A Study on a Method for Supporting Scenario Extraction from Time Series Information

Kazuhisa INABA, Yukio OHSAWA
Graduate School of Business Science, Univ. of Tsukuba
3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
e-mail:{kazuhisa, osawa}@gssm.otsuka.tsukuba.ac.jp

Abstract

In KeyGraph analysis [1], firstly it computes the frequency of each words and the co-occurrence between the words in data. Secondly it analyzes the relative probability between the frequency and the co-occurrence of the words. Finally it chooses keywords from words and displays the relationship among the keywords. In the recent researches, they apply KeyGraph analysis for the scenario extraction of the consensus building in the discussion groups.

But it is difficult to extract the scenario depending on time, because KeyGraph cannot show the time series changes.

In this research a new tool was developed, which overlays additional time series information on KeyGraph output. The algorithms and some results of the tool analysis will be discussed in this paper.

Keywords: KeyGraph, Text data, Time series, scenario extraction

1. Background

KeyGraph is a tool for visualizing the map of event relations in the environment. By visualizing a map where events from data chosen by user appear connected in a graph, one can see the overview of the world of user's concern. It extracts the keywords from words in text data and shows the relationships among the keywords. The outputs of KeyGraph enable to us obtain the scenario in discussion or free style text data. The applications of KeyGraph have several areas such as marketing, business consensus and evaluation of human resources. [2]

But it is difficult to approach the data from the view point of the time series and step-by-step, because the output of KeyGraph shows only static relationship among the words. The black/red and nodes/links on KeyGraph which require our attentions must be changed if the data varies depending on time. e.g., consumers' interests are not static in marketing area, or context in the article depends on logical development.

The additional information is required for us to understand the output of the KeyGraph and obtain the time series scenarios.

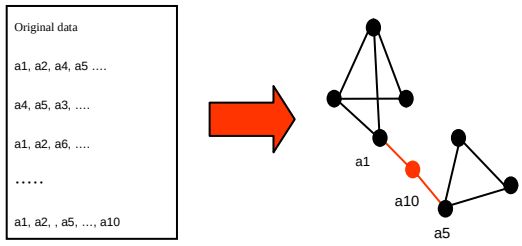

Fig. 1: KeyGraph Analysis.

2. Purpose

The new tool was developed in this research, which visualize the time based changes and enables us to create scenarios。

What this tool provides:

1. Overlaying additional interface on KeyGraph about time series information.

2. Enabling observer to recognize the dynamic varieties.

3. Displaying the strength of co-occurrence between words.

Some samples of the output of additional information are shown in this paper.

3. Algorithm

The points about the algorithms of the tools are:

1. How to spirit data into time series.
2. How to compose the steps to show time varieties.

①Spiriting the data into time series baskets.

In usual case, the chunks such as a line, a sentence and

one paragraph are treated as baskets in analysis. If they have an independency on time series, it will be defined as a basket. If some chunks have same meaning in time series, they are treated as a basket.

②Analyzing whole data and displaying ordinary output of KeyGraph

③Computing the co-occurrence among the words in each basket

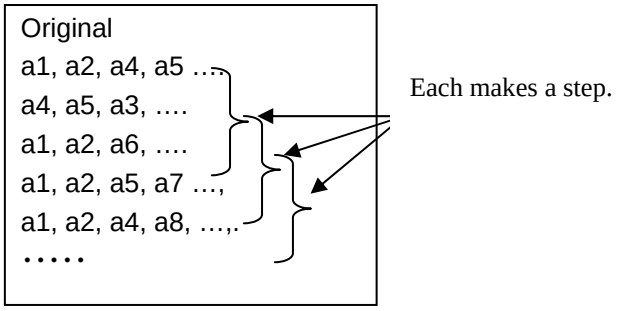

Fig. 2: How to create steps.

By defining the matrix of the co-occurrence as Concurrence（Wa, Wb, ID of the basket), the co-occurrence will be obtain as below

Concurrence（Wa, Wb ,ID of the basket）

=1(when co-occur),

=0(when no co-occur)

④Building steps by assembling baskets and calculating the cumulative co-occurrence.

The reason for assembling baskets is a kind of moving average. It makes the behaviors on the interface smooth.

