An exploratory analysis of the text mining of news articles about "water and society"

S. Hori
C-PIER, Kyoto University, Japan

Abstract

This paper aims to discover the social interest in the issues of water and society from media reports and to compare it in Japanese and international media. This research uses the online databases of two newspapers: the Japan News and the International New York Times. They are used to understand the difference between the two newspapers. The social interest is discovered by cluster analysis that is, to drive clusters that have value with respect to the problem being addressed. Clustering divides the document collection into mutually exclusive groups based on the presence of similar themes. The articles extracted from those databases are analysed using a KH coder and the generated co-occurrence network. Those papers indicate three keywords; "people," "water" and "government." Words in newspapers always create the first impression of their readers about current topics. The first words and phrases used give rise to common knowledge, which turns out to be regarded as common sense. People are used to being influenced by words.

Keywords: text-mining, water and society, online database, newspaper.

1 Introduction

The work presented here is an exploratory analysis of text mining used for news articles about "water" and "society". It has not often been confirmed visually and is not clear how common interests shared by people are in the relationships. It is unusual to show the connection between two words by an analytical study of newspaper databases. People get environmental information from television and newspapers [1–3]. Words in newspapers always create the first impression of the readers about current topics. The first words and phrases used give rise to

WIT Transactions on The Built Environment, Vol 168, © 2015 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/SD150441

common knowledge and turn out to be regarded as common sense. People are used to being influenced by words.

In Japan, it is very popular to analyse the influence of mass media on society. For example, Shunji Mikami *et al.* [4] examine how mass media in Japan influenced public awareness of global environmental issues. Ayagi-Usui [2], for example, revealed that the coverage of global warming by Japanese newspapers had an immediate but short term influence on the public.

The effect of the mass media on public opinion is quite short lived because media coverage changes from issue to issue and often from day to day [6]. As regards the influence on particular issues, mass media campaigns are often used as tools to attempt to influence public opinion [5]. Most people do not think of environmental problems in terms of either their causes or their effects [7]. Some of the initial evidence shows that most people had heard about environmental problems such as air or water pollution but they often failed to draw any connection between the problem and the important causes, such as overpopulation [8].

This research uses the online databases of two newspapers: the Japan News and the International New York Times. The aim is to understand the difference between the two newspapers. The newspapers are surveyed over a period of one year. The articles extracted from those databases are analysed using a KH coder and the generated co-occurrence network.

2 Methodology

The databases of two online newspapers, the Japan News and the International New York Times, were searched using the term – water and society.

The Japan News is published by the Yomiuri Shinbun, which boasts the largest circulation in Japan. It was known as the Daily Yomiuri until it was renamed the Japan News in April 2013. The Japan News is the leading English language newspaper in Japan [9]. The International New York Times is edited from Paris, London, Hong Kong and New York, its news reporting is tailored specifically for a global audience [10].

The articles of the Japan News are extracted from the "Yomidasu" online database run by the group, Yomiuri Newspaper Company. The contents of the International New York Times were extracted through LexisNexis Academic online database. The databases were consulted on 12th May 2015 looking for articles including the words "water" and "society" which dated from 12th May 2014 to 11th May 2015. 37 articles were found in the Japan News and 175 in the International New York Times.

The articles were examined through text mining. Since nouns are most likely to express the contents of water and society issues, the nouns were tagged after parsing the sentences. The frequency of each parse was counted and the relationships between the tagged parses were visualized by calculating Jaccard indices. The Jaccard index J is calculated as shown in equation (1) in which x and y represent parses include in contents of water and society issues. J indicates the degree of co-appearance of words. In this study, the contents consisting of words with large J will be interpreted as the contents expressing the articles in terms of water and society.

$$J(x,y) = |x \cap y| / |x \cup y| \tag{1}$$

The parses were clustered by using the method proposed by Newman and Girvan [11]. The parses which have strong links with each other are divided into clusters based on the "modularity index" [11]. All parses were analyzed as English words, using the text mining software KH coder [12].

Cluster analysis is a popular technique using data analysis. Clustering in the context of text mining divides the collection of a document into mutually exclusive groups based on the presence of similar themes. Themes help in a better understanding of the concepts or events [13].

3 Results

A total 217,126 words were extracted from the International New York Times, and they included 4,184 nouns. Furthermore, there were 52,121 words from The Japan News, giving 10,847 nouns. The results are outlined below:

3.1 The results of the Japan News

Fig. 1 shows the co-occurrence networks of the Japan News, which is visualized by betweenness centrality, the word "hydrogen" tends to co-occur with "people" and "government". Fig. 2 shows the co-occurrence network visualized by betweenness and it also shows the result of cluster analysis.

The first cluster took information such as "nation", "power", "government", "system", and "measure." For example, the article on 29th December 2014 with the news title "Japan in Depth, Stimulus aims to lessen people's burdens." A part of the article writes about "*The government will also help households purchase Ene Farm, a home system to generate electricity and heat by a chemical reaction of hydrogen contained in city gas with oxygen in the air. If a person replaces an existing conventional hot water supply system with Ene Farm, the government plans to subsidize cost up to 350,000 yen. The government expects 60,000 applications for this subsidy system".* This cluster is about environmental policy relevant to water supply.

