A Report on the Size of Information Unit to Extract Contents on the Web Text

Saori Kitahara and Kenji Hatano

***Abstract*—In this paper, we report that our proposed word-based method for extracting the contents of a Web page seems the most effective. Users frequently want to find useful information on Web pages, and they normally utilize a search engine equipped with various methods to extract the contents of Web pages. In most cases, the contents are extracted at sentence level on the assumption that only one content can be extracted from any one sentence. However, this assumption does not always hold, because there are likely to arise some cases in which a sentence contains several contents, especially if the sentence is long. As a result, there is some possibility that users will not recognize the content of a Web page correctly by these methods. In our experiment using an evaluation measure based on recall and precision, we compared the content extraction methods based on sentences with those based on phrases and words.**

***Index Terms*—content-density distribution, the size of information unit for Web page recognition.**

I. Introduction

CURRENTLY, many Web pages are being generated on the Internet one after the other. The Web pages are increasing daily, so it is becoming more difficult for users to search out useful information. The reason that users cannot find useful information is that it is hard to grasp the content of each Web page. To solve this problem, researchers are developing some methods for extracting the contents in Web pages.

For example, tabulation of the result snippets from a Web search engine [1] and visualization of the contents in Web pages are such methods. These help users to identify what kind of contents are present and where the contents are located within the Web pages. The former method summarizes the search results with phrases related to query keywords. The latter method emphasizes where the content is located within a Web page by, for example, highlighting the parts including the content [2] or displaying the words concerning the content in larger type [3]. If search engines apply these kinds of content extraction methods, the users should be able to grasp the content of each Web page. Consequently, the users should be able to choose with ease the Web pages that contain relevant information.

The argument made above is not always false; however, users are sometimes unable to recognize the content of a Web page by such methods. This is because a sentence can have multiple contents but those methods can extract only one content from the sentence. If those methods extract only one content, users cannot browse the other contents of the sentence. Hence, there is a great difference between the judgments of users and computers in relation to extracting contents.

For example, only the first part of the second sentence in the right side of Fig. 1 includes the content "Munich Germany" in the judgment of the user. However, the contents are usually extracted by computers at sentence level, so computers regard "Munich Germany" as the content of the whole sentence, as shown in the left side of Fig. 1, because the contents are extracted at sentence level in the case. Such a great difference occurs especially, when the sentence is long. For various reasons, it can be stated that the longer the sentence is, the higher the possibility of some contents residing in the sentence there are for various reasons.

Methods are implemented with computers on the assumption that only one content can be extracted from any one sentence. There are also some cases in which one sentence has multiple contents because a large unspecified number of people can augment or edit the Web pages freely. For example, they can post to Q&A Site such as Yahoo! Answers [4] or edit Wikipedia [5] without thinking that any one sentence ought to have only one content. Against this background, we should consider altering contents in a sentence on a Web page.

In this paper, we reveal, whether the sentence-based, phrase-based, or word-based method is the most suitable method for extracting one content from a Web page. We conduct experiments to compare the content extraction method based on sentences with those based on phrases and words by the test collection which answers reflect the judgment of users and the measures based on recall and precision.

We believe that the word is a very important unit for judging the content in a Web page. This is because the word is used as the minimal unit of queries by the majority of search engines. Briefly, we can regard each content on the Web pages as the part in a text string of a Web page by a set of more than two words. We call such a text string "Web text" in this paper. The reason that we utilize more than two words is that users find it difficult to resolve the polysemy of a single word.

II. Definition and Related Work

A. Information Unit

First, we have to define the information unit. This is the most important definition related to a content extraction method for Web pages.

In the research field of information retrieval, when users want to find useful information within a document set, a technique for extracting information from the documents normally applied. At that time, it decides a unit of information

Manuscript received December 21, 2012; revised January 30, 2013. This work was supported in part by JSPS Grant-in-Aid for Scientific Research (A) #22240005 and JSPS Grant-in-Aid for Young Scientists (B) #2700248.

