Extraction and Visualization of Trend Information from Newspaper Articles and Blogs

Hidetsugu Nanba
Hiroshima City University
3-4-1 Ozukahigashi,
Hiroshima 731-3194 JAPAN
Phone: +81-82-830-1584
nanba@hiroshima-cu.ac.jp

Nao Okuda
Hiroshima Elpida Memory, Inc.
7-10 Yoshikawa Kogyo Danchi,
Higashi Hiroshima 739-0198
JAPAN

Manabu Okumura
Tokyo Institute of Technology
4259 Nagatsuta, Yokohama
226-8503 JAPAN
Phone: +81-45-924-5295
oku@pi.titech.ac.jp

Abstract

Trend information is a summarization of temporal statistical data, such as changes in product prices and sales. We propose a method for extracting trend information from multiple newspaper articles and blogs, and visualizing the information as graphs. As target texts for extraction of trend information, the MuST (Multimodal Summarization for Trend Information) workshop focuses on newspaper articles. In addition to newspapers, we focus on blogs, because useful information for analysing trend information is often written in blogs, such as the reasons for increases/decreases of statistics and the impact of increases/decreases of statistics on society. To extract trend information, we extract temporal expressions and statistical values, and we devised methods for both operations. To investigate the effectiveness of our methods, we conducted some experiments. We obtained a recall of 6.3% and precision of 31.3% for newspaper articles, and a recall of 44.8% and precision of 60.3% for blogs. From the error analysis, we found that most errors in newspaper articles were caused by misconversion of temporal expressions such as "同年" (the same year) or "前月" (the previous month), into "YYYY-MM-DD" form, although temporal expressions were detected correctly. In contrast to newspaper articles, there are few temporal expressions in blogs for which resolution is required, such as "同日" (the same day) or "前月" (the previous month). As a result, recall and precision for blogs are higher than those for newspaper articles.

Keywords: *Trend Information, Visualization, Newspaper article, Blog.*

1 Introduction

Trend information is a kind of summarization of temporal statistical data, such as changes in product prices and sales. We propose a method for extracting trend information from newspaper articles and blogs, and visualizing it as graphs.

Analysis of trend information has been studied for a long time in various fields, such as stock price prediction and temperature forecasting. A traditional approach is to apply time-series analysis to predict future values of the time-series variable [Doornik and Hendry 2001]. Recently, Gruhl proposed a method to predict the sales rank of a book in Amazon[1] [Gruhl et al. 2005] based on the number of blogs that mention the book. Although these approaches are useful for predicting statistics, they cannot contribute to a deeper understanding of the statistics.

We focus not only on statistical data but also on the information that enables us to understand the statistical data. Therefore, we have studied extraction and visualization of trend information from texts.

As target texts for extraction of trend information, the MuST (Multimodal Summarization for Trend Information) workshop [Kato et al. 2005, 2007] focuses on newspaper articles, and several studies have done in this field [Murata et al. 2006, Watanabe and Kobayashi 2006]. In addition to newspapers, we focus on blogs, because useful information for analysing trends is often written in blogs, such as the reasons for increases/decreases of statistics and the impact of increases/decreases of statistics on society.

The remainder of this paper is organized as follows. Section 2 describes trend information in newspaper articles and blogs. Section 3 explains how to extract trend information from them. Section 4 overviews our system. To investigate the effectiveness of our method, we conducted some examinations and Section 5 reports on these. Section 6 shows system behaviour. We present some conclusions in Section 7.

2 Description of Trend Information in Newspaper Articles and Blogs

Descriptions in newspaper articles and blogs can provide a deeper understanding of trend information. Figures 1 and 2 show sentences extracted from a newspaper article and a blog entry, respectively. Both examples describe trend information. Statistical values and descriptions of the values are shown as underlines and wavy lines, respectively.

In Figure 1, the author of the article described why the yen fell. In Figure 2, the author of the blog entry wrote on the effects of the increase of gasoline prices. The former example is an objective fact, while the

[1] http://www.amazon.com

latter is subjective (or supposition). Thus, descriptions about trend information in newspapers and blogs are different.

八日の東京外国為替市場は、日本経済の先行き不透明感やアジアの経済危機の悪化を懸念した円売り・ドル買いの動きが一段と強まり、円は一時、一ドル＝一四〇円七三銭まで下落した。
(The yen fell to 140.73 yen against the U.S. dollar on June 3 as an increasing number of market players sold the yen and bought the dollar because of uncertainty surrounding the nation's economy and fears of an escalation of the Asian economic crisis.)