The second cluster results from information such as "country", "right", and "security." For example, the article on 1st January 2015 has the news title "2015: A year to break ground on the road to a bright future." A part of the article is about *"There is also a pressing need to reform the social security systems with the aim of sustaining our nation as a society in which people can live in safety."* and *"China must be urged to restrain itself from provocative behavior in waters around the Senkaku Islands in the Okinawa Prefecture."* The subject of this cluster is the issue of water and territory.

The third cluster results from information such as "area", "number", "work", "world", "time", "year", "people", "society," "company", "problem", "effort" and "number." For example, the article on 25th August 2014, has the subject

"Child Poverty SOS, Living with hunger in a candlelight world." This is about the problem of child poverty in Japan. To quote from this article, "One in six children live in poverty in Japan,…*The nation's child poverty rate was at the record high of 16.3% in 2012 according to figures released on July 15th by the Health, Labor and Welfare Ministry". "When her mother came home, they took some empty plastic bottles to a nearby place to fill them with water."* This cluster shows the issue of water and poverty.

The fourth cluster results from information such as "Yen" and "house." For example, the article on 1st April 2015 has the title "Government must support self-reliance of the needy by extending assistance early." This article is about a new system started by local governments to help poor people become self-reliant. A part of article is that, *"Local government employees should find those in need of such assistance by checking the records of arrears in the residence tax or water bills and by utilizing the information gathered by social welfare workers".* This cluster is also the problem of water and poverty.

The fifth cluster results from information such as "hydrogen", "plan", "facility" and "event". For example, the article on 6th January 2015 with the title "Olympic village to be the 1st 'hydrogen town'." This is about the decision of the Tokyo metropolitan government that the Athletes' Village for the 2020 Olympic and Paralympic Games will be made into a "hydrogen town" where electricity and hot water are supplied from hydrogen energy". The fifth cluster is about water utility.

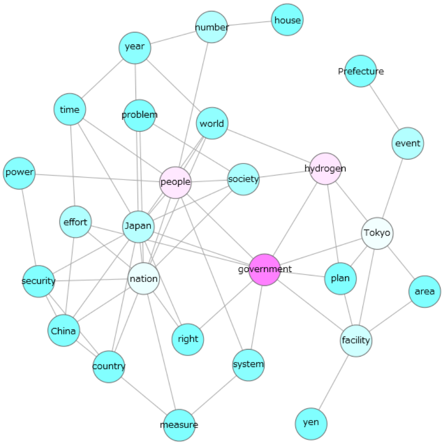

Figure 1: Co-occurrence networks of the Japan News, the word "hydrogen" tends to co-occur with "people" and "government".

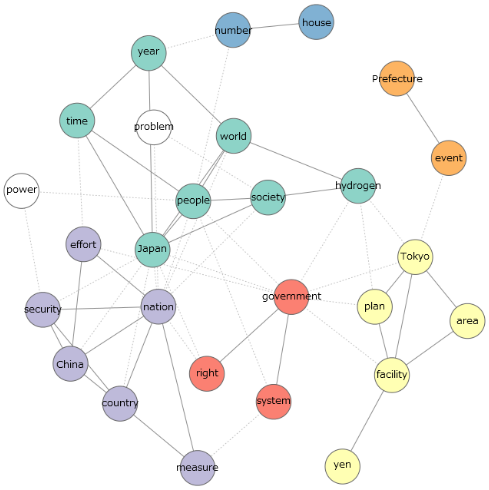

Figure 2: Co-occurrence network of the Japan Times with the result of cluster analysis.

3.2 The results from the International New York Times

Fig. 3 shows the co-occurrence networks of the International New York Times. The word "water" tends to co-occur with "people", "government", "year", "time" and "country".

The first cluster took information such as "time", "people", and "country." For example, the article on 15th November 2014 with the title "With Ebola, family ties fall apart; In Liberia, a mother flees the stigma arising from the deaths of loved one." It is about *"Though Redemption often did not have running water, it was one of the biggest medical centers in Liberia."* The first cluster is about the issue of water shortage and the lack of water utility.

The second cluster took information such as "government", "year", and "water." For example, the article on 27th December 2014 with the title "Mover forward with Cuba." This is about *"Cuba is the largest island neighbor of the United States. Opportunities for joint research, the definition of territorial waters and the prevention and clean up of toxic spills will benefit both the United States and the Caribbean basin."*

Another article about *"the National Audubon Society foresees danger for more than half of the 650 species of birds in North America"* has the title *"Mass climate disruption is forecast for North American birds."* A part of the article is *"Birds could feel the impact of a changing climate in different ways. Drought in Southern California is blamed for a sharp drop in the breeding among California raptors perhaps because the lack of water is killing the insects and small rodents they feed on"*. This cluster is about environmental conservation.