In the case that there are m baskets (B1～Bm) and n(<m) baskets makes each step, M-N steps will be created as result.

Step1： B1～Bn

Step2： B2～Bn+1

Step3： B3～Bn+2

・・・・

Step m-n-1： Bm-n～Bm

Step m-n： Bm-n+1～Bm

e.g. in the case of m=100 and n=10, 90 steps will be created.

In each step, the cumulative co-occurrence will be calculated with co-occurrence matrix. Here the definition of cumulative co-occurrence matrix is:

Accumulation（Wa, Wb, i）

=SUM（Co-occurrence(Wa, Wb ,i)～

Co-occurrence(Wm,Wn,i +l))

⑤Creating additional Layer.

The layer that includes additional information is created with Accumulation matrix in ④. The eclipse links are introduced to show the information about each link and built corresponding with black and red link on KeyGraph. The density of the eclipse links' color depends on Accumulation matrix like elevation of topography. for improve the understanding, thresholds are used so that it display only the eclipse links that have more than a certain value.

⑥Overlaying additional layer on KeyGraph②

4. UserInterface

The user interface is shown Figure 3. By pushing the button, the time series varieties can be traced.

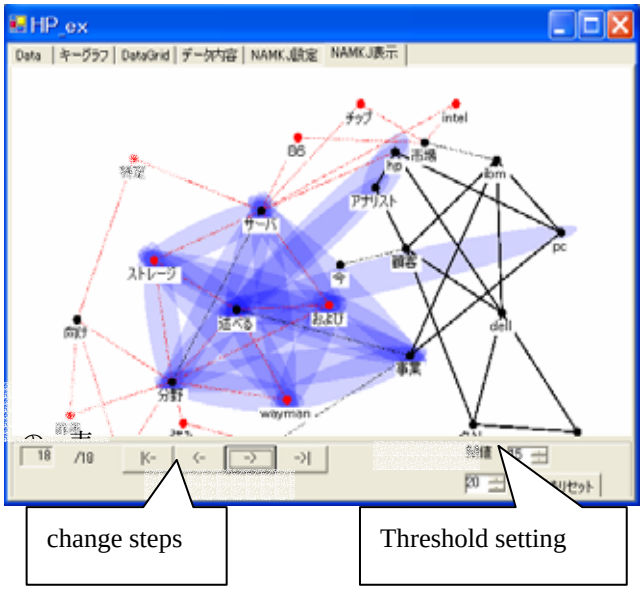

Fig. 3: The User Interface of Tool

5. Example of Analysis

An article on Internet web page was processed with this tool. The title of the article is

“Are sharks circling HP?” (CNET News.com)[3]

Though this article is originally written in English, the translated one into Japanese was used for Analysis.

Then several outputs are obtained from analysis. They show the effective of the tool.

This article has 20 sentences and it was set that each step consist of 3 continual baskets, which means there are 18 steps in this analysis.

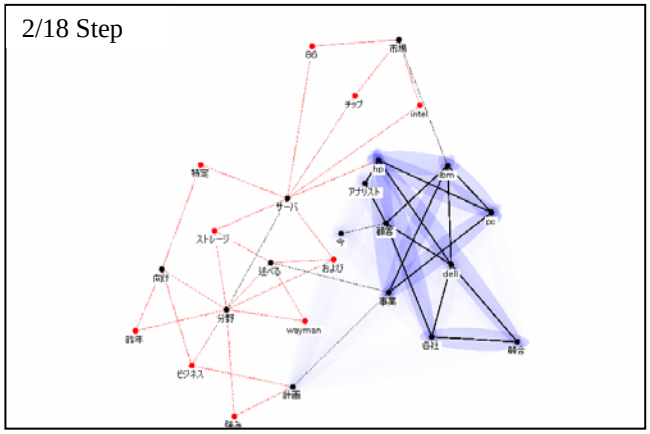

Fig. 4: Output (2/18)

Fig. 4 shows that PC venders such as dell, HP and IBM are now in competition. Recent news about IBM's selling out PC business reminds dell is coming to be strong in this area.

