Summary of Web news based on relation between articles and identity frequency of contents

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 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. We explain the flow of the system. First, the system extracts top part of the articles on topics list. Next, the system analyzes the articles to extract morpheme by using MeCab. Next, the system finds the nouns with high occurrence rate. Then, the system deletes the article without high occurence rate nouns. Next, the system executes processing that deletes the article that the similar content to other article by using the data of the article of the publishing date. We think that it is rare that same topics exist in two days or more. Therefore, the system compares the noun that exists in a new article and the noun that exists in the article on the day before of the publishing day, the system counts the number which the same noun exists. If the number exceeds the threshold value, the system deletes the old article.

We experiment for the topic of "aegis destroyer collision" on Yahoo! JAPAN News by using proposed system. The articles not related to topics of 15% had existed before executing our system. The articles not related to topics of 4% existed after using our system. The articles with the similar content to other article of 37% had existed before using our system. The articles with the similar content to other article of 5% existed after using our system. However, all processing did not succeed. We think that it is a problem to judge the relation only from one noun. Moreover, we think that there is a problem in the algorithm that always deletes old articles. We think that it is important to improve the problem for improvement of result.

II. 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 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.

III. 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. (See Fig. 1) 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 one sentence at the beginning of article for analysis. So, the required processing time can be expected to be shortened to the morphological analysis greatly by assuming the object of the analysis to be one sentence at the beginning. The system is mounted by the Java application.

A. Algorithm 1

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.

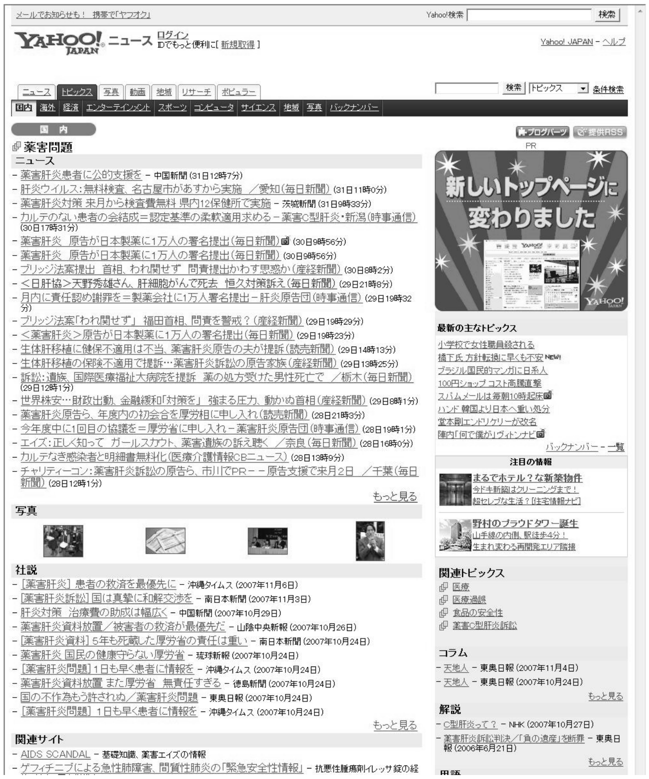
Fig. 2 shows a screenshot of the proposed system.

The flow of the algorithm 1 is shown below.

1) input top-page URL of topics
2) get the beginning sentence and the delivery date
3) process morphological analysis
4) extrace 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.