S. Kitahara is with the Graduate School of Culture and Information Science, Doshisha University, 1-3 Tatara Miyakodani, Kyotanabe, Kyoto, 610-0394, Japan (e-mail: kitahara@ilab.doshisha.ac.jp).

K. Hatano is with Faculty of Culture and Information Science, Doshisha University, 1-3 Tatara Miyakodani, Kyotanabe, Kyoto, 610-0394, Japan (e-mail: khatano@mail.doshisha.ac.jp)

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

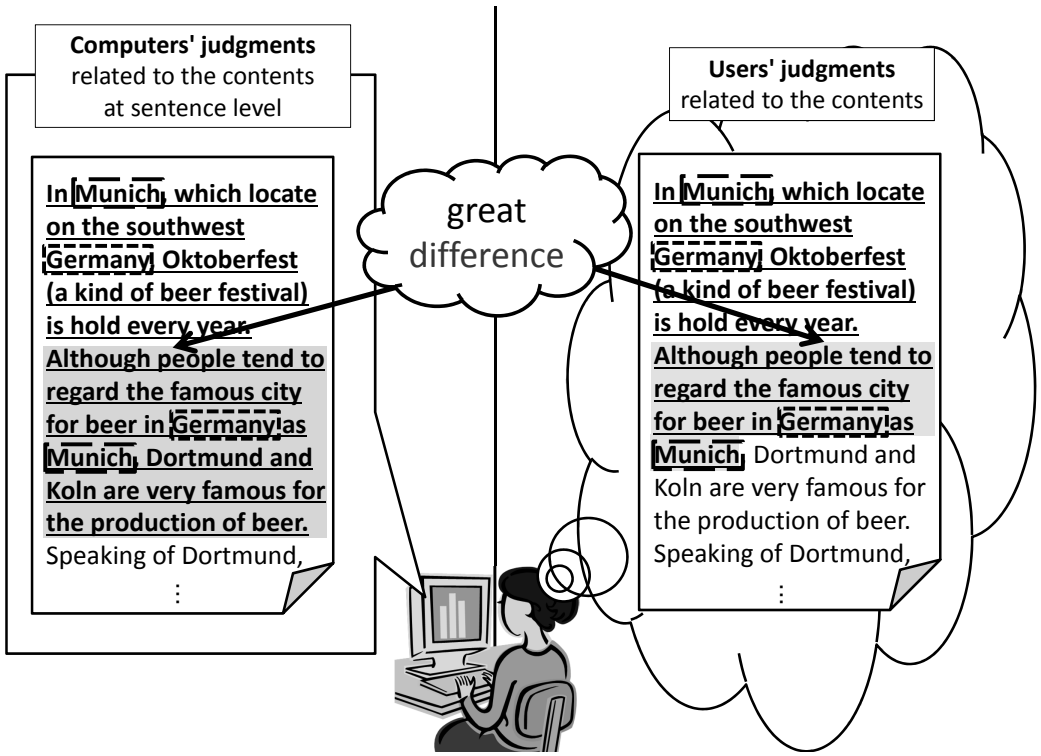

Fig. 1. Great Difference between the judgments by users and computers

extraction (e.g., sentence, phrase, or word) is decided, and we define the unit of information extraction as the information unit

Researchers engaged in work on techniques of information extraction have proposed various information units and have discussed which information units are most suitable in their own research fields. For instance, an XML element that is a part of an XML document is an information unit in the field of XML document retrieval [6], and a set of Web pages that contain query keywords is an information unit in the field of Web document retrieval [7].

The purpose of these studies is not to list specific documents as the search result but only to show users where the useful information exists in the documents. In other words, their studies enable users to browse useful information on their own without spending time to show much information that is insignificant for them.

Based on a combination of the descriptions in Section I and this subsection, we deduce that the sentences is the most common information unit in the research field of content extraction. As we described in Section I, a certain portion of a Web page through which users are unlikely to browse is often extracted as the content of a document when the information unit is the sentence. Therefore, we also need to discuss the size of information unit for extracting the content of Web pages. In this paper, our goal is to determine how large an information unit is most suitable for extracting the content in a Web page, and we achieve this by comparing three information units: the sentence, phrase, and word.

