The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

Web News Summary System with Clustering Algorithm to Identify Same Article

Ayahiko Niimi
Department of Media Architecture
Future University-Hakodate
116-2 Kamedanakano-cho, Hakodate-shi,
Hokkaido 041-8655, JAPAN

Yusaku Saito
Department of Complex Systems
Future University-Hakodate
116-2 Kamedanakano-cho, Hakodate-shi,
Hokkaido 041-8655, JAPAN

Osamu Konishi
Department of Complex Systems
Future University-Hakodate
116-2 Kamedanakano-cho, Hakodate-shi,
Hokkaido 041-8655, JAPAN

Abstract

We propose the system that offers only the article that is the relation to topics to the user in this research. When the user wants to read the article that is the relation to topics, the user must click the link to the article. Therefore, it is difficult for the user to read only the article related to topics. Moreover, there is the article that is similar to each other content or article. Therefore, user must read the article that is similar to other article. We propose the algorithm to find similar articles. For the proposed system, we use the feature of reported articles. There is an outline of the entire article at the beginning of reported articles.

Keyword:

Web news, Natural Language Processing, Morphological Analysis, Summary Generation

1 Introduction

The web news sites become popular, but they are not understood easily. On the other hand, the newspaper and the television are comprehensible. We think that it is a cause that the web news is not arranged. The portal site (such as Yahoo JAPAN News, etc) is news collection site. If news is a little related to other news, it becomes related news and is made a link to related news. Moreover, the portal site publishes the article on a lot of newspapers and news agencies. Therefore, there are a lot of related contents, but it is difficult to read articles that user actually wants to read.

We propose the system that offers only the article that is the relation to topics to the user in this research. When the user wants to read the article that is the relation to topics, the user must click the link to the article. Therefore, it is difficult for the user to read only the article related to topics. Moreover, there is the article that is similar to each other content or article. Therefore, user must read the article that is similar to other article. We propose the algorithm to find similar articles. For the proposed system, we use the feature of reported articles. There is an outline of the entire article at the beginning of reported articles.

2 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, 7]

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 1)

Table 1: 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.

3 Proposed System

This chapter describes a proposed algorithm for low related articles and similar articles are deleted from the list of the news of topics.

This system extracts only a high relativity article from the article list including high/low relativity articles about topics that the user wants to learn. Moreover, the article on a similar contents are searched out, and deleted. In this paper, we decide a high relativity article which includes main content as related to topics. Moreover, we think that same information of the article with the high similarity is contained in other articles.

For the necessity for confirming the content clicking the link to the article to know the relativity of the article to exist, and to read only a high relativity article, it can be said that it is inconvenient under the present situation. Moreover, because a lot of similar articles exist, too the possibility of reading the article on almost the same content is high. It is thought that the site where a lot of volume of information with high possibility that the problem becomes a relief exists is the best for the verification of this system. Yahoo! JAPAN news has a lot of topics, and its source are from many newspaper sites. So, in this paper, we discuss "Yahoo! JAPAN news" site for experiment. It paid attention to the tendency that the entire summary was written in the part at the beginning about the news article when the proposed system was designed. Because the point of the entire article has been brought together in the sentence at the beginning, the outline can be understood. Therefore, we use beginning sentences of article for analysis. The system is mounted by the Java application.

