# Category-based Extracting Keyword Method Selection with Genetic Programming

Ayahiko Niimi*, Takuma Yasunobu† and Eiichiro Tazaki†

*Department of Media Architecture, Future University-Hakodate
116-2 Kamedanakano-cho, Hakodate 041–8655, Japan
E-mail: niimi@fun.ac.jp

†Department of Control and Systems Engineering, Toin University of Yokohama
1614 Kurogane-cho, Aoba-ku, Yokohama 225-8502, JAPAN
E-mail: tazaki@intlab.toin.ac.jp

***Abstract*—Quality of keywords given to each document is important to search documents from a lot of document databases. It is necessary automatically extracting high quality keywords from a document to achieve a document search with high efficiency. Some extracting keyword methods were proposed, but these extracted keywords accuracy is depended on document categories. In this paper, we proposed a keyword extraction approach with selection of extracting keyword method depended on document categories using genetic programming. Keyword extracting methods were applied to five kinds of documents with a different category for the verification, it was confirmed that there was a performance difference corresponding to the category in the keyword extracting method. In addition, keyword extracting methods were combined by using genetic programming, its performance was improved from a single purpose method.**

***Index Terms*— genetic programming, extracting keyword, categories, information retrieval**

## I. Introduction

Recently, various and large amount of information can be obtained easily from Internet. But, it is not easy to look for useful information from among these information sources. Some web search engine can use full text search, but almost results are dust. Keyword search is useful for not only web search but also document retrieval. An efficiency of searching from large amount of documents is strongly depended on a quality of keywords given to each document. It is necessary to achieve document retrieval with high efficiency that high quality keyword is automatically extracted from documents.

Some extracting keyword methods were proposed, but these extracted keywords accuracy is depended on document source categories. It is hard to tuning its algorithm's parameter. In general, keyword extraction method is tuned for specific domain, specific source.

In this paper, we assume that there is a keyword extraction method suitable in each document source category. We propose keyword extraction approach which contains classifying documents into categories, combining keyword extraction methods depended on category. For combining method, we use genetic programming. Genetic programming can use various criteria to evaluate system, and can automatically build combined system.

In section II and III, we first briefly describe genetic programming and keyword extracting method. In section IV, we describe our proposed approach. In section V, we apply our proposed approach to keyword extraction from some kind of document source categories, and discuss the results.

## II. Genetic Programming

Genetic programming (GP) is a learning method based on the natural theory of evolution, and the flow of the algorithm is similar to genetic algorithm (GA). The difference between GP and GA is that GP has extended its chromosome to allow structural expression using function nodes and terminal nodes. [1] In GP, individuals are expressed by S type form of LISP by using functions and the terminals.

To evaluate individuals, GP uses a fitness value function. It is possible to use synthesizing two or more indices like accuracy, size, and calculation time etc. of individuals to fitness value function. One of the synthesizing fitness value functions is based on MDL principle. [2]

The algorithm of GP follows the following procedures.

1) An initial population is generated from a random grammar of the function nodes and the terminal nodes defined for each problem domain.
2) The fitness value, which relates to the problem solving ability, for each individual of the GP population, is calculated.

3) The next generation is generated by genetic operations.
    a) The individual is copied by fitness value (reproduction).
    b) A new individual is generated by intersection (crossover).
    c) A new individual is generated by random change (mutation).
4) If the termination condition is met, then the process exits. Otherwise, the process repeats from the calculation of fitness value in step 2.

## III. Keyword Extracting Method

Various keyword extracting method were proposed. These keyword extractions are using morphological analysis, not using morphological analysis, using document structure, etc. [3] In this paper, we use following methods.

### A. Morpheme Analysis

A morphological analysis is to divide the input sentence into the morpheme which is a minimum unit with the meaning in linguistics, to decide the part of speech of each morpheme, and to allocate the prototype to the morpheme to which the transformation of the word of use. [4] A morphological analysis is important for Japanese documents, because Japanese sentence is not divide words by blank. In English, a morphological analysis is used to analyze end of a word transformation (tense, single or plural), suffix, prefix, etc. For instance, it is analyzed that the morphological analysis is done by the sentence “Happyoukai wo okonaitai.” (This sentence means “I want to hold a symposium”). (Refer to table I)

TABLE I
Examples of Morphological Analysis

| | | |
|---|---|---|
| **Happyou** | Happyou: | Noun |
| **kai** | Kai: | Noun |
| **wo** | Wo: | particle |
| **okonai** | Okonau: | verb-independent |
| **tai** | Tai: | auxiliary verb |
| **.** | . | symbol-period |

The word divided by the morphological analysis is called an element-term. It comes to be able to do the frequency analysis and filtering to a specific part of speech by dividing into the element-term.

### B. Appearance Frequency Analysis

The appearance frequency analysis means examining appearance frequency of each element-term divided by a morphological analysis. Element-terms with high appearance frequency are extracted as keywords. The method which extract element-terms with high appearance frequency as keywords, it is easy to extract keywords from what kind of document. Because element-terms sometimes contain no meaning term, a frequency analysis usually has stop word list(ex. a, an, the, is, get, give, etc.), and part of speech filter.

