Analytical Framework of Relations among Nations using News Articles

Takeru Yokoi, Masato Fukuchi, Michihiro Kobayakawa
Tokyo Metropolitan College of Industrial Technology
Shinagawa, Tokyo, Japan
dr.takeru.yokoi@ieee.org

Roliana Ibrahim, Ali Selamat
Faculty of Computing, Universiti Teknologi Malaysia
Johor Bahru, Malaysia

ABSTRACT

News articles are a type of instantaneous and regional media, and provide the daily concerns of its publication area. This work proposes an analytical framework of the relations among nations using the similarities in the content of news articles published in different nations. Our key idea is that those relations exist in the news articles miss-classified by different nations from their original publication area. In order to clearly illustrate those relations, the classification results are visualized as bar graphs. We also carried out some experiments for the proposed framework using a small collection of news articles.

Categories and Subject Descriptors

M.9 [Knowledge Valuation]; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

Keywords

Text Mining, Relation Analysis, Miss-Classified Items, Spatiotemporal Mining

1. INTRODUCTION

In recent years, we are able to instantaneously obtain news articles not only about our own nation but also from around the world through the internet. The main focus of the news articles concerns matters of relevance to each nation. Thus, if the concern has a common point among multiple nations at the same time, the point is regarded as their relation or association.

Such relations in the texts on geolocation and time have recently been the focus of much attention. For instance, "News Stands"[1] is one of the representative systems. The system searches and shows news articles on a map by queries consisting of the date.

In this work, we have proposed an analytical framework of the relations among nations using the similarities in the content of their respective news articles. In addition, we tried to find the relation with an actual small collection of news articles by the proposed framework.

Copyright is held by the author/owner(s).
WWW 2016 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889363.

2. PROPOSED ANALYTICAL FRAMEWORK

In this section, our key idea for the proposed analytical framework and its procedure are introduced.

2.1 Key Idea of the Proposed Framework

The key idea for the proposed framework can be briefly described as "Miss-classified news articles indicating relations among nations". In this paper, miss-classified news article means what is miss-classified by different nation from its original publication area. Miss-classified news articles denote the common concerns among different nations, they indicate their relation or association which can be analyzed.

In Figure 1, $\triangle$s and $\circ$s denote news articles which are originally published nations A and B, respectively. $\times$s denote the centroid of each cluster. The filled $\blacktriangle$s and $\bullet$s denote miss-classified news articles. In usual classification framework, the number of news articles belonging to the correct cluster is important. However, the proposed framework focuses only on the miss-classified news articles, which are illustrated as $\blacktriangle$s and $\bullet$s in Figure 1, and supposes that they indicate the relation or association between nations.

2.2 Framework Procedure

The proposed framework procedure consists of six main factors: 1) feature extraction, 2) similarity derivation, 3) mapping into Euclidian space, 4) clustering based on the similarities between content features, 5) labeling nations, and 6) finding relations.

First, the framework extracts features from the content of news articles. A set of general and proper nouns in the news article is regarded as its feature content. The news article d_i is defined as a set of undeclined nouns and described as:

$$d_i \equiv \{w_{i1}, w_{i2}, \cdots, w_{iV_i}\}$$

where V_i denotes the sorting number of common and proper nouns in the ith news article. In addition, $w_{im} \neq w_{in}$ for all m and n and their frequency is excluded.

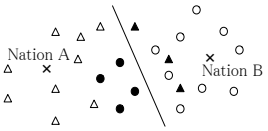

Figure 1: The relational indicators between nations.

Second, the similarity $s_{i,j}$ between ith and jth news articles is derived by the Jaccard coefficient between sets d_i and d_j. Using $s_{i,j}$, the dissimilarity $\bar{s}_{i,j}$ is defined here as follows:

$$\bar{s}_{i,j} = 1 - s_{i,j}.$$

The range of $\bar{s}_{i,j}$ is $[-1, 1]$ and the close value of $\bar{s}_{i,j}$ to 1 denotes that the content of the ith and jth news articles is not similar. To classify news articles, mapping news article into Euclidian space within arbitrary dimension is performed by nMDS[2] based on their dissimilarities. In addition, the news articles are classified by k-means in the mapping space.