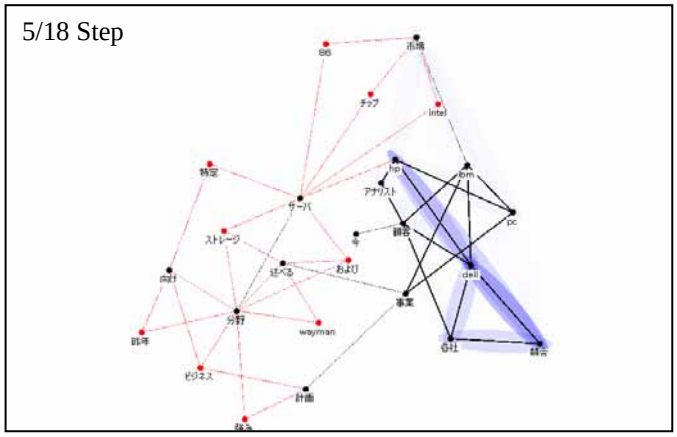

Fig. 5: Output (5/18)

Figure 5 shows that Dell and HP are competitors and they are focusing on PC market area.

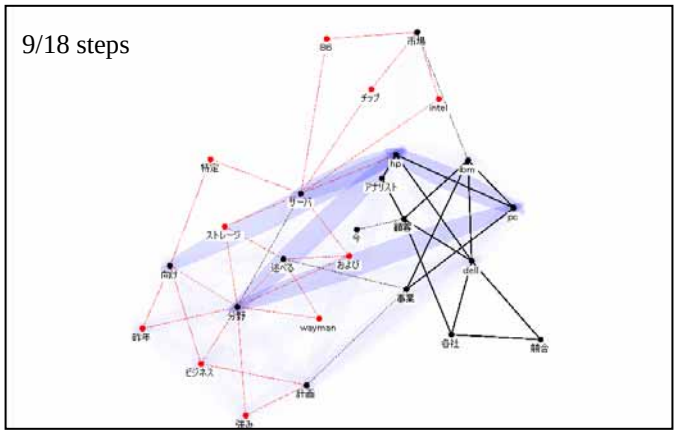

Fig. 6: Output (9/18)

Fig. 6 shows about HP's PC servers and its recent business result and plan. Although the advantage of HP is discussed in this step, it seems to be concerned with storage and business uses. .

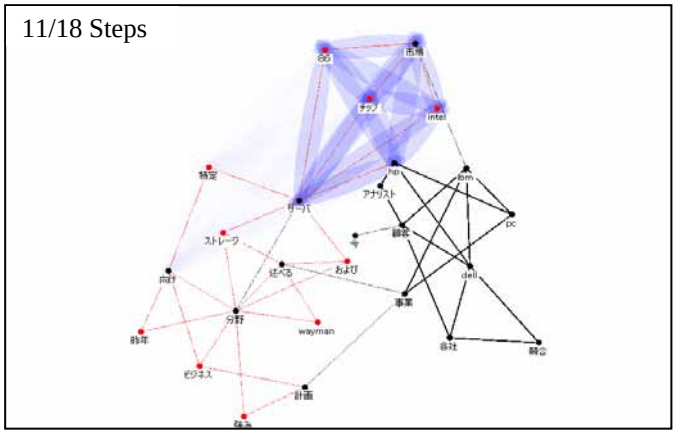

Fig. 7: Output (11/18)

Fig. 7 shows HP's strategy in PC server with special functionality. Intel's 86 series seems to be key product in this area.

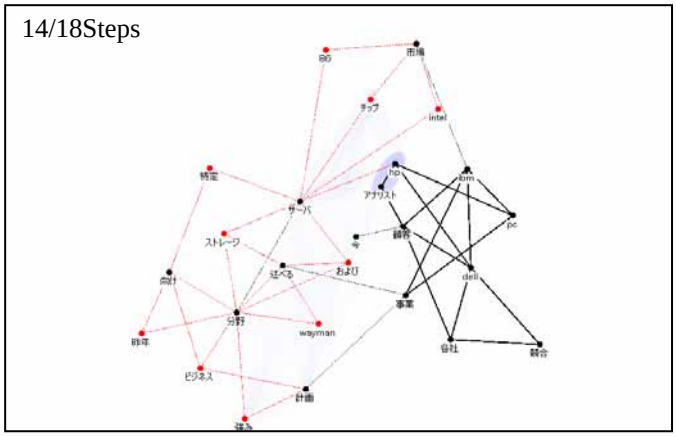

Fig. 8: Output (14/18)

Figure 8 shows that Analysts are discussing about CPU which makes HP's product stronger.

