Fourth International Workshop on Computational Intelligence & Applications
IEEE SMC Hiroshima Chapter, Hiroshima University, Japan, December 10 & 11, 2008

News summary system for Web news site

Ayahiko Niimi, Yusaku Saito, Osamu Konishi
School of Systems Information Science, Future University-Hakodate
116–2 Kamedanakano-cho, Hakodate-shi, Hokkaido, 041–8655 Japan
email: niimi@fun.ac.jp

***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.**

I. Introduction

The web news sites become popular, but the whole image of news is not understood easily because one news has divided into some articles. 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 [1]) 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 large number of newspapers and news agencies. Therefore, there are a large number of related contents, but it is difficult to read articles that user actually wants to read. The text mining is effective for this problem [2], [3], [4]. There are some previous work for the summary of the news article [5], [6], [7]. However, it doesn't go well in the proposal method because the news article on Web is short. There is previous work for web document focusing on topic transition processes [8]. We paid attention to a time connection of the article.

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.

II. 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 large number 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 large number 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 large number 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.

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 [9]
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, the some sentences until punctuation are extracted when the first sentence is shorter than the threshold. It sets it to 25 characters as a result of experimenting on the number of strokes that becomes a standard.

The extracted part of speech is sorted to the lexical order, and a large number 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 [10]. 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.

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

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, the nouns on the day before article 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.

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. 1)

III. Conclusion

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. We think about the use of the parsing tool [11]. Because a relativity judgment is an algorithm that one noun extracted of the deletion only by not existing at the beginning in the sentence, the number of nouns that become standards of the comparison is increased. It is thought that the false detection that deletes a relativity and high article can be eliminated by assuming the algorithm considered that relativity is high if there is those nouns. Moreover, at similar articles detection, articles are compared


Yahoo!ニュース - 薬害C型肝炎訴訟
23個の記事があります。

川田龍平参院議員が結婚発表 <2008年03月03日>
薬害エイズ訴訟の元原告で無所属の川田龍平参院議員(32)は3日、ジャーナリストの堤未果さんと結婚したと発表した。
{ 薬害, エイズ, 訴訟, 原告, 無所属, 川田, 龍平, 参院, 議員, ジャーナリスト, 堤, 果, 結婚, 発表 }

薬害C型肝炎訴訟:東京訴訟和解 県内原告2人、限定「決着」に心境複雑 /埼玉 <2008年03月01日>
◇「全員の救済これから」 薬害C型肝炎訴訟で29日、県内在住2人を含む東京訴訟原告21人と国が和解した。
{ 全員, 救済, 薬害, C型肝炎, 県内, 在住, 東京, 訴訟, 原告, 国, 和解 }

薬害C型肝炎訴訟:県内からも初提訴 患者、遺族11人が福岡地裁へ /大分 <2008年02月29日>
血液製剤を投与されC型肝炎に感染したとして県内に住む患者と遺族11人が28日、国を相手に薬害肝炎救済法に基づく補償(総額2億400万円)を求め福岡地裁に追加提訴した。
{ 血液, 製剤, 投与, C型肝炎, 感染, 県内, 患者, 遺族, 国, 相手, 薬害, 肝炎, 救済, 補償, 総額, 福岡, 地裁, 追加, 提訴 }

薬害肝炎訴訟 東京高裁でも和解 国と11府県の原告21人 <2008年02月29日>
汚染された血液製剤「フィブリノゲン」などを投与され、C型肝炎に感染したとして、患者らが国と製薬会社「田辺三菱製薬」(旧ミドリ十字)などに損害賠償を求めた薬害C型肝炎訴訟で、東京高裁(小林克己裁判長)で係争中の原告21人が29日、国と和解した。
{ 汚染, 血液, 製剤, フィブリノゲン, 投与, 感染, 患者, 会社, 田辺, 三菱, 製薬, ミドリ十字, 損害, 賠償, 薬害, C型肝炎, 訴訟, 東京, 高裁, 小林, 克己, 裁判, 係争, 原告, 国, 和解 }

フィブリノゲン投与記録、9176人に増える…厚労省が確認 <2008年02月29日>
薬害C型肝炎問題で、厚生労働省は29日、感染の原因となった血液製剤「フィブリノゲン」の投与を示す記録について、今月15日時点で9176人分の保管が確認されたと発表した。
{ 薬害, C型肝炎, 厚生, 労働省, 感染, 原因, 血液, 製剤, フィブリノゲン, 投与, 記録, 時点, 保管, 確認, 発表 }

「私たちも」救済探る 京滋の団体 薬害C型肝炎訴訟 原告以外の感染者 <2008年02月29日>
薬害肝炎訴訟の和解手続きが進むなか、不安や無念の思いを抱き続ける原告以外の感染者らがいる。
{ 薬害, 肝炎, 訴訟, 和解, 手続き, 不安, 無念, 思い, 原告, 感染 }

電話相談窓口を1カ月延長=薬害C型肝炎問題で厚労省 <2008年02月29日>
薬害C型肝炎問題で、厚生労働省は昨年11月に開始したフリーダイヤルの相談窓口を3月末まで1カ月延長することを

Fig. 1. Output of proposed system

after the noun of high frequent occurrence is excluded, and there is a possibility the keywords can be extracted.

References

[1] Yahoo! NEWS, http://headlines.yahoo.co.jp/hl
[2] 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)
[3] 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)
[4] Nagata, M., Taira, H.: Text Classification - Showcase of Learning Theories -, IPSJ Magazine, Vol.42 No.1,pp.32–37 (2001). (In Japanese)
[5] 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)
[6] 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)
[7] 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)
[8] 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)
[9] MeCab, http://mecabsourceforge.jp/
[10] 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)
[11] KNP, http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html