(June 8, 1998, *Yomiuri* newspaper article)

Figure 1 Example of a newspaper article that mentions trend information

日本ではレギュラーガソリンの平均価格が 136 円/1 リッターと過去最高水準を記録して話題になっていますが，アメリカ本国でも 1 ガロン 2. 88 ドル（約 84 円/1 リッター）と，これまた最高記録を更新しそうな勢い．2 年と 4 ヶ月でほぼ 2 倍になったアメリカのガソリン価格が，このハマーをはじめとした大排気量大型モデルの販売不振に直結しているのは間違いないみたいですね．
(The average price of regular gasoline reached 136 yen/litre in Japan, hovered at this record level, and got into the news in Japan. On the other hand, the average price in U.S. was 2.88 dollars per gallon (about 84 yen/litre), which was a near-record level. The price of gasoline has almost doubled in the last two years and four months, and there is almost no doubt that this directly affected the sales of high-emission vehicles, such as Hummer.)

Figure 2 Example of a blog entry that mentions trend information

Many blog entries containing trend information are based on some information source, such as newspapers or TV news, so it is not obvious that all blog entries contain original material. To investigate the originality of descriptions in blogs, we classified blog entries according to the degree of citation of newspaper articles. The result is shown in Table 1. As can be seen from Table 1, about 75% of blog entries contain original contents, and more than half of blog entries do not cite newspaper articles. In other words, these entries are fully written by the blog authors.

Table 1 Classification of blog entries

	Fraction (%)
Entries that cite newspaper articles in full, with no original content	25.8 (80/310)
Entries that partially cite newspaper articles	4.8 (15/310)
Entries that mention their information source as newspaper articles, without citing them	50.0 (155/310)
Entries with unspecified information sources	19.4 (60/310)

3 Extraction of Trend Information

The task of extracting trend information can be divided into two subtasks: (1) extraction of temporal expressions and (2) extraction of statistical values from texts. We describe these subtasks in Sections 3.1 and 3.2, respectively.

3.1 Extraction of Temporal Expressions

We use CaboCha[2] [Kudo and Matsumoto 2003] for extracting temporal expressions from texts. CaboCha is a statistical syntactic parser for Japanese texts, and also identifies eight kinds of named entity, such as date, organization, and location, in texts. Here, not all temporal expressions in texts are used to generate a graph. For example, there are three temporal expressions in the sentence in Figure 3, but "1994" is the only expression that is used in generating a graph of "the number of births for one year".

一九九四年の年間出生数は前年より四万七千人も多い百二十三万五千人を記録し，二十一年ぶりに大幅増に転じた．
(The number of births recorded in 1994 was 1,235,000, which was 47,000 greater than in the previous year, and this significant increase was the first in 21 years.)

Figure 3 Example of a sentence that contains more than one temporal expression

To eliminate unrelated temporal expressions from texts, we focus on some cue phrases, some of which are shown in Table 2. In the table, cue phrases are underlined. If these cue phrases appear before or after temporal expressions, they are eliminated.

Table 2 Examples of cue phrases to eliminate unrelated temporal expressions

前年より (in comparison with the previous year)
10 年ぶり (for the first time in 10 years)
昨年以来 (since last year)
3 年連続で (for the third year in a row)

[2] http://chasen.org/~taku/software/cabocha/

In the next stage, the extracted temporal expressions in the prior stage are converted into a specific format "<DATE>YYYY-MM-DD</DATE>", where YYYY, MM, and DD indicate year, month, and day, respectively. For example, if "昨日" (yesterday) is extracted from a text written on January 25, 2007, then the expression "yesterday" is replaced by "<DATE>2007-01-24</DATE>". If the exact date is not provided, such as "昨年 12 月" (last December), this expression is converted as "<DATE>2006-12-??</DATE>".

3.2 Extraction of Statistical Values

We extract statistical values in the following four steps:

1. Split a sentence into separate statistical values;
2. Annotate "NUM" and "UNIT" tags;
3. Eliminate unrelated statistical values;
4. Extract statistical values.

(Step 1) Split a sentence into separate statistical values

Some sentences contain more than one statistical value, for example:

今日のレギュラーガソリンは１３０円，ハイオクは１５０円だった．
(Today's price of regular gasoline was 130 yen, and that of premium gasoline was 150 yen.)