### C. Serial Noun Extraction

Almost keywords contain element-terms of noun. [5] In general, the noun is extracted by using the morphological analysis as keyword. For instance, by morphological analysis “Happyoukai wo okonaitai.” is analyzed 5 element-terms.(Happyou: Noun)(Kai: Noun)(Wo: particle)(Okonau: verb-independent)(Tai: auxiliary verb) In this case, noun “Happyou” and “Kai” are extracted as keyword. (“Wo”, “Okonau” and “Tai” are not extracted as keyword because of not noun.) “Happyou” and “Kai” are lack specifics, connected word “Happyoukai”is better for keyword. Therefore,a connected noun makes keyword more concreteness.

### D. N-gram

N-gram analysis can extract a part of strings from long strings. N express a number which is larger than 2. N-gram algorithm consist of taking serial N characters, calculating its association frequency, sorting characters by its frequency, extracting characters of high frequency as keyword. [5], [6] Beforehand the part of speech putting need not be done to documents, and an arbitrary number of characters can be used for keyword. However, a character string which contains a part of word is extracted as a keyword when analyzing documents without doing the part of speech putting the extraction. We use N-gram algorithm with doing morphological analysis and dividing into element-terms, extracting continuous element-term to improve efficiency.

### E. Association Rule

It is possible to extract keywords from association of characters or words which appears in one sentence. This analysis needs not to do morphological analysis like N-gram algorithm. There is apriori algorithm as a technique for extracting the association rule at high speed. [7] Association rule analysis has same problem of N-gram algorithm. In this paper, keywords were made from association rule between element-terms after doing morphological analysis so that this might also decrease a possibility of extracting only a part of word.

### F. Text Structure Analysis

It is possible to extract keywords by using sentence structure. In most documents of news, a topic sentence is put on a head. Because LaTeX, HTML and XML has description tag of title, section title, etc., extracting keyword can use these information sources. These extra information sources are good for keyword extraction, but all documents do not have tags. The text structure analysis is strongly depended on document source.

## IV. Auto-selection of Extracting Keyword Method with GP

It is thought by each keyword extraction method that there are good and weak in the object documents. It is difficult to analyze same analysis in documents not so structured though the keyword can be extracted to structured documents while analyzing the structure. Documents of E-mail are not so structured and short, documents of paper are structured well and long. Then, in E-mail and paper, it is thought that it is effective to use a different keyword extraction method. (Examples of document source category are shown in table II.) Moreover, in each keyword extraction method, should do parameter tuning to object documents additionally.

TABLE II
Examples of Document Source Category

| | long document | short document |
|---|---|---|
| hard structured | paper | news |
| week structured | column | mail |

In this paper, we assume that there is a keyword extraction method suitable in each document source category. We propose keyword extraction approach which contains classifying documents into categories, combining keyword extraction methods depended on category with GP. In this paper, human chose keyword is assumed to be a correct answer, and the correct answer rate was defined as follows.

$$\text{correct answer rate} = \frac{\text{number of system's correct}}{\text{number of human selection}}$$

Because of GP using, a suitable keyword extraction method for document source category is selected automatically, and keyword can be extracted. A fitness value function is designed becomes possible considering accuracy, a number of keywords, and calculation time until extracting. Moreover, a parameter of keyword extraction method can be learned at the same time.

Function nodes mean condition judgements of which document category is evaluated, terminal nodes mean using keyword extraction method. Parameters of each keyword extraction method are defined as a node. Keyword extraction method which specifies for structured document may be used with function node of “If a document is structured or not.” Fitness value is defined by sum of correct answer rate which is obtained by extracting method depended on GP’s terminal nodes. Then, an individual with highest correct answer rate becomes an individual with high fitness value. The number of extractions and extraction time of keyword are defined as fitness value function.

A weak point of keyword extraction system with GP is difficulty in real time learning. It is thought that an interactive keyword extraction system which calculates fitness value from evaluation of the system user. However, when it is made to learning interactively, waiting time becomes long because fitness value calculation of GP depends on the number of individuals and the number of nodes and it increases. Then, it is thought that it is possible to correspond to learn in real time by waiting learning in parallel with the evaluation by spending time that a evaluation input waiting time from system user and the system are not used etc.

Automatically clustering document methods were proposed, but it is not contained in our proposed approach. [8] In this paper, the specified category was decided by the user.

## V. Experiment Results

We applied it to keyword extraction using two or more algorithms from documents of two or more categories to verify the effectiveness of the proposal technique. We used paper, news, editorial, manual, and E-mail as categories of document source. First of all, keywords were extracted respectively by a hand work, and these were assumed to be a correct answer. Keywords from paper and manual were comparatively long documents, and extracted keywords were also long. Keywords from news and editorial were comparatively short documents, and extracted a lot of short keyword. In E-mail, minimum document was two sentences, maximum was decades sentences. A long keyword was consisted of quoted sentences. Same mail contained many topics in one mail, keywords in one document were sometimes not related each other. We used frequency analysis, extraction of a noun series, N-gram method based on characters, N-gram method based on words, association rule based method with words as a keyword extraction algorithms. In our proposed method, structure analysis algorithm can be used with “testing if structured or not” function nodes, but we didn't use them to compare keyword extraction method's feature.