薬害問題

ニュース

- 薬害肝炎患者に公的支援を - 中国新聞 (31日12時7分)
- 肝炎ウイルス：無料検査、名古屋市があすから実施　／愛知（毎日新聞）(31日11時0分)
- 薬害肝炎対策 来月から検査費無料 県内12保健所で実施 - 茨城新聞 (31日9時33分)
- カルテのない患者の会結成＝認定基準の柔軟適用求める－薬害C型肝炎・新潟（時事通信）(30日17時31分)
- 薬害肝炎　原告が日本製薬に1万人の署名提出（毎日新聞）(30日9時56分)
- 薬害肝炎　原告が日本製薬に1万人の署名提出（毎日新聞）(30日9時56分)
- ブリッジ法案提出　首相、われ関せず　問責提出かわす思惑か（産経新聞）(30日8時2分)
- ＜日肝協＞天野秀雄さん、肝細胞がんで死去　恒久対策訴え（毎日新聞）(29日21時8分)
- 月内に責任認め謝罪を＝製薬会社に1万人署名提出－肝炎原告団（時事通信）(29日19時32分)
- ブリッジ法案「われ関せず」　福田首相、問責を警戒？（産経新聞）(29日19時29分)
- ＜薬害肝炎＞原告が日本製薬に1万人の署名提出（毎日新聞）(29日19時23分)
- 生体肝移植に健保不適用は不当、薬害肝炎原告の夫が提訴（読売新聞）(29日14時13分)
- 生体肝移植の保険不適用で提訴…薬害肝炎訴訟の原告家族（産経新聞）(29日13時25分)
- 訴訟：遺族、国際医療福祉大病院を提訴　薬の処方受けた男性死亡で　／栃木（毎日新聞）(29日12時1分)
- 世界株安…財政出動、金融緩和「対策を」　強まる圧力、動かぬ首相（産経新聞）(29日8時1分)
- 薬害肝炎原告ら、年度内の初会合を厚労相に申し入れ（読売新聞）(28日21時3分)
- 今年度中に1回目の協議を＝厚労省に申し入れ－薬害肝炎原告団（時事通信）(28日19時1分)
- エイズ：正しく知って　ガールスカウト、薬害遺族の訴え聴く　／奈良（毎日新聞）(28日16時0分)
- カルテなき感染者と明細書無料化（医療介護情報CBニュース）(28日13時9分)
- チャリティーコン：薬害肝炎訴訟の原告ら、市川でPR－－原告支援で来月２日　／千葉（毎日新聞）(28日12時1分)

もっと見る

写真

社説

- [薬害肝炎] 患者の救済を最優先に - 沖縄タイムス (2007年11月6日)
- [薬害肝炎訴訟] 国は真摯に和解交渉を - 南日本新聞 (2007年11月3日)
- 肝炎対策　治療費の助成は幅広く - 中国新聞 (2007年10月29日)
- 薬害肝炎資料放置／被害者の救済が最優先だ - 山陰中央新報 (2007年10月26日)
- [薬害肝炎資料] 5年も死蔵した厚労省の責任は重い - 南日本新聞 (2007年10月24日)
- 薬害肝炎 国民の健康守らない厚労省 - 琉球新報 (2007年10月24日)
- [薬害肝炎問題] 1日も早く患者に情報を - 沖縄タイムス (2007年10月24日)
- 薬害肝炎資料放置 また厚労省　無責任すぎる - 徳島新聞 (2007年10月24日)
- 国の不作為もう許されぬ／薬害肝炎問題 - 東奥日報 (2007年10月24日)
- [薬害肝炎問題] 1日も早く患者に情報を - 沖縄タイムス (2007年10月24日)

もっと見る

関連サイト

- AIDS SCANDAL - 基礎知識、薬害エイズの情報
- ゲフィチニブによる急性肺障害、間質性肺炎の「緊急安全性情報」 - 抗悪性腫瘍剤イレッサ錠の経

最新の主なトピックス

小学校で女性職員殺される
橋下氏 方針転換に早くも不安 NEW!
ブラジル国民的マンガに日系人
100円ショップ コスト高騰直撃
スパムメールは毎朝10時起床
ハンド 韓国より日本へ重い処分
堂本剛エンドリケリーが改名
陣内「何で僕が」ヴィトンナビ

バックナンバー - 一覧

注目の情報

まるでホテル？な新築物件
今ドキ新築はクリーニングまで！
超セレブな生活？[住宅情報ナビ]

野村のプラウドタワー誕生
山手線の内側、駅徒歩4分！
生まれ変わる再開発エリア隣接

関連トピックス

医療
医療過誤
食品の安全性
薬害C型肝炎訴訟

コラム

- 天地人 - 東奥日報 (2007年11月4日)
- 天地人 - 東奥日報 (2007年10月24日)