From this sentence, "150 yen" might be extracted mistakenly as the price of regular gasoline, or "130 yen" as that of premium gasoline. We therefore split sentences for the statistical values. First, we analyse the dependency structure of the sentence using CaboCha. Second, we integrate two *bunsetsus* (segments) that have a modification relation and are adjacent to each other. We explain this process using the example shown in Figure 4. This figure is the result from CaboCha for the sentence given above. "Chunk" tags with ID numbers are assigned to each *bunsetsu*. Here, attribute values in each chunk tag indicate ID numbers of *bunsetsus* that have a modification relation. *Bunsetsus* 0 and 1 have a modification relation, and are adjacent to each other. Therefore, these *bunsetsus* are integrated. By conducting the process repeatedly to the end of the sentence, it is split into two parts: (1) "Today's price of regular gasoline was 130 yen," and (2) "and that of premium gasoline was 150 yen".

(Step 2) Annotate "NUM" and "UNIT" tags

We add "NUM" and "UNIT" tags to all candidate statistical values using the pattern "number + (noun phrase | postfix | counter suffix)". For example, NUM and UNIT tags are annotated for an expression "150 円" (yen) as "<NUM>150</NUM><UNIT> 円 </UNIT>".

<chunk id="0" link="1">今日の(Today's)</chunk>
<chunk id="1" link="2">レギュラーガソリンは (price of regular gasoline was)</chunk>
<chunk id="2" link="4">130 円，(130 yen,)</chunk>
<chunk id="3" link="4">ハイオクは(and that of premium gasoline was)</chunk>
<chunk id="4" link="–1">150 円だった．(150 yen.)</chunk>

<chunk>今日のレギュラーガソリンは 130 円, (Today's regular gasoline was 130 yen,)</chunk>
<chunk>ハイオクは 150 円だった． (and premium gasoline was 150 yen.)</chunk>

Figure 4 Splitting a sentence for statistical values

(Step 3) Eliminate unrelated statistical values

Not all expressions assigned "NUM" and "UNIT" tags are used to generate a graph of trend information. For example, there are two statistical values in the sentence in Figure 5, but "９２円" (92 yen) is the only expression that is used to generate a graph.

石油情報センターが２３日発表した給油所石油製品市況調査によると，６月のガソリン価格は全国平均でレギュラー１リットル当たり<NUM>９２</NUM><UNIT>円</UNIT>となり，前月比で<NUM>２</NUM><UNIT>円</UNIT>上昇した．
(According to the Oil Information Centre's survey of market conditions for products sold through service stations on (June) 23, the price of gasoline reached a national average of <NUM>92</NUM><UNIT>yen</UNIT> per litre, regular; <NUM>2</NUM><UNIT>yen</UNIT> higher than the average price of last month)
(June 24, 1999, *Mainichi* newspaper article)

Figure 5 Example of an analysis of a newspaper article

In the same way as extraction of temporal expression, we focus on some cue phrases to eliminate unrelated statistical values from texts, some of which are shown in Table 3. Here, cue phrases are shown as underlines. If they appear before or after statistical values, they are eliminated.

Table 3 Example of cue phrases to eliminate unrelated statistical values

2 円高い (2 yen higher)
昨年比 45.2%上昇 (up 45.2% from the previous year)
0.28 パーセント下落 (down 0.28 percent)

(Step 4) Extract statistical values

All statistical values are extracted when a predefined keyword (the name of the statistic) and unit appear in the same integrated *bunsetsu*. For example, "130 yen" is extracted from the sentence in Figure 3,

when “レギュラー” (regular) and “円” (yen) were given to the system in advance.

4 System Configuration

The system configuration is shown in Figure 6.

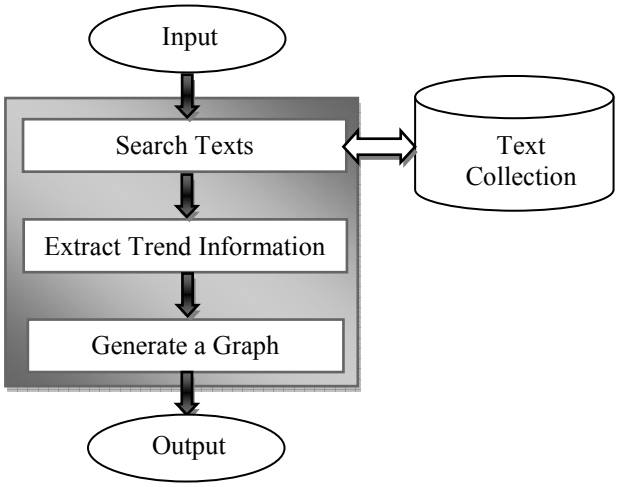

Figure 6 System configuration

Input
The system input is one or more keywords, which indicate a topic (e.g. “cabinet approval rate”) and a unit (e.g. “dollar”, “%”).