B. Related Work

In this subsection, we introduce two studies related with our proposal. The first has the Web page as the information unit. The second has the sentence as the information unit.

Lv et al. [8] studied passage retrieval with a window of variable size, although other conventional methods of passage retrieval involve a window of fixed size. The window in a Web text is the range processed by the method at one time. They extended the information retrieval model based on the idea of the language model named the positional language model [9]. That is, they executed passage retrieval with passages of the flexible size (i.e., soft passage retrieval), and this model was enabled them to conduct passage retrieval without a restriction on the size of the retrieval-result texts. By virtue of the model, users can view passages commensurate with each Web text.

In addition, the model was also enabled them to find the position and proximity of query keywords in some cases. However, they calculated the passage retrieval score representing the estimated word count at each position in a Web text. This score is the indicator not for part of a Web page but for an entire Web page, because the score is compressed by Web pages. Therefore, we regard the positional language model as a method whose information unit is the Web page. We do not utilize this method, because it was designed for comparing each Web text by calculating the score of each Web text, whereas our goal is to determine what size of information unit is suitable for extracting the content in a Web page.

Sunayama et al. [2] developed a tool for visualizing parts of Web pages, which helps users to search for useful information. They regarded the principal subjects of Web pages as a set of nouns, because nouns constitute about 70% of all words in Web pages and they considered the nouns to represent subjects.

They also considered that the parts more relevant to the

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

subject of the Web page can be shown more strikingly to users. However, they calculated the indicator of this method by considering the words as parts of the sentences and summing up the weights, because their goal was helping users to recognize the sentences in the texts. We regard the indicator as the weight of each sentence in the Web page and their visualization method as one whose information unit is the sentence. We do not utilize their method in our comparative experiment, because they applied the concept of associated words to the method, and treated a value derived from the content of these words. This is different from our goal of comparing purely the size of the information unit.

In this paper, we discuss what is the most suitable size of information unit for extracting one content from a Web page. As mentioned above, the most suitable size is certainly smaller than a sentence, because we should consider the changes of contents within a sentence on a Web page. Therefore, we set the information unit of our content extraction methods to the phrase and the word which are smaller than the sentence.

In addition, we suggest the manner in which our proposed methods can be used to judge where the content lies in a Web page. Specifically, we extract the portion of a Web page where the content exists in a phrase by the method whose information unit is the phrase. As proposed before, we also extract the portion where there is content on the basis of content-density distribution. In Section III, we explain these proposed methods in detail.

III. Our Proposal

In this section, we explain our methods for extracting the content of each Web page based on two sizes of information unit. As we mentioned in Section I, we compare our method with the existing method whose information unit is the sentence. The size of information unit in our methods is smaller than that in the existing method; that is the phrase" or the word is used in these cases. In the following subsections, we explain how to extract the content of Web pages by each method.

A. Method Whose Information Unit is the Word

If we extract the portions of a Web page where words representing content are located, it can be said that we can never extract the content in the page. This is because a single word does not constitute a content and exerts a kind influence on a portion of the Web page. Therefore, we should consider positions and influence of the words. In this study, we utilize the content-density distribution through the content extraction method applied with the word as the size of the information unit [10]. This method, which is our previous proposal for constructing a content-density distribution of Web pages, can be used for grasping the content of each Web page. The content-density distribution indicates the position where a content exists and is constructed on the basis of the word-density distribution, which represents a portion of an extracted Web text and is influenced by one word in the Web page. Using the content-density distribution of Web pages, we can extract the content of the Web pages, recognize the content as parts of Web pages, and consider the positions

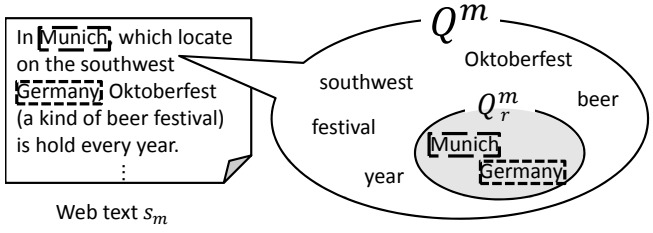

Fig. 2. An example of Q^m and Q_r^m

of the words as well as the influences of the words at each position.