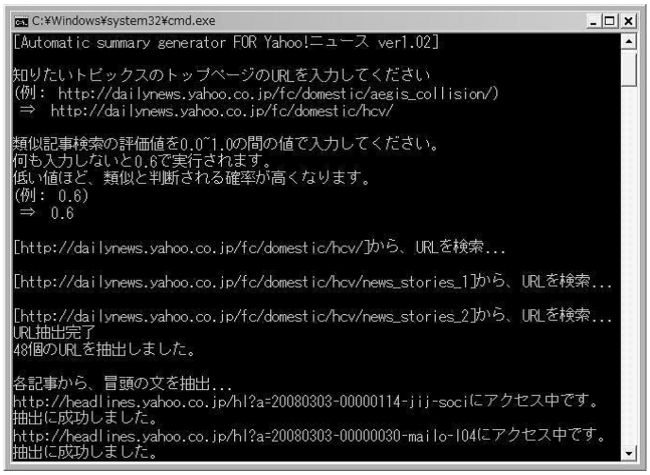
もっと見る

解説

- C型肝炎って？ - NHK (2007年10月27日)
- 薬害肝炎訴訟判決／「負の遺産」を断罪 - 東奥日報 (2006年6月21日)

もっと見る

Fig. 1.   Topics list of Yahoo! JAPAN News



```
[Automatic summary generator FOR Yahoo!ニュース ver1.02]

知りたいトピックスのトップページのURLを入力してください
(例:  http://dailynews.yahoo.co.jp/fc/domestic/aegis_collision/)
 ⇒  http://dailynews.yahoo.co.jp/fc/domestic/hcv/

類似記事検索の評価値を0.0~1.0の間の値で入力してください。
何も入力しないと0.6で実行されます。
低い値ほど、類似と判断される確率が高くなります。
(例:  0.6)
 ⇒  0.6

[http://dailynews.yahoo.co.jp/fc/domestic/hcv/]から、URLを検索...

[http://dailynews.yahoo.co.jp/fc/domestic/hcv/news_stories_1]から、URLを検索...

[http://dailynews.yahoo.co.jp/fc/domestic/hcv/news_stories_2]から、URLを検索...
URL抽出完了
48個のURLを抽出しました。

各記事から、冒頭の文を抽出...
http://headlines.yahoo.co.jp/hl?a=20080303-00000114-jij-sociにアクセス中です。
抽出に成功しました。
http://headlines.yahoo.co.jp/hl?a=20080303-00000030-mailo-l04にアクセス中です。
抽出に成功しました。
```

Fig. 2.   Screenshot of proposed system

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.

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 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 II, 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 II, "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.

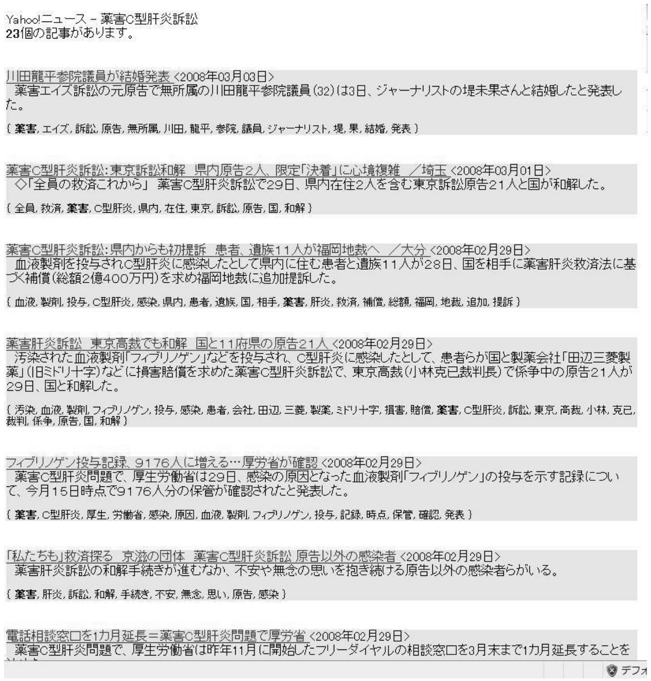
TABLE II
DATE AND ARTICLES

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. 3)