Search Texts
The system searches texts using keywords from a text collection.

Extract Trend Information
The system extracts temporal expressions and statistical values from texts searched in the prior stage.

Generate a Graph
The system generates a graph using a Perl module, “GD Graph”[3].

Output
The system outputs the graph and the documents.

5 Experiments

To investigate the effectiveness of the method given in Section 3, we conducted some examinations.

5.1 Data

We used the MuST data [Kato et al. 2007] and blog data in the examinations. The MuST data consists of newspaper articles about 27 topics written in Japanese, and were extracted from the *Mainichi* newspaper database for the years 1998 and 1999. Among these 27 topics, we manually selected eight that were also mentioned in blogs. The blog data was collected using blogWatcher[4] [Nanno 2004], a Japanese blog search engine. All the blog entries were written in 2006. Details of the data are shown in Table 4.

[3] http://search.cpan.org/dist/GDGraph/
[4] http://blogwatcher.pi.titech.ac.jp/

Table 4 Topics, newspaper articles and blog entries used in the experiments

Topics	Number of News Articles (MuST data)	Number of Blogs
Shipment volume of beer	22	68
Vehicle shipments	16	265
Sales of communication devices	26	225
Gasoline prices	20	578
Attendance at movies	6	192
Nikkei stock average	37	195
Domestic shipment volume of personal computers	20	57
Cabinet approval rate	17	108

5.2 Experimental Method

We evaluate our method using the following equations.

$$\text{Recall} = \frac{\text{The number of entries that the system extracted correctly}}{\text{The number of entries that should be extracted}}$$

$$\text{Precision} = \frac{\text{The number of entries that the system extracted correctly}}{\text{The number of entries that the system extracted}}$$

5.3 Results

The experimental results for newspaper articles and for blogs are shown in Tables 5 and 6, respectively. We calculated recall and precision for each of the following three cases: (1) temporal expressions extraction, (2) statistical values extraction, and (3) pairs of temporal expressions and statistical values extraction.

Table 5 Results of trend information extraction from newspaper articles

	Recall (%)	Precision (%)
Time	6.3 (31/491)	31.3 (31/99)
Statistics	19.6 (96/491)	97.0 (96/99)
Time and Statistics	6.3 (31/491)	31.3 (31/99)

Table 6 Results of trend information extraction from blogs

	Recall (%)	Precision (%)
Time	55.5 (239/431)	74.7 (239/320)
Statistics	59.9 (258/431)	80.6 (258/320)
Time and Statistics	44.8 (193/431)	60.3 (193/320)

5.4 Discussion

As can be seen from Tables 5 and 6, extraction from newspaper articles is worse than from blogs. In particular, both recall and precision of temporal expression extraction from newspapers are very low. From the error analysis, we found that most errors were caused by misconversion of temporal expression into "YYYY-MM-DD" form, although temporal expressions were detected correctly in most cases. Figure 7 shows a typical example of such errors.

総市場のシェアは，キリンビールが３８・５％で首位を守り，前月はキリンと１・３ポイント差に迫ったアサヒは３４・７％.
(*Kirin*'s share in the total market is 38.5%, and the company maintained the top ranking, while *Asahi*'s share is 34.7%, although *Asahi* got within 1.3% of *Kirin* in the previous month.)

(March 12, 1999, *Mainichi* newspaper article)

⇩

総市場のシェアは，キリンビールが
<NUM>38.5</NUM><UNIT>%</UNIT>で首位を守り，
<DATE>1999-2-??</DATE>はキリンと
<NUM>1.3</NUM><UNIT>ポイント</UNIT>差に迫ったアサヒは<NUM>34.7</NUM>
<UNIT>%</UNIT>.
(*Kirin*'s share in the total market is <NUM>38.5</NUM> <UNIT>%</UNIT>, and the company maintained the top ranking, while *Asahi*'s share is <NUM>34.7</NUM> <UNIT>%</UNIT>, although *Asahi* got within <NUM>1.3</NUM><UNIT>%</UNIT> of *Kirin* in <DATE>2006-02-??</DATE>.)

Figure 7 Example of error

In this example, "前月" (the previous month) is mistakenly converted into "1999-2-??", because the system inferred that "the previous month" was the month previous to March, when this article was written. However, "the previous month" indicates January 1999, because this article describes the shares of beer sales in February.