However, the language model explained in Section II-B cannot include the positions of the words. Although the positional language model introduced in Section II-B as the expanded language model can include the positions of the words, this model is utilized for calculating the importance of each Web page to compare the Web pages. Therefore, we utilize the content-density distribution as a method whose information unit is the word. The subsequent processes in our method are for extracting Web page content. We construct the content-density distribution in the following three steps:

1) Calculation of the word-density distributions for a Web text.
2) Construction of the content-density distribution by using the word-density distributions.
3) Extraction of content from the Web text by using the content-density distribution.

1) Calculation of Word-density Distributions for a Web Text: In order to construct a content-density distribution from each Web page, we first calculate the word-density distribution of each word in the Web page. A word-density distribution represents the influence exerted on the Web page by one word. To calculate the word-density distribution, we pick out from a Web page a text string called a Web text.

When we extract the Web text $s_m (m = 1, 2, \cdots)$, Q^m denotes the set of words in the Web text s_m, and t_i^m denotes a word in s_m. Therefore, a content $Q_r^m (r = 1, 2, \cdots)$ in s_m is defined by a subset of Q^m (i.e., $Q_r^m \subset Q^m, r = 1, 2, \cdots$) (see Fig. 2). In this case, we denote the j th word of t_i^m in s_m $(j = 1, 2, \cdots)$, and $hw[t_{i,j}^m](k)$ refers to a value of the word-density distribution, which is taken for $t_{i,j}^m$ at the k th word in the Web text s_m. If $t_{i,j}^m$ appears at the position $l[t_{i,j}^m]$, then $hw[t_{i,j}^m](k)$ becomes the maximum. That is, the greater the distance of the position from $l[t_{i,j}^m]$, the smaller the value of $hw[t_{i,j}^m](k)$ is.

In addition, we can consider that the influence of the content may suddenly change at the end of a sentence. We call such points sentence separators, and these include periods, exclamation marks, and question marks. We define $a[t_{i,j}^m]$ as the sentence separator just before $l[t_{i,j}^m]$, and we define $b[t_{i,j}^m]$ as the sentence separator just after $l[t_{i,j}^m]$. Hence, the Web text from $a[t_{i,j}^m]$ to $b[t_{i,j}^m]$ constitutes one sentence. Therefore, if k is farther away from $l[t_{i,j}^m]$, then $hw[t_{i,j}^m](k)$ is smaller.

According to these definition, $hw[t_{i,j}^m](k)$ is given by the following equation:

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$hw[t_{i,j}^m](k) = \begin{cases} \frac{1}{2}(1+\cos 2\pi\frac{k-l[t_{i,j}^m]}{W}) \\ \qquad (\, a[t_{i,j}^m] < k < b[t_{i,j}^m]) \\ \frac{1}{2}D(1+\cos 2\pi\frac{k-l[t_{i,j}^m]}{W}) \\ \qquad (\, a[t_{i,j}^m] \geq k,\, b[t_{i,j}^m] \leq k) \end{cases} \quad (1)$$

$$(|\, k - l[t_{i,j}^m] \,| \leq \tfrac{W}{2},\ 0 \leq D \leq 1)$$

where D is a weighting parameter $(0 \leq D \leq 1)$ in order to consider the change in the content of a Web text. it has been suggested that this kind of function is the most useful function for extracting the influence of words [11]. The influence of $t_{i,j}^m$ exists only within the range of the window W and thus is defined by $|\ k - l[t_{i,j}^m]\ | \leq \frac{W}{2}$. We call $hw[t_{i,j}^m](k)$ the weighted Hanning window function (see Fig. 3).

