

EXTRACTING AND VISUALIZING PEOPLE'S NEEDS AND TOPIC TRENDS FROM USERS' COMMENTS ON VIDEO STREAMING SITES OR BLOG POSTS

David Ramamonjisoa, Tatsuya Suzuki, Basabi Chakraborty

Faculty of Software and Information Science, IPU

152-52 Sugo Takizawa Iwate, Japan

ABSTRACT

This paper describes a method and prototype deployment for extracting people's needs from social blogs or by analyzing trends in social media. Case studies presented in this paper include (1) video comments from a streaming site in Japan for a week in March 2011 during the TEPCO Fukushima accident and (2) the SAVE IWATE blog where people from the region Iwate in northern Japan were able to post their needs and comments relating to their daily life after the Great Tohoku Earthquake in 2011.

KEYWORDS

text data mining, blogs and comments topic analysis.

1. INTRODUCTION

Analyzing dynamically evolving textual data is a task for social knowledge and information management. Thanks to the rapid and exponential growth of social sites in the last few years, this need has suddenly arisen and taken on dramatic urgency. Live video streaming sites, Twitter, Facebook, and blog sites such as USTREAM and NICOVVIDEO in Japan, allow users to post comments in real-time, accumulating an important amount of social data on these platforms. A social knowledge task gathers and records the concerns of people and problems following an event. These concerns may consist of the general trends or needs of individuals or groups participating in video comments or blog articles. Generally, this task is performed manually by collectors and poll media staff to devise business marketing strategies or political and social action. This is because users' comments can produce a new consensus among users and this consensus has a bearing on users' thoughts. Thanks to text mining techniques, topic trends or users' needs can be analyzed and summarized autonomously.

This paper describes an experiment to replace a human with intelligent systems to perform the task (or part of it) for him. Intelligent systems have a natural language processing capacity, a feature extraction module, text mining techniques for co-occurrence, topics extraction function, pattern-matching of social concerns according to a social pattern database, and a temporal change of focuses on a problem.

In this paper, we will first present the methods used during the experiment and the datasets. Next, we will detail the experiments and discuss the results obtained. Finally, we will draw conclusions and discuss future work.

2. APPROACHES

The general architecture of the system is depicted in the flowchart in Fig. 1. The data model for the experiments is described as follows:

Users' comments or blog posts are designated as document collections. The model of the document collection is described below:

$$D = \{d_i\} \quad \text{where } d_i = (\text{docID}, \text{day or month}, \text{title}_i, \text{content}_i)$$

A natural language processing (NLP) task was processed to extract important keywords such as nouns or adjectives from the content_i of each d_i . A bag-of-words model was constructed by attaching a weight to the extracted words. A weight may be just the frequency of a term, $tfidf$ [Salton, G. et McGill, M.J., 1986] or LSA [Landauer, T. K. and Dumais, S., 2008] results. The content of the document is then a set of tuple keywords and weights as follows:

$$\text{content}_i = \{(k_{ij}, w_{ij})\} \quad j \in [1..n], n : \text{number_of_keywords_in_the_content}, w_{ij} > \tau(\text{threshold})$$

A document collection is therefore a table where rows consist of the weights of each keyword in each document and columns list the documents. This document list is arranged as time-series data so that old posts and comments are the first element of the list and the newest comments and posts are the last. The document table is formalized as follows:

$$D^T = [\text{content}_i(\text{row}) \times d_j(\text{column})] \quad i \in [1..m], m : \text{number of documents in the collection}$$

Next, topic extraction and graph analysis programs based on text mining techniques and data analysis algorithms with visualization tools were executed to build a visual output for the end user. This process outputs the top n keywords, charts, or graphs representing topics or trends during a given period.

Related research on Twitter analysis is described in [Yang, S. and Kavanaugh, A. L., 2011].

