CF-007

社会的イベントに対する関心抽出のためのトピックトレンド分析法に関する検討
A Note on a Topic Trend Analysis Method for Extracting Public Interest in Social Events

Hoan Le Huu[1)]　原川 良介[1)]　岩橋 政宏[1)]
Ryosuke Harakawa　Masahiro Iwahashi

Abstract

This study introduces a novel method for topic modeling by combining trend features and semantic features. Previous methods extracted topics based on only semantics, and the trend representation was simply counting the frequency of occurrence over time. Therefore, these methods could not eliminate the indirect correlations among a topic. To address these, our method introduces trend clustering into topics extraction that enables us to remove indirect correlations and provide an efficient topic-trend visualization solution. Experimental results show our method outperforms previous methods in enhancing the interpretability of trend-topic in revealing human interest in social events.

1 Introduction

In social science, it is difficult to measure human behavior by a machine. Traditionally, we need to conduct surveys or interviews with in human behavior analysis. However, it is challenging to conduct research with a wide range. Information from the Internet is a potential source to explore human patterns. Taking advantage of information from the Internet allows us to explore new knowledge efficiently.

One of the popular methods for data mining is topic modeling [1]. Topic modeling is an automatic method to extract the latent topics from vast amounts of documents. Most conventional methods use the statistical method or natural language processing technique [2, 3, 4] to extract latent topics.

Trend of topics is evolution of topics over time that can support in analyzing the changes in public opinion under social event affect. Previous methods only use semantic features and trends of topics are obtained by simply counting the frequency of topics over time. Therefore, trends representation is discrete and impossible to understand how events affect topics.

In addition, these methods use only semantic features to produce topics; therefore, indirect correlations between keywords in terms of time remain. This makes the trend of the topic difficult to analyze the actual influence of society on human behavior. For example, the Christmas event often falls in the winter season. The temperature in the winter season is low and leads to an increase in the number of seasonal influenza cases. Thus, the correlation between the Christmas event and seasonal influenza is indirect, as it is influenced by temperature.

The research [5] introduced a method for clustering topics based on time-series features. This method provides an effective solution for representing trends. Moreover, by leveraging of direct correlation, this method eliminate the indirect correlations between keywords in topics that previous methods (e.g, k-means [6, 7, 8], k-shape [9]) could not. Therefore, it improves the explainability of the extracted topics. However, due to clustering based on time series characteristics only, topic coherence lacks.

Changes in public interest in social events can provide important indicators for understanding changes in human behavior under the impact of social events. These indicators are needed in many fields such as market change analysis, disaster response, behavioral psychology trend analysis, etc. Therefore, a more effective method is needed to extract information about trending topics that can be interpretable. To overcome this limitation, we propose a new approach that combines trend features and semantic features for trend-topic representation to explore public interest changes in social events. Specifically, we use Graphical Lasso-guided Iterative Principal Component Analysis (GLIPCA) [5] for trend clustering and use BERT [10, 11] to produce the word embeddings. Following that, Affinity Propagation [12] is applied to cluster into topics.

The novelty of this study includes three main points:

- Our method improves the topic coherence compared with a method that uses only trend features (Novelty 1)
- The trend representation of our methods is more interpretable than BERTopic [2] method. (Novelty 2)
- Benefiting from removal of indirect correlation, our methods could remove irrelavent words from topics. This improves the interpretation of the actual relationship (direct correlation) between the keywords of the extracted topic. (Novelty 3)

To demonstrate these, we conduct experiments with a Vietnamese newspaper dataset [13] to analyze how Tet holiday [1)] affects on public interest in Vietnam.

2 Methodology

The architecture of the proposed method is illustrated in Fig. 1 and divided into three phases: (1) preprocessing and keyword extraction, (2) trend clustering, (3) semantic clustering.

2.1 Preprocessing and Keyword extraction

Preprocessing includes removal of HTML format, lowercase, and punctuation removal, morphological analysis, and stopword removal. Vietnamese is the isolating language; therefore, we do not need to conduct stemming and lemmatization for documents. Nouns are more informative and representative than other types of words (e.g, verbs, adjectives, etc). Moreover, research [14] has shown that topic modeling with preserving only nouns could improve the topic coherence and reduce the time processing. Therefore, in this study, we only keep nouns to process. Next, TF-IDF [15] is used to score the importance of words from documents. The cut-off threshold is set, and only words satisfy the threshold are kept.

1) Nagaoka University of Technology, Graduate School of Engineering

1) Tet holiday, also known as Lunar New Year, is the main holiday that starts from January 1st to January 10th of the lunar calendar. Preparations for Tet holiday usually start earlier. This is the biggest festival in Vietnam with cultural activities, entertainment, visiting relatives, shopping, etc.