B. Algorithm 2

Algorithm 1 was mounted, and it experimented. As a result, the article that the entire summary was not written in the top sentence was more than the expectation. Because there is no noun that characterizes the article to the sentence at the beginning, even if the relativity of the content was high, it was occasionally deleted. Moreover, when a similar article is extracted based on the number of nouns, if a base article has only little words, the possibility to judged "similar" become high. The possibility to be judged the resemblance when a similar article is extracted oppositely based on the article that the number of nouns is a lot of lowers. The algorithm that improves these two points is assumed to be algorithm 2, and the improvement is described in this section.



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(小林克已裁判長)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, 克已, 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. 3. Output of proposed system

For the situation that there is an article that putting the entire article together is not written at the beginning, the sentences are not extracted when it is fewer than the constancy number at the beginning, with the number of strokes to the punctuation of the sentence, and sentences to the following punctuation are extracted in improved algorithm 2. It sets it to 25 characters as a result of experimenting on the number of strokes that becomes a standard. Details of the experiment are described in section IV-B.

In algorithm 1, when the algorithm judges whether the article is similar, the algorithm was not considered the number of extracted words of the beginning sentences. However, we use of the expression agreement technique in algorithm 2 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. Details of the experiment are described in section IV-B.

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

IV. Experimental Results

This section describes the experimental methodology and the results.

A. Experiment 1

In this section, the evaluation value setting of algorithm 1 and the experiment by algorithm 1 are described.

1) Evaluation threshold setting of algorithm 1: The evaluation threshold used when a similar article is retrieved is decided. 10 groups (20 articles) to which the content was similar were searched out from genres of the education, sports, and the science, etc in Yahoo! JAPAN news. And, the morphological analysis of those articles was done, and nouns were compared. The ratio with the noun to which an old article was very common was examined among nouns of the article in new one.

Fig. 4 shows this result. The alphabet shows the group of each article. The mean value of the agreement rate became 69.5%. It can be judged that 90 % is similar with the noun of 50% or more from this result. Therefore, in algorithm 1, if an old article having more than 50% same words which are noun of new article, it is judged that two articles are similar.

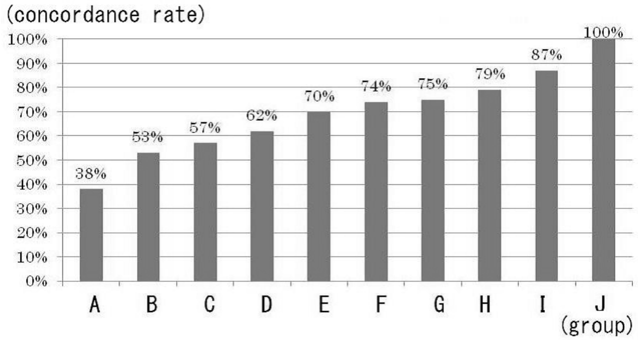


Fig. 4. concordance rate

2) URL extraction: It experimented in the experiment for topics of the damage of crops by chemicals hepatitis C lawsuit now in Yahoo! JAPAN news. 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.

3) Delete low relativity articles and extract high reratibity articles: The appearance frequency has extracted 10 nouns in a lot of order from 59 articles which extracted. Table III shows the result. The article with "Damage of crops by chemicals" extracted a lot in the first sentence is judged to be an article that relativity is high in the experiment and we extracts it. Even if the relativity of the content was high because at the beginning, there was no noun that the article that the entire summary was not written in the sentence characterizes the article more than the expectation to the sentence at the beginning, it was occasionally deleted. (Former sentences are English translations of the word to which the word in the table is extracted for Japanese.) (See Table IV)