When average correct answer rates of keyword of each document category was calculated as a preliminary experiment, it was able to be confirmed that the difference was in each category at the correct answer rate. (Refer to table III) The correct answer rate of N-gram in E-mail was especially low. It is thought that the reason that only short N could be extracted because there were a lot of documents constructed several sentences in E-mail.

Next, we applied our proposed method with GP for keyword extraction. The parameter of GP used the following. (Refer to table IV) Fitness value was calculated from a correct answer rate. After the correct answer rate had been calculated of keyword extraction in the experiment beforehand by each keyword extraction method, GP was learned using its rate, because it took time when the individual was evaluated with to be extracted keywords every time.

In the selection of each keyword extraction method based on the category which used GP, it became the following results, and this result average of correct answer rate became 0.58.

```
(if_news associate-w_key
    (if_editorial connect_noun_key
        (if_mail associate-w_key
            ngram-w_key)))
```

The result same as a preliminary experiment was obtained by some categories. However, it did not become the best result in paper and manual. When documents were analyzed, a big difference was not seen at the correct answer rate in documents with a high keyword correct answer rate as for any method.

TABLE III

RESULTS OF CORRECT ANSWER RATE AVERAGE BY EACH METHOD

| | paper | news | edit. | man. | mail |
|---|---|---|---|---|---|
| frequency | 0.43 | 0.24 | 0.83 | 0.37 | 0.67 |
| connect_noun | 0.60 | 0.47 | 0.85 | 0.46 | 0.50 |
| ngram-c | 0.18 | 0.03 | 0.03 | 0.11 | 0.05 |
| ngram-w | 0.39 | 0.24 | 0.28 | 0.34 | 0.02 |
| associate-w | 0.52 | 0.70 | 0.41 | 0.09 | 0.64 |

## VI. CONCLUSION

In this paper, we proposed a keyword extraction approach in which GP automatically selects a pair of doc-

TABLE IV

PARAMETER OF GP

| | |
|---|---|
| GP Population | 500 |
| Reproduction Probability | 0.1 |
| Crossover Probability | 0.8 |
| Mutation Probability | 0.1 |
| Selection Method | tournament method |
| Function Nodes | 5 types in table V |
| Terminal Nodes | 5 types in table VI |
| Number of Training Data | each 25 documents with different category (total 125 documents) |

TABLE V

GP FUNCTION NODES

| Display | Meaning |
|---|---|
| if_paper | if category is paper, evaluate arg1, otherwise evaluate arg2 |
| if_news | if category is news, evaluate arg1, otherwise evaluate arg2 |
| if_editorial | if category is editorial, evaluate arg1, otherwise evaluate arg2 |
| if_manual | if category is manual, evaluate arg1, otherwise evaluate arg2 |
| if_mail | if category is mail, evaluate arg1, otherwise evaluate arg2 |

TABLE VI

GP TERMINAL NODES

| Display | Meaning |
|---|---|
| frec_key | using appearance frequency analysis |
| connect_noun_key | using serial noun (connected noun) |
| ngram-c_key | using N-gram with characters |
| ngram-w_key | using N-gram with words |
| associate-w_key | using association rule with words |

ument source category and extracting method with high correct answer rate. To verify our approach, we applied keyword extraction experiment, and evaluate its result.

As a result, it was confirmed that the performance of keyword extraction method was depended on each category. And, improvement accuracy was able to seem by GP combining keyword extraction algorithms compared with case to use keyword extraction method alone. However, it was not possible to learn well when there were little differences at the correct answer rate.

In the future, we will discuss an approach for parameter tuning of keyword extraction algorithm at the same time, combining structure analysis approach, and keyword extraction with high accuracy or more.

## References

[1] J.R. Koza: Genetic Programming, MIT Press (1992).

[2] H.Iba: Genetic Programming, Tokyo Denki University Press (1994). (In Japanese)

[3] Y. Ichimura, T. Hasegawa, I. Watanabe, M. Sato: Text Mining: Case Studies, Journal of Japanese Society for Artificial Intelligence, Vol.16 No.2,pp.192–200 (2001). (In Japanese)

[4] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, M. Asahara: Morphological Analysis System ChaSen version 2.2.1 Manual (2000). [Online] Available: http://chasen.aist-nara.ac.jp/chasen/bib.html.en

[5] T. Nasukawa, H. Kawano, H. Arimura: Base Technology for Text Mining, Journal of Japanese Society for Artificial Intelligence, Vol.16,No.2,pp.201–211 (2001). (In Japanese)

[6] M. Nagao, S. Mori: A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, In Proceedings of the 15th International Conference on Computational Linguistics pp.611–615 (1994).

[7] R. Agrawal, R. Srikant: Fast Algorithms for Mining Association Rules, the 20th International Conference on Very Large Databases, Santiago, Chile, September 1994:32pages (1994).

[8] M. Nagata, H. Taira: Text Classification - Showcase of Learning Theories -, IPSJ Magazine, Vol.42 No.1,pp.32–37 (2001). (In Japanese)