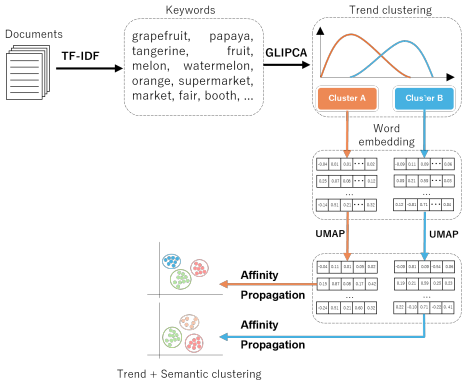

Figure 1 Architecture of the proposed method

2.2 Trend Clustering

Semantic features only reflect the meaning of keywords. Therefore, it can not formulate the relationship between social events. We use GLIPCA for trend clustering because GLIPCA can formulate the correlation relationship between keywords. To address that, GLIPCA establishes the partial correlation matrix L by using the graphical lasso algorithm [16].

$$\hat{\Sigma}^{-1} = \arg\max_{\Sigma^{-1}} \left(\ln \det \Sigma^{-1} - \mathrm{tr}(S\Sigma^{-1}) - \rho\Sigma_1^{-1} \right) \quad (1)$$

L is a matrix that calculate the correlation between two keywords while eliminate the effect of other keywords (direct correlations). Next, we solve eigenvalue problems of the partial correlation matrix L and keep the first component $\mathbf{u} = [u_1, u_2, ..., u_M]^T$ where M is the number of keywords.

$$L^{\mathrm{T}} L\mathbf{u} = \lambda\mathbf{u}, \quad (2)$$

Each value of the vector $\mathbf{u}$ corresponds to the membership degree of a keyword belonging to the cluster. We set up the threshold to determine which keywords belong to clusters. After that, we remove the value of these keywords from L and repeat the equation (1) until no keywords satisfy the threshold.

2.3 Semantic clustering

BERT [10, 11] is one of the Transformer-based methods to compute sentence embeddings. The target of this study focuses on the word. We used the pretrained model "VoVanPhuc/sup-SimCSE-VietNamese-phobert-base" as BERT model to create embeddings.

The advantage of BERT lies in context-awareness. In other words, we need to provide the context of the keyword rather than just the input with only keywords. Meanwhile, keywords extracted from TF-IDF are representative of documents. Therefore, we input documents that satisfy the TF-IDF threshold into the BERT model instead of keywords. Specifically, we first only keep documents where the keyword satisfies the TF-IDF threshold. Then, we feed these documents into the BERT model. Finally, we average these vectors to get embeddings for the keyword. For efficiently clustering, we apply UMAP [17] to reduce the dimensionality of embeddings.

Finally, embeddings are clustered by Affinity Propagation (AP) [12]. AP performs clustering through a similarity matrix and sets a preference score to determine whether a point is an exemplar (center). In this study, we set the preference score as the median value of the similarity matrix. With this setup, AP can automatically determine the number of clusters (topics) without predetermining.

3 Experiments

3.1 Data and Ground Truth

In this study, we collect 9813 Vietnamese newspapers from the Binhvq News Corpus [13] with the query "Tết" ("Tet holiday" in English) from January 25 to February 22, 2016 (corresponding to 14 days before and after the first day of Tet - the most important day). These documents include 15 topics based on the category of dataset and manual adjustment. We use humans' interest from Google Trends as trend features.

To evaluate the trend and semantic clustering task, we define the ground truth as the procedure in Fig.2. We first extract keywords by TF-IDF with a threshold (in this study, the threshold is 0.4). Next, the trend of each keyword (crawled from Google Trends) is defined into one of three trend types (early, during, and after Tet holiday) based on the peaks of trends. The Tet holiday occurs from February 8 to February 18, 2016. Therefore, we define trend with a sharp peak that falls into this range of time, as during Tet Holiday. Meanwhile, the trend has clear peaks before and after these days, corresponding to early and after the festival. From each trend, the topic of keywords is clustered by using the document topics.

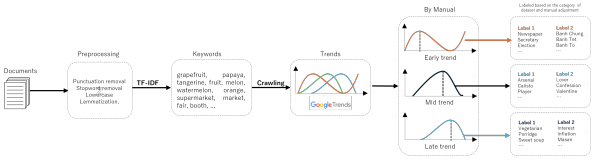

Figure 2 The procedure of defining the ground truth.

3.2 Experimental results

For Novelty 1, we conduct a comparision in topic coherence between our method and GLIPCA. For Novelty 2, we compare the trend representation of our methods and BERTopic. For Novelty 3, we use F_{score} [18], AMI [19], and ARI [20] to measure the accuracy of the topic extraction task and compare with the combination method between semantic clustering and other trend clustering methods that could not formulate the correlation relationship. The description of each comparative methods are listed in Table. 1 We set $\rho = 0.02$ in (1) and applied a moving average filter to smooth the trend feature with length of 2 because this setup gains the best accuracy with ground truth. The result is shown in Fig.3. Our method extracted three main trends corresponding to the red, blue and the green curves. To analyze these trends, the red curve indicates that it peaks around February 5th. This peak is equivalent to before the Tet holiday. Before the Tet holiday, it can be shown that extracted topics related to food, clothing, plants, politics, etc. This is suitable for the context in Vietnam. Because before the festival, people tend to spend a lot of money on new clothes (keywords: coat, shoes, etc) and Tet foods (keywords: candy, candied fruit, etc). In addition, displaying trees and flowers (keywords: bonsai, cherry, etc) is also a prominent feature of Tet. This is also the occasion for the government (keywords: communist party, congress, etc) to conduct annual summaries and have activities