TABLE III
OCCURRENCE RATE OF NOUNS

nouns	times
medication scandal	44
lawsuit	37
hepatitis C	29
hepatitis	25
libelant	22
government	20
settlement	19
drug product	16
settle	15
medicine manufacture	14

TABLE IV
RESULTS OF EXTRACTED AND DELETED ARTICLES

	extracted	deleted
high related articles	44	8
low related articles	0	7

4) Delete similar articles and extract articles: The article that was judged that there was no similarity as a result of processing it for 44 articles that were judged that relativity was high, and extracted, and extracted became 22. (See Table V) When a similar article is extracted based on the article that the number of nouns is little, the possibility to be deleted is high. The fault that the possibility to be deleted lowered when a similar article was extracted oppositely based on the article that the number of nouns is a lot of was seen.

TABLE V
RESULTS OF DELETED AND EXTRACTED ARTICLES

	extracted	deleted
similar articles	22	5
non similar articles	0	17

B. Experiment 2

In this section, we described, the experiment on the number of strokes setting of the top sentence about algorithm 2, the experiment on the evaluation threshold setting, and the comparative experiments by algorithm 1 and algorithm 2.

1) The experiment on the number of strokes setting of the top sentence about algorithm 2: 20 articles that putting the entire article together was not written from genres of the education, sports, and the science, etc. to the punctuation of a start the article opening in sentences were searched out from among Yahoo! JAPAN news. And, the number of characters of sentences to the punctuation of an opening start of those articles was shown in the graph. (See Fig. 5) The alphabet shows each article. It can be judged that putting the entire article together is not written from this data as for 80 percent at less than 25 characters. Therefore, the sentences are not extracted when the number of strokes to the punctuation of the sentence is at the beginning fewer than that of 25 characters, and sentences to the following punctuation are extracted in algorithm 2.

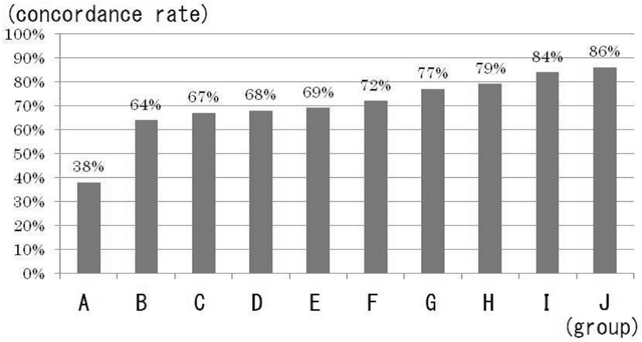


Fig. 5. words of first sentence

2) The experiment on the evaluation threshold setting: It experimented the evaluation threshold setting as well as algorithm 1. (See Fig. 6) The article on the object used the same one. It can be judged that 90 % is similar from the graph when the agreement rate of the noun is 60% or more. Therefore, it is judged that the article on X and Y is similar in algorithm 2 when Score(X, Y) is 60% or more.

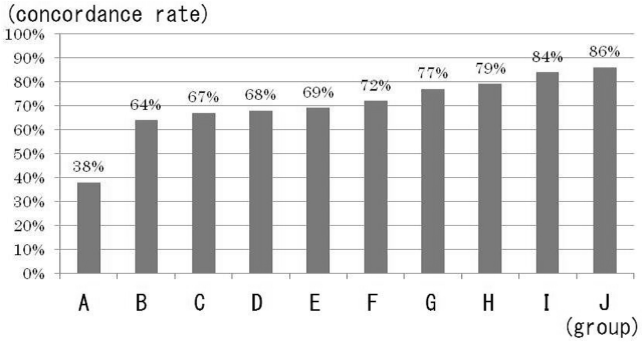


Fig. 6. occurrence rate of nouns using Score(X,Y)

3) URL extraction: It experimented on topics of the damage of crops by chemicals hepatitis C prosecution in the experiment for the article on the clocking as well as experiment 1 simultaneously now. Time had passed since time when it experimented 1, and the content of topics had been updated. It experimented on algorithm 1 again to compare the outcome of an experiments of algorithm 1 and algorithm 2. 45 URL has been both extracted as for algorithm 1 and algorithm 2 as a result of the URL extraction.