After classification with k-means, new nation labels are assigned to each news article. The new label is decided as the original nation which mainly dominates the cluster. In addition, each cluster is assigned to a different label. After labeling a new nation to each news article, analysis is carried out by comparing the new label with its original label. If the new label is different from the original label, a relation can be supposed to exist between those nations.

3. EXPERIMENTS AND RESULTS

In this section, we introduce the experiments of the proposed analytical framework with the small news article collection.

3.1 Experimental Condition

We carried out some experiments of the proposed framework with a small collection of news articles. In these experiments, 50 news articles written in English from 5 nations– Japan, USA, UK, China and Malaysia– were collected for the experimental data. The original label of those news articles were determined by the location of the publisher. All of the news articles were published in 2 days: 11/14/2014 and 11/15/2014.

The number of clusters for k-means was set at 5, i.e., equal to the number of nations targeted in these experiments. In addition, the dimensions of Euclidian space to map the news articles by nMDS were set at 8, 16 and 64.

3.2 Experimental Results

Figure 2 visualizes the mixing number of the k-means results for each dimension. The vertical axis denotes the original label which describes the nation where those articles were published. The color labels in the bar graphs denote the new label assigned to each news article. The horizontal axis denotes the number of articles.

3.3 Discussion

From the results illustrated in Figure 2, the proposed analytical framework enabled the determination of relations by considering the mixing of articles from different nations. The remarkable relations between the USA and UK could be observed independently of dimension. This denotes that

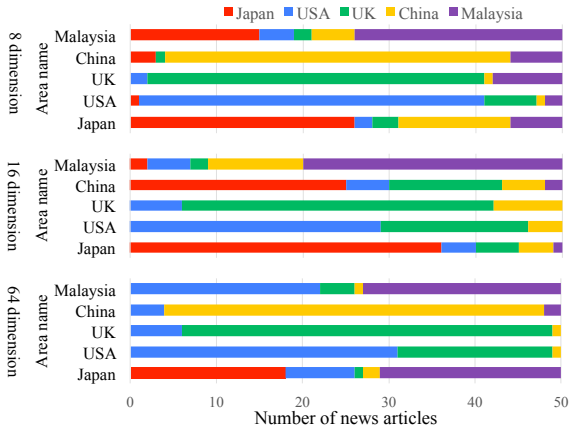

Figure 2: Color bar visualization of the mixing number of news articles by k-means results within 8, 16 and 64 dimensions.

the USA and UK have common concerns and indicates their association points during this period.

Some relations were observed depending on the dimension of Euclidian space. For instance, the relation between Japan and China was observed in lower dimension 8 but hardly in higher dimension. This denotes that news articles were assigned to their original nation in high dimension space even though those news articles included some relations. A more complex hyper plane to classify news articles can be constructed in high dimension space than in low dimension space. More discussion is necessary to obtain data that the relation is stable.

It is also necessary to mention asymmetric relations. In Figure 2, the asymmetric relation between Japan and China was observed in the results of dimension 64. Japan includes news articles labeled as China but China does not include news articles labeled as Japan at all. For instance, Japanese news articles raised the problem of red coral fishing, however, even though it is a common concern, the Chinese news articles hardly mentioned it. Like this case, the proposed analytical framework elucidated the differences in concerns as well as asymmetric relations.

4. CONCLUSION

We have proposed an analytical framework of relations among nations using the similarities in news article contents. The experiments were conducted with a small collection of news articles published in 5 nations in 2 days, thus indicating the existence or non-existence of relations among nations.

5. REFERENCES

[1] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. Reading news with maps by exploiting spatial synonyms. *Communications of the ACM*, 57(10):64–77, 2014.

[2] Y. h. Taguchi and Y. Oono. Relational patterns of gene expression via non-metric multidimensional scaling analysis. *Bioinformatics*, 21(6):730–740, 2004.