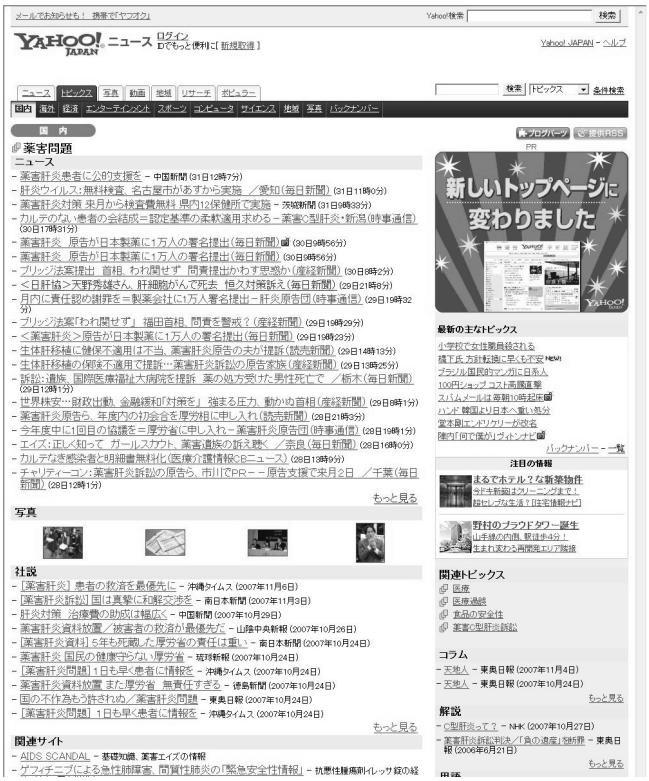

Figure 1: Topics list of Yahoo! JAPAN News

We describe the proposed algorithm of extracting high relativity article from the article group of topics of Yahoo! JAPAN news, and deleting URL of a similar article.

The flow of the algorithm is shown below.

1. input top-page URL of topics
2. get the beginning sentence and the delivery date
3. process morphological analysis using MeCab
4. extract keywords
5. extract high relativity articles
6. delete similar articles
7. outout results

The user acquires URL of topics that the user wants to learn from the top page of topics of news and the

program outputs URLs to the text file. At this time, we think the article only in the image thought that the content is low relativity, then that URL is excluded. Moreover, the link is not acquired when there is a page such as other newspapers because it targets only Yahoo! JAPAN news in this paper.

It accesses acquired URL, and the sentence to the punctuation of the start of the text and the delivery date is acquired. Because the noun decreases when one sentence of the start is short, the following punctuation is acquired in addition, and it outputs it to the text file for 20 characters or less. Moreover, delivery time of the article is acquired, and it outputs it to the text file with URL of the article. But, for the situation that there is no abstract sentences at the beginning. Then, when the sentences are not extracted when it is fewer than the threshold number at the beginning with the number of strokes to the punctuation of the sentence, we extract sentences until the following punctuation. It sets it to 25 characters as a result of experimenting on the number of strokes that becomes a standard.

Using MeCab that is the morphological analysis tool, the morphological analysis of the sentence is done at the beginning, and the result of the acquired each article is output to one text file.

The part of speech that doesn't show the feature of the article easily is excluded from the text file that does the morphological analysis and is output, and only a part of noun is extracted. The extracted part of speech is output to the text file.

The extracted part of speech is sorted to the lexical order, and a lot of consecutive nouns are found. It thinks this noun to be a noun that characterizes the relativity of topics, and only the article with this noun is output to the text file. However, when the same in one article two nouns or more exist, it counts with one. Moreover, the article not extracted is output to another text file. We use of the expression agreement technique to consider the number of extracted words. The following equation is used for the expression agreement technique. [9] In the equation, x is a number of words of sentences X that become standards, y is a number of words of sentences Y that become the object of comparisons, and m is a number of words that appears in both X and Y. It experimented to set the evaluation value as well as algorithm 1. As a result, if Score(X, Y) is larger than 60%, it is judged that two articles are similar , and deletes an old article.

$$\text{Score(X,Y)} = \frac{\frac{m}{x} + \frac{m}{y}}{2} \times 100 \qquad (1)$$