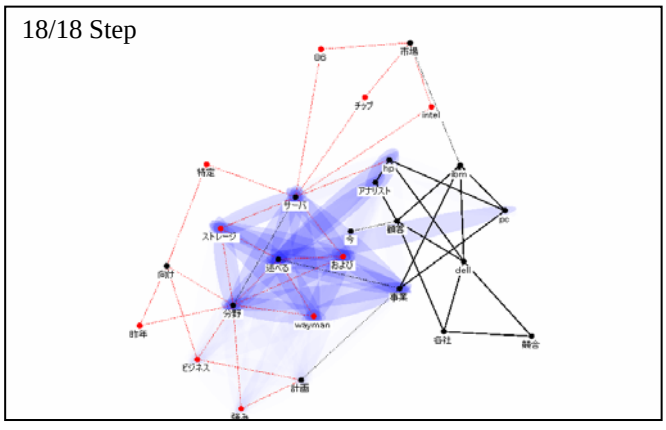

Fig. 9: Output (18/18)

Fig. 9 shows the conclusion of this article. Analysts and HP former CEO Wayman has commented focusing on PC, server and storage business of HP.

Japanese	English	Japanese	English
競争	Competition	市場	Market
各社	Each company	チップ	CPU chips
事業	Business	サーバ	Server
顧客	Customer	ストレージ	Storage
アナリスト	Analyst	特定	Focus
今	Current	および	And
分野	Area	述べる	Describe
強み	advantage	ビジネス	business
向け	Purpose for	昨年	Previous year
計画	Planning		

Table 1: Japanese-English matrix

The conclusion of article is different from the cluster of KeyGraph. In not only conclusion but also other paragraphs, it shows different part of KeyGraph. This means that the tools can show the important part of KeyGraph so that observer can reach the scenarios.

6. Conclusion

The developed tool in this research can deal with time series information on KeyGraph output and show the strength of the co-occurrence on each link that also changes depending on time.

The output of this tool makes it easier to extract scenarios from KeyGraph analysis, because recognizing which part of KeyGraph is important at the moment.

In the example of analysis about the article, it shows context varieties and conclusion on KeyGraph. The conclusion is different from part of cluster.

This implies that the developed tool can point out which red nodes/links are important for scenario extraction.

7. Future works

Firstly, more analysis examples and expansion of applied area are needed to verify the interface. The marketing, business consensus, pos data, chat in online community and etc. are considered as applications.

Besides I think about the improvement as below.

① Currently it deals with only text data. But adding attribute information to normal text data, it is available to approach from multiple view points. E.g. in questionnaires, some times men have different tendency from that of women. If we could grasp this difference it leads to new scenarios.

② In pre-analysis, baskets are created from original data for each time. But the time range should vary according to view points. There are different units in time e.g. year, month, season, week, day and so on. These time range changes could bring us diverse scenarios. In that sense, time range should be free to change.

③ The rough context of the article is obtained in this paper. It is difficult to focus on details about the part of KeyGraph. Selection of words and re-generating localized KeyGraph will make it available to focus on where we notice and recognize scenario deeply.

④ It is difficult to seize the differential information. is needed.

8. References

[1] Ohsawa Y, KeyGraph: Visualized Structure Among Event Clusters, in Ohsawa Y and McBurney P. eds, Chance Discovery, Springer Verlag: 262-275 (2003).

[2] Takamura M, A Study on Support Framework for Sales Business to Obtain Customer's Information for Products Development, The 4th Workshop on Scenario Emergence (2005)

[3] John G. Spooner (CNET News.com)「フィオリーナの CEO 辞任で気になる、HP とコンピュータ業界の行方」(Are sharks circling HP?) (2005/02/10) URL:http://japan.cnet.com/news/biz/story/0,2000050156,20080626,00.htm