n/w_n \end{bmatrix} \tag{1}$$

At this time, $a_{ij}$ is ideally as follows,

$a_{ij} = w_i/w_j$, $a_{ji}$=1/$a_{ij}$

$$w = \begin{bmatrix} w_1 \\ | \\ w_n \end{bmatrix}, i, j = 1, 2, \dots, n \tag{2}$$

At this time, if $a_{ij} \times a_{jk} = a_{ik}$ holds for *i*, *j*, and *k*, it can be said that the decision-makers' judgments are perfectly consistent. Next, multiplying Eq. (1) by *w* from the right yields Eq. (3).

$$Aw = \begin{bmatrix} \frac{w_1}{w_1} & - & \frac{w_1}{w_n} \\ | & | & | \\ \frac{w_n}{w_1} & - & \frac{w_n}{w_n} \end{bmatrix} \begin{bmatrix} w_1 \\ | \\ w_n \end{bmatrix} = n \begin{bmatrix} w_1 \\ | \\ w_n \end{bmatrix} \tag{3}$$

Then,

$$Aw=nw \tag{4}$$

Equation (4) is obtained as a solution to the eigenvalue problem as follows,

$$(A\text{-}nI)w=0 \tag{5}$$

At this time, n must be the eigenvalue of $A$ for $w = 0$. When n becomes the eigenvalue of $A$, $w$ becomes the eigenvector of A. Also, from rank ($A$) = 1, the eigenvalues $\lambda_i$ (i = 1, 2, ..., $n$) except for 0 are given the maximum eigenvalue $\lambda_{\max}$, and other eigenvalues $\lambda_i = 0$. Since the sum of the main diagonal elements of $A$ is $n$, $\lambda_{max}$ satisfies $\lambda_{max} = n$. Therefore, $w$ is the eigenvector normalized to $\lambda_{\max}$ of A. In other words, it can be said that they are completely consistent. However, it is extremely difficult for the decision maker to decide the matrix $A$ that gives the same weighting coefficient as w. Therefore, when the one-to-one comparison matrix obtained from the decision maker is $A$ 'and the weighting coefficient obtained from the one-to-one comparison matrix is $W$ ', Eq. (4) is replaced as Eq. (6).

$$A'W'=\lambda'_{\max} W' \tag{6}$$

Therefore, $W'$ is the normalized eigenvector for the maximum eigenvalue $\lambda_{max}$ of $A$ '.

*2) Method of consistency check in one-to-one comparison:* When an inconsistency occurs in an $n$-by-$n$ one-to-one comparison matrix, the maximum eigenvalue λ max becomes larger than $n$. This is called Satty's theorem [25] and is expressed by Eq. (7).

$$\lambda'_{max} = n + \sum_{i=1}^{n} \sum_{j=j+1}^{n} (W_j\, a_{ij} - W_i')^2 / W_i' W_j' a_j\, n \tag{7}$$

From Eq. (7), it is found that $\lambda_{\max}$ always satisfies $\lambda_{\max} \geq n$. Therefore, *C.I.* (Consistency Index) is defined as an index of consistency check for one-to-one comparison. *C.I.* is expressed by Eq. (8).

$$C.I. = \frac{\lambda'_{max} - n}{n-1} \tag{8}$$

Consistency increases as *C.I.* = 0 approaches, and conversely decreases with distance from 0. Generally, if *C.I.* is 0.1 or less, it is said that there is consistency.

*3) Confidence synthesis method:* The *CF* that combines $CF_1$ and $CF_2$ when deriving the same consequent part is as follows,

$$CF(CF_1, CF_2) = \begin{cases} CF_1 + CF_2 - (CF_1 * CF_2) \\ CF_1 > 0 \; and \; CF_2 > 0 \\ CF_1 + CF_2 + (CF_1 * CF_2) \\ CF_1 < 0 \; or \; CF_2 < 0 \\ \frac{(CF_1 + CF_2)}{1 - \min(|CF_1|, |CF_2|)} \\ other \end{cases} \tag{9}$$

## AUTHOR'S PROFILE

**Kohei Arai,** He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html