AN EXTRACTION METHOD OF TIME-SERIES NUMERICAL DATA FROM ENTERPRISE PRESS RELEASES

Masanori Akiyoshi, Mayu Gen, Masaki Samejima, Norihisa Komoda
Osaka University
Yamadaoka 2-1, Suita, Osaka 565-0871, Japan

Keywords: Time-series numerical data, enterprise press releases, automatic extraction.

Abstract: This paper addresses an extraction method of time-series numerical data from enterprise press releases for business strategy design. Business strategy consists of logical actions for continuously producing enterprise outcome. The business strategy design process that is partially based on competitive environment analysis may extremely resort to professional skills so far. To enhance and accelerate the competitive environment analysis, we focus on press releases of competitors in order to extract numerical data related to products or services. Sentences in press releases are well organized and grammatically correct. Therefore such extraction is simply done by identifying the keywords of products or services and the unit indicator co-occurrence. In addition to such simple approach, we clarify the specific rules to applying our method to practical press releases.

1 INTRODUCTION

Recently business is considered to be under unpredictable environment, which means rapid changes of consumers' demand cause sudden drop of profit and weaker positioning in the target market. When present corporate management is in excellent state, future corporate management is no longer on the elongation of it. To judge investment adequately, it is indispensable to analyze various factors, for instance, not only financial aspects but also aspects of organizational growth, customer satisfaction, and so forth. On the other hand, along with the development of enterprise information systems, it becomes possible to judge business operations rapidly and adequately, for instance, sales planning by using OLAP (on line analytical processing) of data warehouse or data mining technique. Even if such business operations are succeeded, continuous growth of business is not attained without business strategy. Business strategy has become more critical to corporate management (Loebbecke, 2003).

Business strategy consists of logical actions for continuously producing enterprise outcome. Such design may extremely resort to professional skills so far, because the design process involves repetitive work of various analysis, goal setting, and investigation through simulation, which are inherently human-centric work. It takes much time to design business strategy only by hand of human professionals. Therefore there is a strong need to make it possible to design business strategy by newly emerging information technology, which is expected to support human professionals.

While there are several possible approaches to support human professionals, we focus on document analysis-based support for repetitive analysis. Business documents such as enterprise annual reports and press releases indicate applied results of business strategy. Market-related documents such as consumer opinion collected through antenna shops or a callcenter include potential needs for products or services. Therefore we think analysis on business documents is inevitable to design business strategy (Akiyoshi, 2005).

In this paper, as analysis on business documents, we focus on press releases of competitors in order to extract numerical data related to products and services, and so forth. We first describe some features of press releases, and then propose our extraction method in detail. Finally experimental results are also discussed.

Akiyoshi M., Gen M., Samejima M. and Komoda N. (2006).
AN EXTRACTION METHOD OF TIME-SERIES NUMERICAL DATA FROM ENTERPRISE PRESS RELEASES.
In *Proceedings of the First International Conference on Software and Data Technologies*, pages 221-225
DOI: 10.5220/0001318202210225

2 EXTRACTION OF NUMERICAL DATA FROM PRESS RELEASES

2.1 Approach Based on the Feature of Press Releases

As mentioned in the previous section, it is indispensable to investigate competitors' status such as the amount of shipment related to some specific products and settlement of accounts in the last quarter, and so forth. Such numerical data are appeared in press releases, currently via Web pages. Therefore we can use such Web pages as analysis data source. These documents have the following features.

- Sentences are well organized and grammatically correct.
- Similar expression and format is used in the same contents type of releases.
- Published date is explicitly included.

These features make it possible to find the target portions systematically to some extent, for instance, shipment data co-occurred with the same keywords is identified in the different press releases. Figure 1 shows an overview of the tool based on our proposed method.