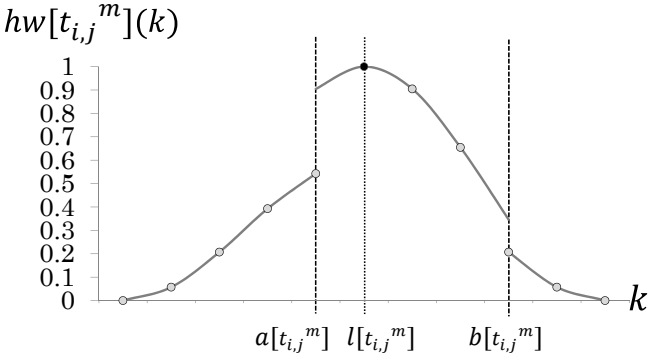

Fig. 3. Weighted Hanning window function

When we utilize the weighted Hanning window function, we must set two parameters W and D. In this paper, we set D to 0.6 and W to three times the average length of the sentences in each Web text. That is, these parameters are set in accordance with our past experiments on Web pages [10].

To calculate the word-density distribution, we sum all $hw[t_{i,j}^m](k)$ at each position k and then normalize $hw[t_{i,j}^m](k)$ by dividing each $\sum_j hw[t_{i,j}^m](k)$ by the largest among the values of the weighted Hanning window function $hw[t_{i,j}^m](k)$. Accordingly, the word-density distribution $hw[t_i^m](k)$ is defined by the following equation:

$$hw[t_i^m](k) = \frac{\sum_j hw[t_{i,j}^m](k)}{\max_k \sum_j hw[t_{i,j}^m](k)} \quad (2)$$

2) Construction of a Content-density Distribution by using the Word-density Distributions: Because we regard the content in a Web page as a set of words on the page, we combine the word-density distributions of the words in Q_r^m to construct a content-density distribution of Q_r^m as a set of word-density distributions in s_m. If we calculate the content-density distribution related to Q_r^m, we sum $hw[t_i^m](k)$ at each k and divide the result by the number of words in Q_r^m. This is because we simply combine the words into a content. We denote a content-density distribution of Q_r^m as follows:

$$hw[Q_r^m](k) = \begin{cases} \frac{1}{n}\sum_i hw[t_i^m](k) \\ \qquad \left(t_i^m \in Q_r^m, (\forall i)\ hw[t_i^m](k) > 0\right) \\ 0 \\ \qquad (others) \end{cases} \quad (3)$$

We extract the parts of a Web text to grasp the content by utilizing the content-density distribution.

3) Extraction of content from the Web Text by using the Content-density Distribution: If the value of the content-density distribution at a position k exceeds a threshold $\tau(0 \leq \tau \leq 1)$, we can describe the content related to the words in Q_r^m located at this position. We consider the threshold because there are some differences in construction between kinds of text.

For example newspaper texts have stricter configurations than Web texts. When we describe the case as $hasContent[Q_r^m](k) = 1$, we regard the position k as part of a portion included in a content. We can define the function $hasContent[Q_r^m](k)$ as follows:

$$hasContent[Q_r^m](k) = \begin{cases} 1 & (hw[Q_r^m](k) > \tau) \\ 0 & (others) \end{cases} \quad (4)$$

we need to set the threshold τ for Web texts. In this paper, we set the threshold as $\tau = 0.1$ for Web texts on the basis of our past work [12].

B. A Method Whose Information Unit is the Phrase

We have considered the word as the size of the information unit in Section III-A. Next, we regard the sentence as the most popular size of information unit to extract contents from a Web text. Therefore, we consider a phrase-based method, because a phrase represents an intermediate size of information unit between a word and a sentence.

We utilize a simple method whose information unit is the phrase by extracting phrases including the words forming the content Q_r^m in a Web text s_m. The portion of s_m where there is content Q_r^m is extracted by the method in the following two steps:

1) Detection of phrase separators in a Web text.
2) Extraction of phrases including the content.

We first detect the phrase separators, which are points at which phrases change in a Web text. Specifically, phrase separators are periods, exclamation marks, question marks and commas.

