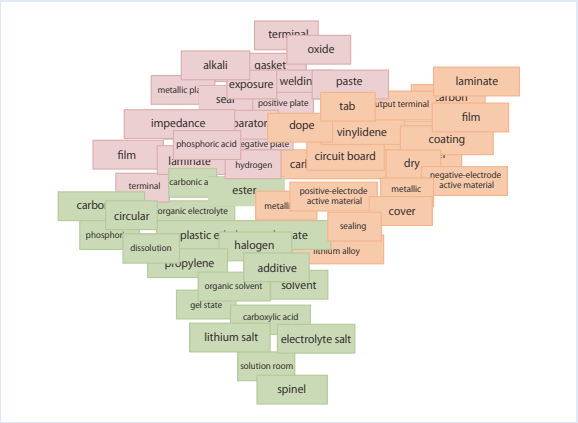
Japanese Patent Publication No. 2008-21052

Information Extracting Devise, Information Extracting Method, and Information Extracting Program

Invented by: ***MURATA Masaki***



Text mining results example
(Example of the patent related key word extraction)

Overview of the technology

The purpose of this invention is to automatically extract paired information relating to a certain topic from several groups of electronically recorded information and put such information in graph form. As shown in Fig. 1, this devise contains a key expression extracting component that extracts key expressions from a related article DB and a paired information extracting component that extracts multiple paired information from articles contained in the related article DB based on the key expressions extracted using the above key expression extracting component. Extracted information consists of multiple content expressions and corresponding target unit expressions (Fig. 2). For example, if the content expression is the Nikkei average share price or maximum temperatures, the corresponding unit expression would be the word “yen” in 9,100 yen or “degrees” in 35 degrees. The paired information extracting component extracts multiple paired information from articles contained in a group of articles based on the key expressions extracted by the key expression extracting component. For example, the paired information extracting component specifies the location where key expressions extracted from the key expression extracting component simultaneously appear in a group of articles stored in a related article DB, extracts the paired numeri-

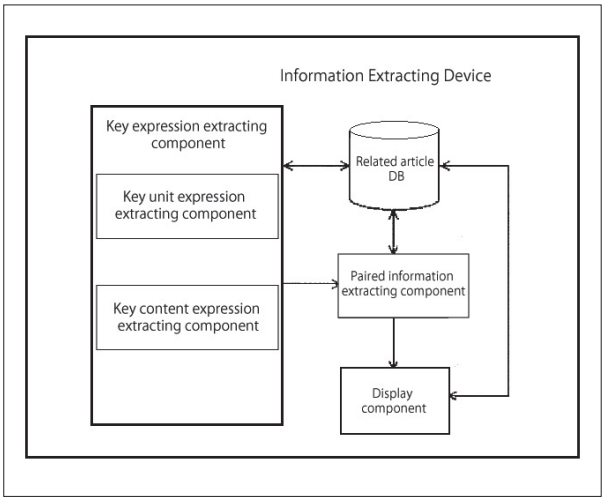


Fig. 1 System structure example

Movie data	Typhoon data	Beer data
Content expression		
Movie	Typhoon	Expected retail price
Box-office revenue	Maximum wind speed	Low-malt beer
Work product	Center area	Beer
Chihiro	Japan Meteorological Agency	Business information
Kamikakushi	Speed	Can
Unit Expression		
Yen	Number	Yen
Person	Meters	Milliliters
Dollars	Kilometers	%
Age	Hectopascals	Case
Number	Millimeters	Number

Fig. 2 Example of key expressions

cal information entered in such locations and generates paired information from the corresponding extracted target numerical information and content expressions of the above key expressions. In regard to the unit expressions contained in the key expressions, the paired information extracting component simultaneously extracts the numerical value relating to unit expressions (for example, the numerical value appearing adjacent to the unit expression in the article) and extracts the numerical value and unit expression as a numerical expression.

The display component organizes and displays paired numerical information extracted by the paired information extracting component. For example, the paired numerical information regarding "box-office revenue" "audience numbers" extracted from the paired information extracting component is displayed in a graph showing the "audience number" in the horizontal axis and "box-office revenue" in the vertical axis. When the key expression extracting component extracts multiple key expressions, the display component creates a graph of the key paired information selected from multiple types of paired information extracted by the paired information extracting component based on each expression, on the basis of evaluated values calculated using a specific evaluation value computation formula for each key expression. In addition, the paired information extracting component can also select key paired information from multiple types of paired information with the input specified by the user and, and the display component can also display the bubble chart where a displayed circle size corresponds to the numerical value.