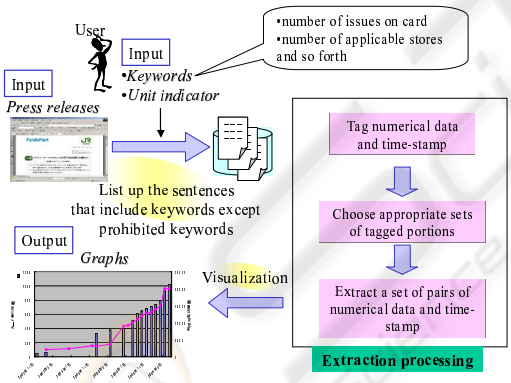

Figure 1: Overview of the extraction tool.

However the contents structure demands different methods to extract such numerical data from the identified target portions in documents, for instance, the portions are expressed in plain text or in listed items, and so forth. In addition, even if a portion includes the target keyword, the extracted numerical data might be irrelevant for the sake of the description for the other related keyword. Therefore extraction of numerical data has to be considered based on such variations of contents structure and description in documents. Also extracted pairs of time-stamp and numerical data are not always complete from graph charts points of views, because expression on time-stamp varies from documents, and numerical data at some time-point is missing. Against these problems, our approach is as follows.

- Identification of the target portions is executed by both target keywords and prohibited keywords, for instance, "card name" as the target keyword and "card name" indicating electronic money usage as the prohibited keyword.
- Extraction from the target portions is categorized into two cases such as plain text portions and listed item portions.
- Pre-process for the parenthesised portions is used.
- Post-process for the unified expression of time-stamp is sometimes used.
- Post-process for filling up the interval numerical data is sometimes used.

2.2 Basic Extraction Method

Extraction is processed sentence by sentence. The keywords and the unit indicator for numerical data given by a user are input to the tool. Basic extraction is executed when the keywords except prohibited ones and the unit indicator co-occurred in a sentence. We use tagging tool called NExT that automatically identifies specific nouns such as person name, organization name, location name, and numerical data (Fukumoto, 2002). In order to tag arbitrary numerical number expressions, it is necessary to register unit indicators in its dictionary. Therefore we add online registration functionality as pre-process using NExT. Figure 2 shows an example of the basic extraction processing flow.

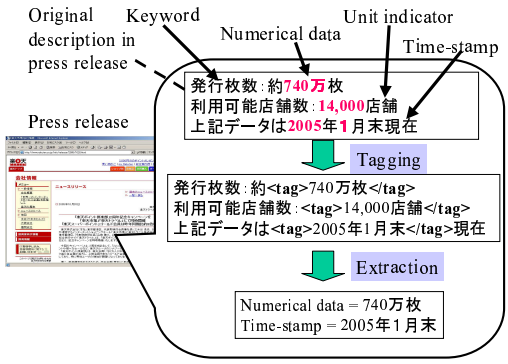

Figure 2: Basic extraction processing flow.

The basic extraction is effective to the sentences that include only one set of keywords and numerical data. However, most of sentences include plural

numerical data or optional expressions. Here we introduce processing rules in addition to the basic extraction.

2.3 Extraction by Words Distance

Figure 3 shows an example of plural set of keywords and numerical data. In this sentence, basic extraction processing makes two pairs as indicated in Figure . To distinguish the correct pair from the incorrect, we introduce words distance rule. As mentioned before, one of the characteristic features appeared in press releases is ”Similar expression and format is used in the same contents type of releases”. Therefore the same criteria of words distance such as ”nearest” is applicable to target sentences. In this example, the first pair indicating time-stamp as ”1l (the first quarter)” and numerical data as ”1,983,739 (1,983,739 thousand yen)” is extracted as correct one.

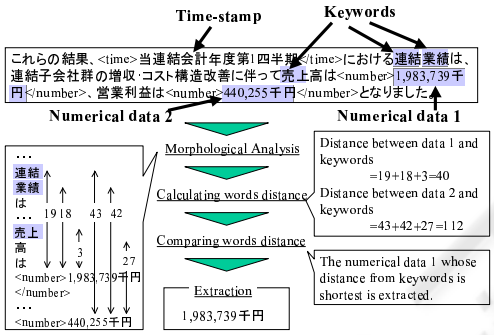

Figure 3: Example of extraction by words distance.