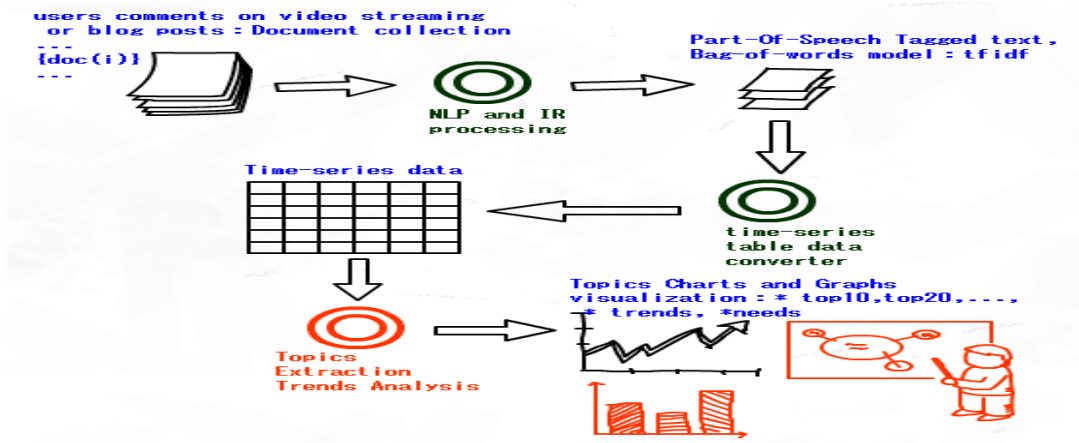


Fig. 1: System Architecture

The next section describes the method used for the extraction of topics.

3. TOPICS EXTRACTION AND ANALYSIS ALGORITHMS

Table 1: Procedure for extracting topics and analyzing topic trends

<ol style="list-style-type: none"> 1. We assume that the documents are gathered together and that we do not need to crawl the web or process raw html data. $D = \{d_i\}$ is given. 2. Build the document table D_1^T by setting the weight as a term frequency $w_{ij} = tf_{ij}$ 3. Remove high and low frequency words and predefined stopwords in a list (slang used in video stream comments or specific terms in blogs such as single character terms)
--

4. Plot keywords representing the top n keywords according to their weight. This figure represents the general topics of the documents.
5. Rebuild the document table D_2^T by changing the weight w_{ij} $w_{ij} = tf_{ij} \times idf_i$
6. Plot charts representing the Top n keywords according to their weight. This figure depicts topics varying along the interval of days or months and shows rare topics within the collection.
7. Calculate the correlation between keywords over the timeline, and extract keywords that exceed the threshold
8. Search for the co-occurrence of words of a keyword within the document using a Dice or Jaccard coefficient measure. Temporal changes on focused topics can be traced.

The algorithm in Table 2 is a modified version of the one proposed by Fukuhara et al. which they used for the analysis of weblog articles [Fukuhara, T. et al., 2007].

Table 2: Algorithm for finding daily or monthly topics

Let **DM** be the set of days or months. If we want to know about topics during Q days or months, $\mathbf{DM} = \{\mathbf{dm}_1, \mathbf{dm}_2, \dots, \mathbf{dm}_Q\}$
Let **W** be the set of words that have appeared through Q days or months. If we seek P words during Q days or months, $\mathbf{W} = \{w_1, w_2, \dots, w_P\}$
For each $w_i (1 \leq i \leq P)$ in **W**, repeat as follows
 For each $\mathbf{dm}_j (1 \leq j \leq Q)$ in **DM**, repeat as follows
 1. Let \mathbf{cpi}_j be the number of comments or posts containing w_i on \mathbf{dm}_j
 2. Calculate $\mathbf{sum}(\mathbf{cpi}) = \sum_{j=1}^Q \mathbf{cpi}_j$, max value: $\mathbf{max}(\mathbf{cpi})$, and SD(Standard Deviation)/average ratio: $\mathbf{sd}(\mathbf{cpi})/\mathbf{avg}(\mathbf{cpi})$
 3. Print w_i as a topic word of day/month \mathbf{dm}_j if $(\mathbf{sum}(\mathbf{cpi}) \geq \mathbf{th1})$ and $(\mathbf{max}(\mathbf{cpi}) \geq \mathbf{th2})$ and $(\mathbf{sd}(\mathbf{cpi})/\mathbf{avg}(\mathbf{cpi}) \geq \mathbf{th3})$ where $\mathbf{th1}$, $\mathbf{th2}$, and $\mathbf{th3}$ are thresholds.

4. EXPERIMENTS

Our experiments were based on two datasets. The first dataset was obtained from previous research in [Hashimoto, T. et al., 2012]. The second was retrieved manually from the Iwate Banya Nippo website¹.

4.1 Dataset 1: Tokyo Electric Power Plant accident interview comments

The first dataset is a collection of users' comments obtained from NICOVIDEO during live press conferences of the Tokyo Electric Power Company (TEPCO) between March 15th and March 20th, 2011. On average, there were 2400 comments a day and 14450 in total.