We think that it is rare that same topics exist in two days or more. It is based on the newest article in the extracted relativity and high article. Nouns that are to the article on the day before are compared. If the noun more than the evaluation value of nouns that exist in the article that became a standard exists in the article on the object of comparison, it is judged that two articles are similar and deletes an old article. Next, a new similar article is secondarily operated, and repeated this operation. We show its example. In Table 2, the alphabet presents one article. If D and E, G and H, J and K, L and M are judged as same contents. In Table 2, "similer to" means"judged as same contents". Using our proposed algorithm, the comparison is done in order of (A,B) → (A,C) → (A,D) → (A,E) → (A,F) → (B,D) → (B,E) → (B,G) → (H,I) → (L,M). When (A,C), (A,D), (A,F), (D,E), and (D,G) are compared, the article on C, D, F, E, and G is deleted. Therefore, A, B, D, H, I, J, K, L, and M are extracted. The extracted article is output to the text file. Moreover, the deleted article is output to another text file.

Table 2: date and articles

date	articles	similer to
12, Feb.	A	-
12, Feb.	B	-
11, Feb.	C	A
11, Feb.	D	-
11, Feb.	E	D
11, Feb.	F	A
10, Feb.	G	D
9, Feb.	H	D
8, Feb.	I	-
6, Feb.	J	A
31, Jan.	K	A
15, Jan.	L	B
15, Jan.	M	B

As an output result, the extracted URLs are written to the html file with the title of the article, newspaper site name in delivery origin, delivery date, extracted nouns and opening sentences. Moreover, URL of low relativity articles and similar articles are output to the text file respectively. (See Fig. 2)

4 Experimental Results

This section describes the experimental methodology and the results. We use two topics, " the damage

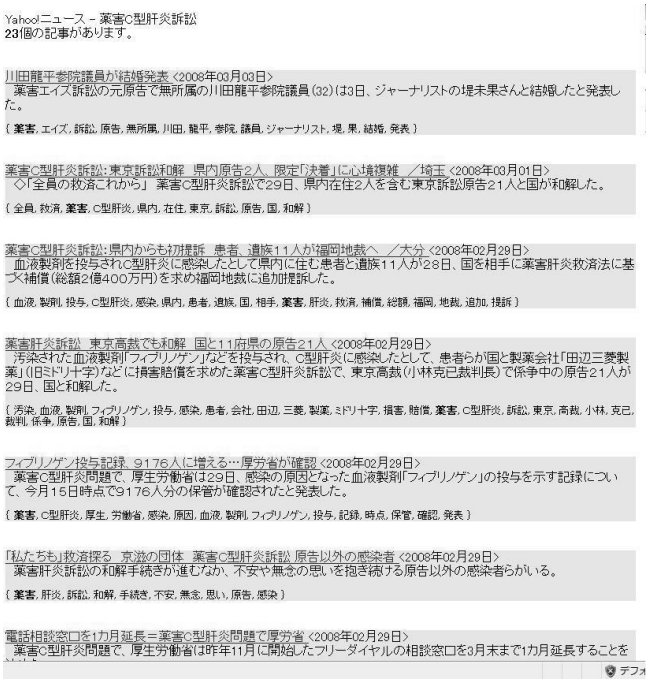

Figure 2: Output of proposed system

of crops by chemicals hepatitis C lawsuit" and "the Aegis destroyer collision in the experiment".

64 articles existed in topics of the damage of crops by chemicals hepatitis C prosecution when experimenting. 59 pieces in 64 pieces have been extracted because it was judged that five pieces were image links.

Using the topics of the Aegis destroyer collision in the experiment, as a result of the URL extraction, 336 URL has been extracted. However, 20 articles have been deleted from making the system work to doing this verification. The article that has been deleted is disregarded in the experiment. Therefore, it experimented assuming that 316 URL was extracted.