2.4 Extraction by Pre-process on Parentheses

Most of additional explanation expressions are appeared in parenthesis-style format such as the use of round brackets. In such cases, parenthesis portions play several roles in the sentence. Based on preliminary analysis, we categorize these cases into three types as follows.

Case A: The parenthesis includes the target keywords, time-stamp, and numerical data.

Case B: The parenthesis includes time-stamp, and numerical data except the target keywords.

Case C: Neither case A nor case B occurs.

In the case A, the extraction using tagging is executed on the parenthesis, and extraction is done in advance. Then, removing the parenthesis and its contents from the sentence is processed for further extraction.

In the case B, the parenthesis and its contents are just removed from the sentence for further extraction.

In the case C, only the pair of parenthesis is removed from the sentence.

The above procedures are repeated until all parentheses are removed. Figure 4 illustrates an example of this procedure.

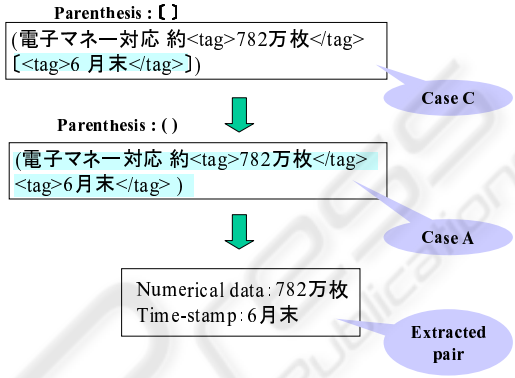

Figure 4: Example of extraction from parenthesis portions.

2.5 Unified Expression of Time-stamp as Post-process

As expression on time-stamp, concrete description is not always used in target documents, for instance, ”the first quarter of this year” or ”previous month” are sometimes used. If such relative description is intermingled with concrete one, unified expression is necessary to generate a set of pairs of time-stamp and numerical data for generating graph charts.

To unify this expression, one concrete description is hooked first. Then, based on relative description patterns, the relative description is substituted by using the hooked concrete time-stamp.

2.6 Filling Up the Interval Data as Post-process

Suppose that extracted data have the year-based numerical data and the first half year-based one as shipping products. Then the second half year-based numerical data is easily calculated from these two data. In case of generating half year-based graph charts, this filling up procedure is executed as post-process of extraction.

3 EXPERIMENTAL RESULTS

Here we discuss experimental results when applying our method to practical press releases. Three sets of enterprise press releases are used. Table 1 shows such experimental conditions; each English meaning is in round brackets.

Table 1: Experimental conditions.

	Number of documents	Keywords	Unit indicator
Company-A	100	Edy,J[h(card), s(number of issues)	
Company-B	48	Suica,J[h(card), s(number of sheets of issues)	
Company-C	43	A(consolidates), (sales results),(sales)	(yen))

Figure 5, Figure 6, and Figure 7 show the result graphs indicating the retained problem. As shown in some graphs, some numerical data are not extracted, and wrong extraction also occurs.

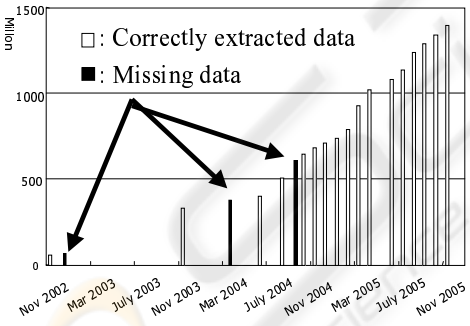

Figure 5: Extracted results on company-A.

As for the missing extraction, it is caused by slightly different keywords such as ”s (number of issues)” instead of ”s (number of sheets of issues)”. In our method, a keyword is matched exactly, which relies on the characteristics of press releases ”Sentences are well organized. Similar expression and format is used”. Therefore the extension to handle similar words of input keywords is necessary in more accuracy of extraction.