4.2 Dataset 2: Save Iwate blog posts

The second dataset is a collection of users' posts on a blog for the refugee camp or disaster area after the Tohoku Great Earthquake on March 11th, 2011. This blog enabled people to express their needs and report their concerns, so that the organization responsible for the site could understand their requirements and provide assistance. The site was also used to report social activity conducted by the organization in order to raise funds from readers. The website for the blog was called "SAVE IWATE" because it focused only on the disaster area of the Iwate prefecture. The collection has 700 posts from June, 2011 to December, 2012, which are categorized into various categories such as "voices from victims", "voices from inland evacuees", "voices from coastal residents", and "wanted to start support."

4.3 Experimental environment

We used the Python programming language to code scripts for basic processing. We also used a ready-made text mining tool [Higuchi, K., 2012] to deal with topic extraction and trend analysis for steps 7-8 in the

¹ SAVE IWATE BLOG: <http://sviwatebanya.wordpress.com/>

the dataset. Manually, we discovered topics such as ‘Chernobyl’, ‘Three miles’, ‘Caesium’, ‘Iodine’, and ‘Sievert’ in the last phase of the document collection. Despite overseas panic, repatriation, and heated discussions, the people in Japan had been overwhelmed by successive events such as the power outage, and the shortage of gas, energy, and food supplies. Therefore, the long term problem was excluded from community discussion during the interview with the TEPCO president.

The topic extraction results according to the algorithm in Table 2, and implemented using a Python script, were compared with the results of the algorithm in Table 1. We set the thresholds according to the document collection size. For example, for two documents (two days), as a set of 4000 comments and 500 keywords, we set th1, th2, and th3 to 5, 3, and 0.2, respectively, for a good size and quality of topics, enabling us to compare with the bag-of-words models where weights are set to $tf > 5$ and $tfidf > 0.5$ (i.e. top150).

5.2. Results from Dataset 2: SAVE IWATE topics

The collection has on average 40 blog posts a month. Posts were categorized manually by the administrator. We disregarded those categories, arranged all posts chronologically, and divided them by month. We ran the procedure detailed in Section 3 for each month’s data and drew charts of the top 50 keywords. An example plot of the top 50 keywords with the weight set as $tfidf$ can be done easily. The document is labeled ‘June 2011’ as the starting point of the blog. During this period, disaster areas were full of rubble and victims needed basic life assistance (three meals a day, pans, cup noodles, maternity care). Lifelines were disrupted and people required assistance to meet certain basic needs. Organizations gathered donations from all over Japan and the ‘Donation’ keyword has the highest rank in the top 50. The distribution of goods is also relevant as is volunteer (staff) recruitment to assist victims. Kitchen tools like ‘kettle’ and ‘pan’, and goods such as battery, hat, mat, etc. are also among the top 50 keywords. Peoples’ needs in the community were reflected in the blogs and were extracted by our system.

Another result of our experiment is the ability to extract the top n keywords from each category defined by the blog administrator, allowing him to verify and autonomously index the blog posts. We run the program for the category “wanted to start support subjects” for the entire period (18 months). Topics related to support for the elderly and reconstruction are apparently popular requests. The results reveal that some people were starting to seek funding for the reconstruction of homes but the government doesn’t want them to build on the same place (coastal area) for future Tsunami prevention. A cluster of keywords [Banya – continue – support – you – more] has been found and it means that the SAVE IWATE organization is continuing to support.

The results of the algorithm in Table 2 for this dataset during the month June 2011 is presented in Fig. 3.

```

>>> f=open('topics_1106month_eng.txt')
>>> topics=l.strip().split(',') for l in f]
>>> topTopics = [topics[l][0] for l in range(len(topics))]
>>> f.close()
>>> topTopics
[['Goods', 'House', 'Home', 'Donations', 'Car', 'Goods', 'Informa
tion', 'Foot', 'Housing', 'Destination', 'Original 4', 'Town', '
Food', 'Shop', 'Supplies', 'Hat', 'Government', 'Local', 'Hand',
'Food', 'Next', 'Food', 'Children', 'Miso', 'Bank', 'Apartment',
'Home', 'Money', 'Personally', 'Rubble']]
>>> Topics_ja_en
[['Goods', '物資'), ('House', '家'), ('Home', '宅'), ('Donations',
'義捐'), ('Car', '車'), ('Goods', '品'), ('Information', '情報'),
('Foot', '足'), ('Housing', '住宅'), ('Destination', '先'), ('Orig
inal 4', '元'), ('Town', '町'), ('Food', '食料'), ('Shop', '店'),
('Supplies', '用品'), ('Hat', '帽子'), ('Government', '行政'), ('L
ocal', '地元'), ('Hand', '手'), ('Food', '食糧'), ('Next', '隣'),
('Food', '食べ物'), ('Children', '子供'), ('Miso', '味噌'), ('Bank
', '銀行'), ('Apartment', 'アパート'), ('Home', '実家'), ('Money',
'金銭'), ('Personally', '個人'), ('Rubble', 'ガレキ')]
>>>

```

Fig. 3: Top topics during the month June 2011 in Save Iwate blogs

The problem with the algorithm in Table 2 is that we have to run the program several times by adjusting the values of the thresholds (th1,th2,th3) in order to obtain representative topics for the document (month). For example, the results in Fig. 3 were obtained with very low thresholds (th1=3, th2= 2, th3=0.0). Unsurprisingly, the results are similar to that in algorithm 1 results.