In contrast to newspaper articles, there are few temporal expressions in blogs for which resolution is required, such as "同日" (the same day) or "前月" (the previous month). As a result, recall and precision of blogs are higher than those of newspaper articles.

6 System Behaviour

Figure 8 shows a graph of "*Asahi*'s share" in the national beer market. We used "アサヒ" (*Asahi*) and "%" as a keyword and a unit, respectively. Four statistical values, which were extracted from *Mainichi* newspaper articles in 1998 and 1999, are shown. The *x*-axis and *y*-axis indicate "year" and "share in the national beer market", respectively. The graph was generated automatically using GD Graph, a Perl module.

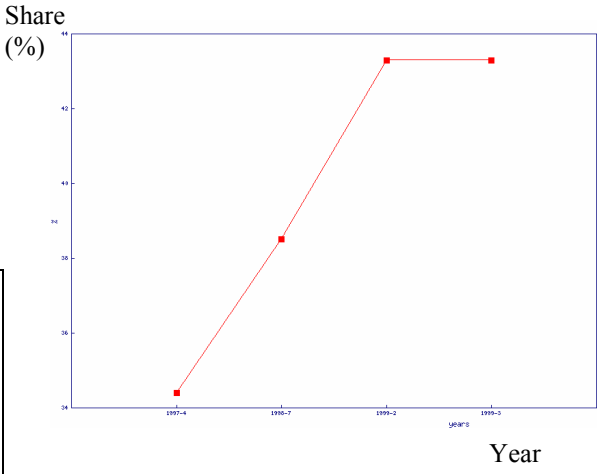

Figure 8 A graph of "*Asahi*'s share" in the national beer market

7 Conclusions

We have proposed a method for extracting statistical values and temporal expressions from newspaper articles and blogs, and visualizing them as graphs. In our experiments, we obtained a recall of 6.3% and a precision of 31.3% for newspaper articles, and a recall of 44.8% and a precision of 60.3% for blogs. From the error analysis, we found that most errors in newspaper articles were caused by misconversion of temporal expression, such as "同年" (the same year) or "前月" (the previous month), into "YYYY-MM-DD" form, although temporal expressions were detected correctly in most cases. In contrast to newspaper articles, there are few temporal expressions in blogs that require resolution, such as "同日" (the same day) or "前月" (the previous month). As a result, recall and precision of blogs are higher than those for newspaper articles.

8 Acknowledgements

The authors would like to express their gratitude to the organizers of the MuST for providing the data set. This work was supported by Grant-in-Aid for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

J.A. Doornik and D.F. Hendry. Modelling Dynamic Systems Using Pcgive. *Timberlake Consultants Press*, 2001.

D. Gruhl, R. Kumar, J. Novak, and A. Tomkins. Predictive Power of Online Chatter. In *Proceedings of ACM Special Interest Group on Knowledge Discovery in Data* (*SIGKDD*), pp.78–87, 2005.

T. Kato, M. Matsushita, and N. Kando. MuST: A Workshop on Multimodal Summarization for

Trend Information. In *Proceedings of the 5th NTCIR Workshop Meeting*, pp.556-563, 2005.

T. Kato, M. Matsushita, and N. Kando. Expanding of Multimodal Summarization for Trend Information – Report on the First and Second Cycles of the MuST Workshop. In *Proceedings of the 6th NTCIR Workshop Meeting*, 2007.

T. Kudo and Y. Matsumoto. Fast Methods for Kernel-based Text Analysis. In *Proceedings of the 41th Annual Meeting on Association for Computational Linguistics*, pp.24-31, 2003.

M. Murata, K. Ichii, Q. Ma, T. Shirado, T. Kanamaru, S. Tsukawaki, and H. Isahara. Development of an Automatic Trend Exploration System using the MuST Data Collection. In *Proceedings of the Workshop on Information Extraction Beyond the Document, COLING-ACL Workshop*, pp.1-11, 2006.

T. Nanno, T., Fujiki, Y. Suzuki, and M. Okumura. Automatic Collection, Monitoring, and Mining of Japanese Weblogs. In *Proceedings of the 13th International World Wide Web Conference*, 2004.

C. Watanabe and I. Kobayashi. Intelligent Information Presentation Corresponding to User's Request based on Collaboration between Text and 2D Charts. In *Proceedings of International Symposium on Computational Intelligence and Industrial Applications 2006*, 2006.