After that, we separate the Web text into phrases by using phrase separators. We extract the phrases including the words that form the content Q_r^m in the Web text s_m in separated into phrases that we mentioned above. We regard the p th phrase u_p in s_m as the portion that we can extract the content Q_r^m. All words of the subset Q_r^m are in u_p.

For example, there are two phrases u_1 and u_p in the text shown in Fig. 4. When we extract the content "Munich Germany" from the text, we check each phrase to determine whether the words "Germany" and "Munich" occur in one phrase. Accordingly, we will extract the content in

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

u_1, because u_1 includes "Germany" as well as "Munich". Conversely, we cannot find both words in u_p, since there is only "Germany" in the phrases.

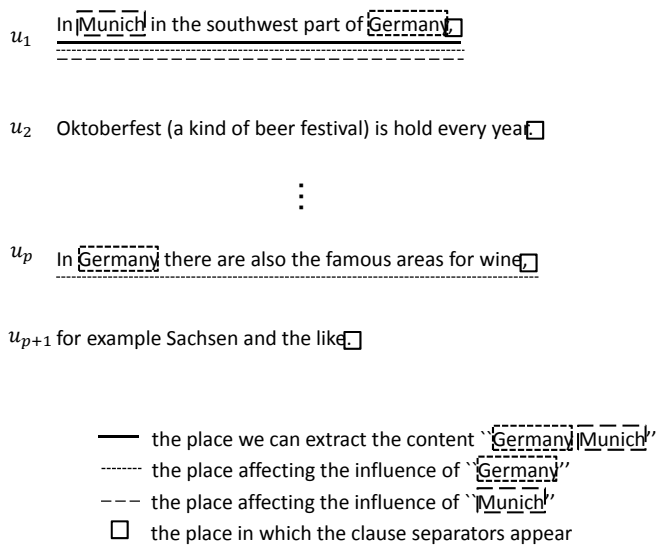

Fig. 4. An example of the method whose information unit is the phrase

IV. Experiment

In this section, we compare several content extraction methods in terms of the portion judged as the portion that includes a content, and we reveal which is the most suitable method for extracting content from a Web page. In order reach our ultimate goal, we need to formulate the evaluation measure at the outset. We explain the evaluation measure in Section IV-A.

Moreover, in order to carry out a comparative experiment, we have to create a test collection, because no specific test collection exists for this kind of experiment. Therefore, in Section IV-B we explain how to construct a test collection that consists of a Web page set, topics, and responses.

Furthermore, to determine which size of information unit is the best, we conduct our comparative experiment with content extraction methods by using the abovementioned measures, threshold τ, and test collection. We explain the comparative experiment in Section IV-C.

A. Formulation of Evaluation Measure

When users examine the content of a Web page as snippets or a visualization, only the portions including the desire content should be extracted comprehensively and precisely. Therefore, we define an evaluation measure based on the F-measure [13], with the following two considerations:

- How much of the Web page is covered by the content extracted (comprehensiveness).
- How much of the extracted part includes the content (preciseness).

We call this measure f_m the evaluation value. To calculate it, we extract a Web text from a Web page. The greater the average of the evaluation value $\overline{F}$, the better the method can grasp the content in Web texts.

By using the abovementioned Web page set, we can calculate is a kind of evaluation value for the Web text s_m as follows:

$$f_m = \frac{n(C_m \cap A_m)}{\frac{n(C_m)}{2} + \frac{n(A_m)}{2}} \tag{5}$$

In (5), C_m denotes the parts of s_m including content extracted by the method, A_m denotes the parts of s_m answered by the Web page set, and $C_m \cap A_m$ describes the parts satisfying the conditions on both C_m and A_m. In addition, $n(C_m)$ denotes the number of words in C_m, $n(A_m)$ denotes the number of words in A_m, and $n(C_m \cap A_m)$ denotes the number of words in $C_m \cap A_m$. We call A_m the answer part. The answer part is the portion of a Web text where the participants recognize the content.