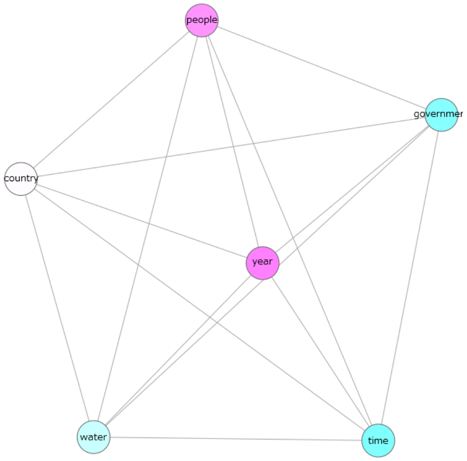

Figure 3: Co-occurrence networks of the International New York Times, the words “water” tends to co-occur with “year”, “people”, “government”, and “country”.

4 Discussion

The Japan News shows mostly 4 subjects:
(1) Environmental policy (relevant to water supply);
(2) Water and poverty;
(3) Water utility;
(4) Water and territory.

The International New York Times indicates mainly 4 subjects:
(1) Environmental conservation;
(2) Water shortage;
(3) Water utility;
(4) Water and territory.

The categories that overlap are:
(1) Water utility;
(2) Water and territory”.

The Japan Times has the tendency to cover domestic news. The issue of water and poverty in Japan is unexpectedness. The International New York Times covers more global environmental problems and international interests. Two online databases were picked up from different media sources. However, it is interesting to note that both results draw pictures using almost similar keywords for co-occurrence networks. The significant keywords in the issues of water and society are, in general, three: “people”, “water”, and “government”. This draws a conclusion from evidence of Fig. 1 and Fig. 3 which is co-occurrence network visualized by betweenness centrality. According to this research, the press tend to write articles in relation to those words.

This study covers the current period of 1 year. The date of the extraction of the data was 12th May 2015 in both cases. However, it is obvious that news articles are not written regularly throughout the year. The total number of articles using words of “water” and “society”: was quite different. There were only 37 from The Japan News whereas 175 articles came from the International New York Times. If it were possible to use a longer period for the study, the results might be different. This merits further research.

There is a limitation of this database research. The accumulated data for analysis is not only articles about water resource, but also the territorial issue, which is a highly political problem. To enhance the objectiveness of the assessment process, those territorial issues are not avoided in those research data.

5 Conclusions

Those articles relating to the issues of water and society are composed of the same material as those of “people”, “water” and “government.” These composition elements are important keys to understanding the situation of articles written in a newspaper.

The analysis of newspapers using text mining can improve at beginning the social understanding of the appropriate and effective solutions for water management and social activity. Despite their simplicity, the results obtained by these methods give relatively accurate results which are sufficient for the policy and management of water resources.

References

[1] Schoenfeld, A.C., Meier, R.F., Griffin, R.J., 1979. Constructing a social problem: the press and the environment. Social Problems 27(1), 38-61.

[2] Ayagi-Usui, M., 2008. An analysis of the effective factors for promoting pro-environmental actions from the information gain and social capital point of view. Review of Environmental Economics and Policy Studies 1(2), 37-50 (in Japanese).

[3] Slovic, P., 2000, Informing and educating the public about risk, In: Solvic, O. (Ed.), The Perception of Risk, Earthscan Publications Ltd, London.

[4] Shunji Mikami, Toshio Takeshita, Makoto Nakada and Miki Kawabata, “The media coverage and public awareness of environmental issues in Japan (1995), Kluwer Academic publisher: 209-226.

[5] Y. Sampei, M. Aoyagi-Usui, Mass-media coverage, its influence on public awareness of climate-change issues, and implications for Japan’s national campaign to reduce greenhouse gas emissions, Global Environmental Change 19 (2009), 203-212.

[6] Driedger, S.M., 2007. Risk and the media: a comparison of print and televised news stories of a Canadian drinking water risk event. Risk Analysis 27(3), 775-786.

[7] Richard F. Carter, Keith R. Stamm, and Katharine Heintz-Knowles, "Agenda-setting and consequentiality," Journalism Quarterly 69, no. 4 (1992): 869-877.

[8] Regina A. Simon, "Public attitudes toward population and pollution", Public Opinion Quarterly 35 (1971): 95-102.

[9] The Japan News (2015, May 1st), The world inbound, Japan outbound, The Japan News by Yomiuri Shinbun. Retrieved from http://adv.yomiuri.co.jp/thejapannews/index.html

[10] Arthur Ochs Sulzberger Jr. (2015, May 1st), Introducing The International New York Times, Retrieved from http://www.nytimes.com/2013/10/15/businesss/media/introduceing-the-international-new-york-times.html?_r=0

[11] Newman, M., and Girvan M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69, 026113.

[12] Higuchi, K. (2013). KH Coder. http://khc.sourceforge.net/en/

[13] Goutam Chakraborty, Murali Pagolu and Statish Garla (2013), Text Mining Analysis, Practical Methods, Examples, and Case Studies using SAS, SAS Press.