4) Delete low relativity articles and extract high reratibity articles: A lot of nouns extracted as well as experiment 1 to which URL was extracted became ”Damage of crops by chemicals”. In algorithm 1, extracted articles were 25 pieces in 45 pieces. (See Table VI) In algorithm 2, extracted articles were 22 pieces in 45 pieces. (See Table VII) In ths paper, a correct detection rate is defined the ratio of the total of number of extracted high relativity article URLs and deleted low relativity article URLs among the numbers of URL extracted at the start. As a result, a correct detection rate of algorithm2 became 67%, which is better than algorithm1’s (64 %). Therefore, algorithm 2 can be judged that detection with high accuracy is more possible than algorithm 1 in a relativity judgement. (See Table VIII)

TABLE VI
RESULTS OF EXTRACTED AND DELETED ARTICLES (ALGORITH1)

	extracted	deleted
high related articles	21	12
low related articles	4	8

TABLE VII
RESULTS OF EXTRACTED AND DELETED ARTICLES (ALGORITH2)

	extracted	deleted
high related articles	20	13
low related articles	2	10

TABLE VIII
COMPARISON OF ALGORITHM1 AND ALGORITH2

	correct extracted	miss extracted
high related articles	64 %	36 %
low related articles	67 %	33 %

5) Delete similar articles and extract articles: The extracted article became 20 in 25 pieces in algorithm 1. The article deleted by an article not similar did not exist. (See Table IX) The extracted article became 17 in 22 pieces in algorithm 2. The article that had been extracted in an article deleted by an article not similar and a similar article did not exist. (See Table X) As a result, it became 100% in algorithm 2 while the ratio of the total of the number of articles that were able to be deleted by a number of articles that were able to be extracted in an article not similar and similar article had become 96% in algorithm 1 among the numbers of articles that became positive detection that were judged that relativity was high and extracted. Therefore, algorithm 2 can be judged that detection with high accuracy is more possible than algorithm 1 in the similar article judgment. (See Table XI)

C. Experiment 3

The experiment by algorithm 2 with a high positive detection rate is described in both similar to a relativity judgment

TABLE IX
Results of deleted and extracted articles (Algorith1)

	extracted	deleted
similar articles	19	0
non similar articles	1	5

TABLE X
Results of deleted and extracted articles (Algorith2)

	extracted	deleted
similar articles	17	0
non similar articles	0	5

in experiment 2 compared with algorithm 1 recently. In topics of the damage of crops by chemicals hepatitis C prosecution targeted by experiment 1 and experiment 2, the number of articles was comparatively little. It experiments for the Aegis destroyer collision that is the thing news when there are a lot of numbers of articles because it measures the utility of the system in the experiment now. Moreover, the execution time of this system is measured while experimenting.

1) URL extraction: It experimented by using algorithm 2 for the article on topics of the Aegis destroyer collision in the experiment now. 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.

2) Delete low relativity articles and extract high reratibity articles: A lot of extracted nouns became ”Collision”. Articles that were judged that relativity was high, and extracted were 221 pieces in 316 pieces. (See Table XII)

3) Delete similar articles and extract articles: The article that was judged that there was no similarity as a result of processing it for 221 articles that were judged that relativity was high, and extracted, and extracted became 114. (See Table XIII)

4) Result: 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 XIV) 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 XI
Comparison of Algorithm1 and Algorith2

	correct extracted	miss extracted
high related articles	96 %	4 %
low related articles	100 %	0 %

TABLE XII
Results of extracted and deleted articles “aegis destroyer collision”

	extracted	deleted
high related articles	212	58
low related articles	9	37

TABLE XIII
Results of deleted and extracted articles “aegis destroyer collision”

	extracted	deleted
similar articles	103	36
non similar articles	11	71

TABLE XIV
Extracted result of “aegis destroyer collision”

	correct extracted	miss extracted
high related articles	79 %	21 %
low related articles	79 %	21 %

V. 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