Fig. 5 shows an example of these symbols for the content "Munich Germany" in s_m. In the figure, a method extracts the content as the highlighted part in Fig. 5(a), and then $n(s_m)$ is thirteen. In the same way, the answer part is extracted as the highlighted part in Fig. 5(b), and then $n(a_m)$ is eight. Based on the highlighted parts in Figs. 5(a) and 4(b), the part satisfying the conditions on both C_m and A_m is that shown in Fig. 5(c), and then $n(C_m \cap A_m)$ is eight. Thus, we can calculate $f_m = \frac{10}{\frac{16}{2} + \frac{10}{2}} = 0.77$.

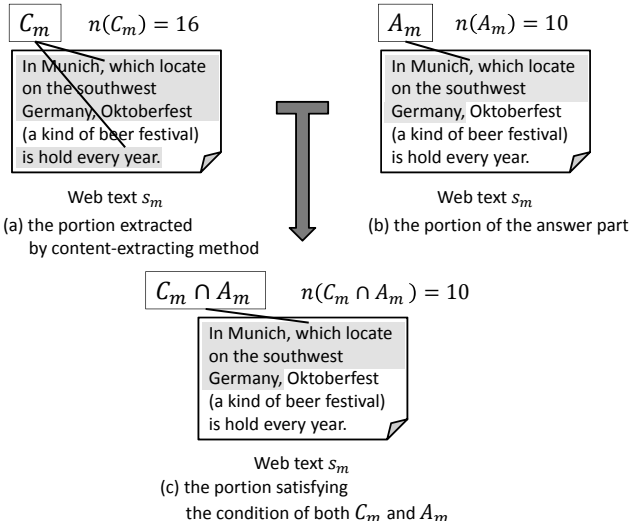

Fig. 5. Example of C_m and A_m for the content "Munich Germany"

Considering each f_m for the test collection described in Section IV-B, we can define the average of f_m as follows:

$$\overline{F} = \frac{1}{m} \sum_m f_m \tag{6}$$

B. Creation of Test Collection for Comparative Experiment

We created the topics of the test collection for this experiment, because the content extraction method should reflect the judgments of the content by the users and there are no test collections reflecting such a viewpoint.

To create the topics, we utilized the known items of NTCIR-5 WEB Navigational Retrieval Subtask 2 [14]. These items consist of some searching situation on the Web and query keywords utilized in each situation. We randomly chose 12 tasks with a combination of search terms specified by the known items. Tasks specified by the known items means that participants can understand what the content is

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
EVALUATING THE SIZE OF THE INFORMATION UNIT TO EXTRACT THE CONTENTS FROM WEB PAGES

Information unit	Word	Phrase	Sentence
$\overline{F}$	**0.163**	0.091	0.107

and the documents explaining each task. Because multiple participants can judge the same content with the same understanding, we chose the tasks specified by the known items. The combination of search terms is the content consisting of more than two words. Because we define the content to consist of more than two words, we utilize a combination of search terms for our experiment. Since we should have nonarbitrary Web texts instead of artificial ones in the existing test collection, we submit the content to the Google AJAX Search API [15] for extraction of Web texts. For each content, we utilize the first 20 hits of the Web text in the search results of the Google AJAX Search API. We gathered 205 Web texts for our experiment, because there are Web pages without Web text [1].

For our comparative experiment, we asked our participants to use Web texts to create a test collection without any blemishes. In this study, three participants chose the answer parts of each Web text, because this let us reduce the biases in the judgments by the participants. Specifically, we first asked them whether the content was included in a Web text. If they thought that the content was in the Web text, they chose the answer parts by word. We regarded the parts chosen by more than two participants as the answer parts of the test collection.

C. Comparative Experiment

In this subsection, we compare $\overline{F}$ of the methods explained in Section III with $\overline{F}$ of the method using sentence-based information unit by using the test collection created in Section IV-B. As mentioned in Section IV-A, we can regard the method with a certain size of information unit can grasp the content, if $\overline{F}$ of the method with the size of information unit is high.