Application

In companies, most information such as product and service survey results and complaints made to consumer advice centers is stored in electronic form. Consequently, it is necessary to ascertain the changes of content and trends of the huge amount of text data accumulated in data bases, reflect such information in strategies for future product sales and the provision of services and formulate policies to increase sales revenue. However, it takes much time to read each page of customer survey results. Text mining enables users to quickly extract only the necessary information from a large amount of text information. Although customer trends and satisfaction levels maybe easily ascertained if answers in surveys are made in multiple choice form, this is may not be the case for customer comments made at the end of such surveys. In order for machines to perform this function, it is essential that the content of the sentences be understood. However, this cannot be done using tools that can only search keywords. Consequently, this device classifies words according to their parts of speech in sentences contained in the related article DB and extracts, for instance in the case of time expressions, the nouns that follow numerical values such as "hours" and "minutes". A graph created by such results is shown in Fig. 3. This graph enables users to see the popularity of movies by box-office revenue represented by the vertical axis and the audience numbers represented by the horizontal axis. In the case, since the data of the movies at the commencement of the screening show naturally that both the box-office revenue and audience numbers are low, it is necessary to compare their data at the end of the screening. Since the entrance unit price appears to be roughly constant, the box-office revenue and audience numbers develop a proportional relationship. In addition, looking at the graph more closely, since the entrance unit price shows stepwise

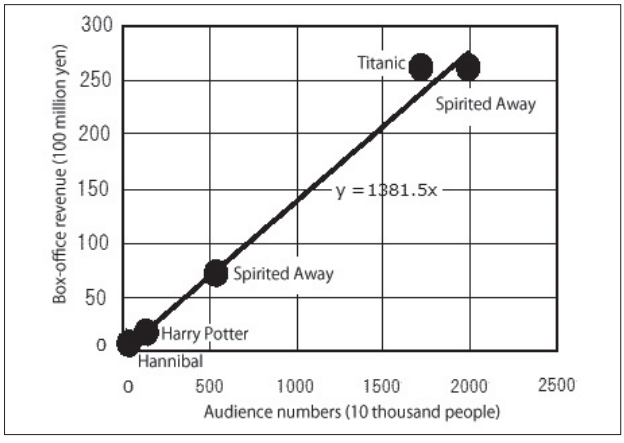


Fig. 3 Movie box-office revenue and audience numbers

changes according to the age of the audience, in regard to the movies “Titanic” and “Spirited Away (*Sen to Chihiro no Kamikakushi*)”, both with around the same box-office revenue, since the “Titanic” has a smaller audience number, its entrance unit price is conversely higher. In other words, it can be presumed that “Titanic” has a higher adult audience ratio than “Spirited Away”.

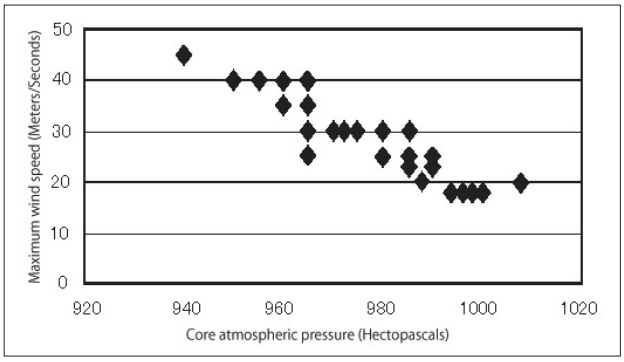


Fig. 4 Typhoon maximum wind speed and core atmospheric pressure

As another sample, information relating to typhoons that has been put into a graph is shown in Fig. 4. By looking at the “maximum wind speed” represented by the vertical axis and “core atmospheric pressure” represented by the horizontal axis, an inverse proportional relationship can be seen. In other words, users can see that when core atmospheric pressure is low, the maximum wind speed is high (large-scale typhoons).

Conclusion

We believe that text mining will become more important as a method to efficiently extracting information required by the user from increasing amount of digitalized information. Of course, current text mining technology is still not capable of extracting without search noise. However, we believe that the day when this technology can extract information in an instant with the same amount of accuracy as a human is not so far away.

(Article written by SAWADA Fumitake, Expert, Research Results and Intellectual Property Management Office, Outcome Promotion Department)