4.0.1 Result of "the damage of crops by chemicals hepatitis C lawsuit"

The ratio of the total of the number of articles to be able to delete the number and the relativity of the article to which relativity was able to be extracted high low became 67% among the numbers of URL extracted to the start. Moreover, the ratio of the total of the number of articles that were able to be deleted by a similar article became 100% with the number of articles that were able to be extracted in an article not similar among the numbers of articles that were judged that relativity was high and extracted. (See Table 3)]

Table 3: Extracted result of " the damage of crops by chemicals hepatitis C lawsuit"

	correct	miss
high related articles extract	67 %	33 %
similar articles extract	100 %	0 %

4.0.2 Result of "aegis destroyer collision"

The ratio of the total of the number of articles to be able to delete the number and the relativity of the article to which relativity was able to be extracted high low became 79% among the numbers of URL extracted to the start. Moreover, the ratio of the total of the number of articles that were able to be deleted by a similar article became 79% with the number of articles that were able to be extracted in an article not similar among the numbers of articles that were judged that relativity was high and extracted. (See Table 4) Moreover, the execution time from the input of URL to the output of the result was 55 seconds of two minutes. (We used a notePC with Intel Core2 Duo CPU(1.2GHz, 1GB RAM, Windows XP.)

Table 4: Extracted result of "aegis destroyer collision"

	correct	miss
high related articles extract	79 %	21 %
similar articles extract	79 %	21 %

5 Discussions

It can be said that the algorithm of the relativity judgment proposes in the present study was able surely to pick up the noun that becomes the key to topics. However, there was a difference at the positive detection rate of a relativity judgment of the result of targeted damage of crops by chemicals hepatitis C lawsuit and Aegis destroyer collision because of being only not existing at the beginning in the sentence, and the logic that one noun had extracted of the deletion. Similarly, there was a difference at the positive detection rate of the similarity judgment in the algorithm of the similarity judgment. There was information that had been described only to the deleted old article though

this system was an algorithm that old when nouns were compared, and it was judged that it resembled it deletes the article, too. The method such as deleting a short article is devised as an idea that improves this problem in consideration of the entire amount of the sentence of each article.

6 Future Works

It is thought that the extraction result in which accuracy is high can be generated by considering the shake of the synonym and the mark of the morpheme not considered in the algorithm of this system, and the appearance order. In assumption as the logic considered that relativity is high if the noun that becomes a standard is expanded because only it doesn't exist at the beginning in the sentence, and the logic in a relativity judgment that one noun extracted of the deletion, and there is one, is the false detection that deletes a relativity and high article?It is thought that it is possible to eliminate it. Moreover, it compares after the noun of frequent occurrence is excluded when similar judged, and there is a possibility to understand the noun that the content of the article on the object or more characterizes.

References

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

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

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

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

[5] Yahoo! NEWS, http://headlines.yahoo.co.jp/hl

[6] Juman, http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[7] MeCab, http://mecabsourceforge.jp/

[8] Ohtake, K., Okamoto, D., Kodama, M., Masuyama, S.: A Summarization System YELLOW for Japanese Newspaper Articles, IPSJ Magazine, Vol.43 No.SIG02, TOD13,pp.37–47 (2002). (In Japanese)

[9] Fujie, Y., Watabe, H., Kawaoka, T.: Article classification method using the calculation of the degree of association between articles and category attributes extracted from Web information, The 21st Annual Conference of the Japanese Society for Artificial Intelligence, 1G3-5 (2007). (In Japanese)

[10] Iguchi, T., Kaminaga, H., Yokomaya, S., Miyadera, Y., Nakamura, S.: Proposal of a Web Exploring Support Method Focusing on Topic Transition Processes, IEICE Technical Report, ET2007-54, pp.33–38 (2007). (In Japanese)

[11] Matsuo,. Y., Ishizuka, M.: Keyword Extraction from a Document using Word Co-occurrence Statistical Information, Journal of Japanese Society for Artificial Intelligence, Vol.17,No.3,pp.217–223 (2002). (In Japanese)

[12] Toda, H., Kataoka, R., Kitagawa, H.: Clustering News Articles using Named Entities, IPSJ SIG Technical Report, 2005-DBS-137, pp.175-181 (2005). (In Japanese)

[13] KNP, http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html