We utilize the $\overline{F}$ of each method in Section III and the method using the sentence-based information unit to compare the content extracted by the methods. The latter method resembles the method using the phrase-based information unit, but it differs in that we do not utilize commas as separators in the sentence-based method. Table I lists the results of the evaluation in which the average evaluation value $\overline{F}$ was calculated for each Web text by each method.

In this experiment, the average evaluation value $\overline{F}$ calculated by the word-based method was the highest of all. This is because the method using the word-based information unit can extract the content more flexibly than any other method beyond the range of sentences and phrases. In fact, 96% of the Web texts in the Web page set of the test collection included more than two contents. As a result, we found that the word is the best size of information unit for extracting content from a Web page.

[1]For example, we regard a page consisting of a flash movie as one that has no Web text.

V. CONCLUSION

In this paper, we discussed whether the sentence' is the best size of information unit by comparing the smaller sizes of information unit (e.g., the phrase and the word) in the method for extracting the contents from Web pages. In order to approach the judgment of users, we compared the method using each size of information unit with our test collection. As a result, we understand that the word is the best size of information unit for extracting contents from Web pages.

In addition, with our previous proposal and the contribution of this study, we expect much further accuracy improvement in extracting contents, because we can set a more suitable threshold. We aim in the future to develop a system for supporting user searches on the Web by our content extraction method.

ACKNOWLEDGMENT

The authors would like to thank Prof. Akiyo Nadamoto, who provided us with valuable comments at the 2012 Academic Year IPSJ Kansai Branch Convention.

REFERENCES

[1] C. D. Manning, P. Raghavan, and H. Schuetze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
[2] W. Sunayama and Y. Nishihara, "Text Visualization Service for Creating Comprehended Texts," in *Knowledge-Based and Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science. Springer, 2011, vol. 6883, pp. 265–274.
[3] A. Woodruff, A. Faulring, R. Rosenholtz, J. Morrsion, and P. Pirolli, "Using thumbnails to search the Web," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2001, pp. 198–205.
[4] "Yahoo! Answers," http://answers.yahoo.com/.
[5] "Wikipedia," http://www.wikipedia.org/.
[6] A. Keyaki, K. Hatano, and J. Miyazaki, "Relaxed global term weights for XML element search," in *Proceedings of the 9th international conference on Initiative for the evaluation of XML retrieval: comparative evaluation of focused retrieval*. Springer-Verlag, 2011, pp. 71–81.
[7] W.-S. Li, K. S. Candan, Q. Vu, and D. Agrawal, "Retrieving and organizing web pages by "information unit"," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 230–244.
[8] Y. Lv and C. X. Zhai, "Positional language models for information retrieval," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 299–306.
[9] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 275–281.
[10] S. Kitahara, K. Tamura, and K. Hatano, "Extraction of the contents in the web texts by content-density distribution," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 3, no. 2, pp. 108–120, 2011.
[11] K. Kise, H. Mizuno, M. Yamaguchi, and K. Matsumoto, "On the Use of Density Distribution of Keywords for Automated Generation of Hypertext Links from Arbitrary Parts of Documents," in *Proceedings of the Fifth International Conference on Document Analysis and Recognition*. IEEE Computer Society, 1999, pp. 301–304.
[12] S. Kitahara and K. Hatano, "A study of extracting contents on the Web text based on the position of words (in Japanese)," in *Proceedings of The 2012 Academic Year IPSJ Kansai Branch Convention*. IPSJ Kansai Branch, 2012.
[13] R. Baeza-Yates and G. Navarro, *Modern Information Retrieval*, 2nd ed. Addison-Wesley, 2011.
[14] K. Oyama, M. Takaku, H. Ishikawa, A. Aizawa, and H. Yamana, "Overview of the NTCIR-5 WEB Navigational Retrieval Subtask 2 (Navi-2)," *NTCIR-5*, 2005.
[15] "Google Web Search API Developer's Guide," https://developers.google.com/web-search/docs/.