As for the wrong extraction, it is caused by irrelevant numerical data co-occurred with an input keyword in the sentence, for instance, a sentence of ”We commemorate the number of issues over 1000

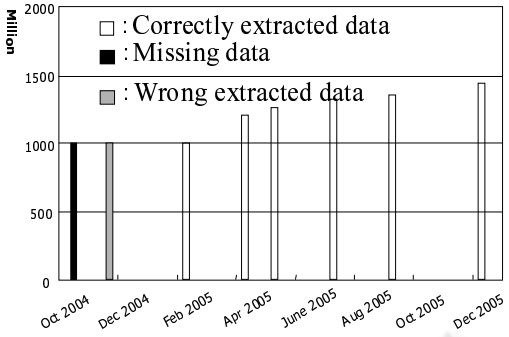

Figure 6: Extracted results on company-B.

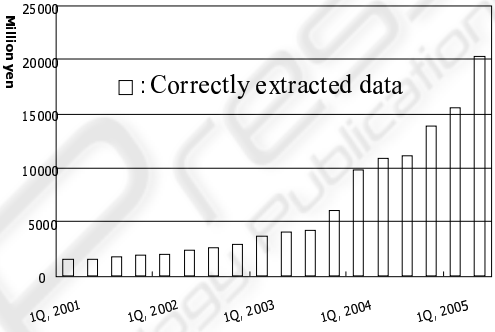

Figure 7: Extracted results on company-C.

as shopping card xxx, the next January 2006 c” includes ”January 2006” as time-stamp and ”1000” as numerical data. This is quite difficult without analysis of sentence meaning, so further study is necessary to handle it.

4 RELATED WORKS

Keywords extraction is frequently discussed in several types of documents such as newspapers, academic papers, Web pages, and so forth. Work in (Kobayashi, 2005) discusses automatic generation of keywords map related to technology. TF-IDF (term frequency and inverse documents frequency) automatically finds specific keywords that characterize documents. ThemeRiver visualize topic stream of online news by using a river flow metaphor (Harve, 2002). However, time-series numerical data extraction is not discussed enough so far.

Other works intended to extraction of key phrase or time-series data do not cover the set of ”keyword”, ”time-stamp”, and ”numerical data” (Zhang, 2005; Sripada, 2003). Of course our used tool NExT handles the extraction of numerical data, but it is general-use tool. Therefore we investigate what is significant

method to extract time-series numerical data for business analysis process points of views. In addition, we discuss post-processes for the sake of generating graph charts.

5 CONCLUSION

In this paper, we proposed an extraction method of time-series numerical data from enterprise press release to enhance and accelerate the competitive environment analysis. Our method simply uses the tagging techniques to extract appropriate portions of target documents. Through experimental results, time-series data are extracted appropriately, however, the extension on handling similar keywords is necessary for more accuracy.

REFERENCES

Loebbecke, C. et al. (2003) The impact of eBusiness and the information society on 'STRATEGY' and 'STRATEGIC Planning': An assessment of new concepts and challenges. J. of Information Technology and Management, vol.4, pp.165-182

Kobayashi, S. et al.(2005) Time series analysis of technology trends based on the Internet resources IEEJ trans. On EIS, vol.125, no.5, pp.720-729

Harve, S. et al.(2002) ThemRiver: Visualizing thematic changes in large document collections IEEE trans. On Visualization and Computer Graphics, vol.8, no.1, pp.9-20

Akiyoshi, M. et al.(2005) An analysis framework of enterprise documents for business strategy design In *Proc. of 5th Int. Conf. on Computational Intelligence for Modelling Control and Automation (CIMCA'2005)*, vol.2, pp.330-335

Fukumoto, J. et al.(2002) A named entity extraction tool (NExT) for text processing (in Japanese) In the *Proc. of 8th Annual Meeting of The Association for Natural Language Processing*, pp.176-179

Zhang, Y. et al.(2005) Narrative text classification for automatic key phrase extraction in web document corpora In *Proc. of the 7th annual ACM international workshop on Web information and data management*, pp.51-58

Sripada, S. G. et al.(2003) Generating English summaries of time series data using the Gricean maxims In *Proc. of the 9th ACM SIGKDD international conference on knowledge discovery and data mining*, pp.187-196