Another experiment to be conducted is to extract compound nouns in the sentences as they are common in the Japanese language. We can work out the needs of people with more details such as “fan for summer”, “hat for summer”, “stove during the winter”, “money distribution”, “shops next home”, “staple foods”, and so on. We can define the patterns of people’s needs according to the season, the time interval starting from June 2011 to December 2012. After one and a half years, our analysis shows that the majority of people tended to forget the situation and returned to a normal life of work, education, leisure activities (hobbies, sports, and amusements), and participation in social events (concerts, festivals, and parties).

6. CONCLUSIONS AND FUTURE WORK

We conducted two experiments for extracting and analyzing people's needs and topic trends from online social data in a time-series. Our researches on topics discovery and topic maps in the past [Ramamonjisoa, D, et al., 2001] [Ramamonjisoa, D., 2003] [Ramamonjisoa, D. and Tan T., 2011] have motivated us to pursue this text data mining research on time series data. We used natural language processing and text mining techniques to process the data and adapted algorithms for the task. Text mining techniques, such as bag-of-words, co-occurrence networks, and topic extraction based on statistical features show the same results as described by the algorithms in Table 1 and Table 2. We can dynamically visualize these topics during a given period using an animation of daily or monthly graphs of topics, and we achieved our goal of visualizing the needs of people by using topic graphs and topic clouds.

At the moment, we are analyzing data offline and so we cannot predict the evolution of a topic into the future. Topic transition discovery can be implemented in the future to autonomously discover the needs of people in blogs for a particular event and report them to administrators. An extension of the system using linked data such as Wikipedia or Wordnet [Medelyan, A., 2008] is possible for categorizing peoples' needs into defined and higher level class labels, thereby allowing abstract views for larger situations in the virtual community.

ACKNOWLEDGEMENT

This work is supported by the Iwate Prefectural University. We acknowledge Dr. T. Hashimoto to provide us the Nicovideo data.

REFERENCES

Book or Part of Book

Salton, G., and McGill, M.J., 1986. *Introduction to modern information retrieval*. McGraw-Hill, Columbus, USA.

Ramamonjisoa, D. et al., 2001. *Research Topics Discovery from WWW by Keywords Association Rules*. In *Rough Sets and Current Trends in Computing*, LNAI 2005. Ziarko, W. and Yao, Y. (Editors), Springer, Berlin, Germany.

Journal

Fukuhara, T. et al., 2007. Analyzing concerns of people from Weblog articles. In *Journal of Artificial Intelligence and Society*, vol. 22, p.253–263, Springer Ed.

Conference paper or contributed volume

Hashimoto, T. et al., 2012. Discovering Emerging Topic about the East Japan Great Earthquake in Video Sharing Website. *Proc. of IEEE TENCON 2012*, Cebu, Philippines.

Landauer, T. K. and Dumais, S., 2008. Latent semantic analysis, *Scholarpedia*, 3(11):4356.

Medelyan, O. et al., 2008. Topic indexing with Wikipedia. In *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, AAAI Press, Chicago, USA, 13 July, 2008. pp. 19–24.

Ramamonjisoa, D., 2003. Research Topics Discovery from WWW by Agent Systems. *Proc. of International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, e-Medicine on the Internet*, Aquila, Italy.

Ramamonjisoa D. and Tan., T., 2011. Modeling, Extracting and Visualizing an Organization's Knowledge with Topic Maps. *Proc. volume II of the 2nd International Multi-Conference on Complexity, Informatics and Cybernetics*. Orlando, USA, pp. 242–247.

Yang, S. and Kavanaugh, A. L., 2011. Collecting, analyzing and visualizing tweets using open source. *Proc. of Digital Government Innovation in Challenging Times*, pp.374–375.

Other Resources

Feinberg, J. 2012. Tags cloud visualization tool. www.wordle.net

Higuchi, K., 2012. KH Coder Text Mining Tool, Free GPL Software