Method	Description
k-means+ Modified BERT +AP	The trend clustering method is k-means with dynamic time warping distance. From each trend, keywords were converted into embeddings by computing the average of BERT embeddings of documents that contain keywords satisfying the TF-IDF threshold. Finally, these embeddings were clustered by AP.
k-shape+ Modified BERT+AP	The trend clustering method is k-shape. From each trend, keywords were converted into embeddings by computing the average of BERT embeddings of documents that contain keywords satisfying the TF-IDF threshold.. Finally, these embeddings were clustered by AP.
GLIPCA	Keywords are clustered into trends via GLIPCA.
BERTopic	Documents are clustered via BERT and HDBSCAN. From each cluster, a collection of keywords is extracted by c-TF-IDF.
GLIPCA+BERT+AP	The trend clustering method is GLIPCA. From each trend, keywords were converted into word embeddings by inputting the keywords into BERT. Finally, these embeddings were clustered by AP.
GLIPCA+Modified BERT+AP	This is our proposed method. The trend clustering method is GLIPCA. From each trend, keywords were converted into embeddings by computing the average of BERT embeddings of documents that contain keywords satisfying the TF-IDF threshold. Finally, these embeddings were clustered by AP.

Table 1 Description of comparative methods. The proposed methods are shown in bold.

to visit people to celebrate Tet. While, the blue curve indicates during Tet holiday, and the green curve is after Tet holiday. During the festival season, cultural activities take place, such as giving lucky money (keywords: luck money, gift, etc) or going to the pagoda (keywords: historical site, pagoda, etc). In addition, because Tet coincides with Valentine's Day, there is also interest in love topics (keywords: lover, couple, etc). After Tet, the focus shift into daily activities (keywords: school, staff, etc).

To demonstrate that our method improves the interpretation of trends, we conduct a comparison between the trend representation by each method in Fig.4. It can be observed that the trend representation of BERTopic is dense. The reason is that BERTopic only counts the frequency of topics over time. Therefore, the trends of topics are represented discretely and thus, it is difficult to analyze the topics in the stages of the event.

To evaluate the improvement of a combination of trend features and semantic features, we compare our method with methods using only semantic features (BERTopic) and only trend

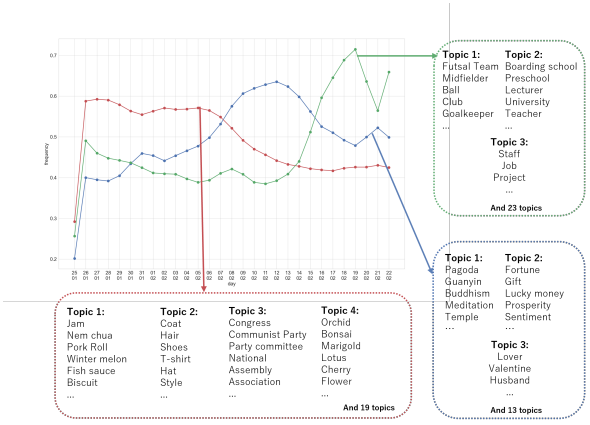

Figure 3 The trend and topic extracted from the Vietnamese newspaper from January 25th to February 22nd, 2016. The vertical axis is the frequency of humans' interest (scale from 0 to 1) and the horizontal axis is the date (first line: day, second line: month)

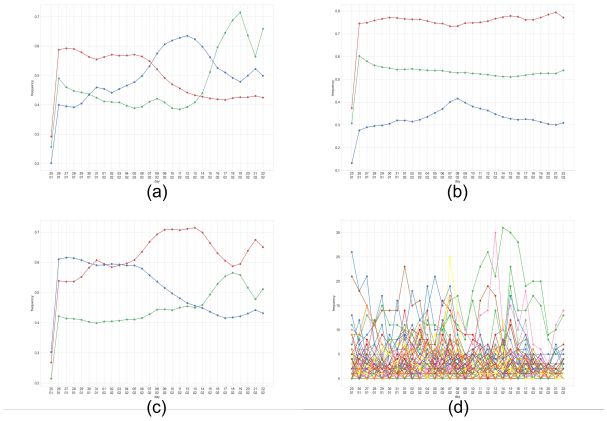

Figure 4 Trend representation comparison between the proposed method and comparative methods:
(a) GLIPCA, (b) k-means, (c) k-shape, (d) BERTopic

features (GLIPCA). To validate, we use topic coherence and topic diversity. Topic coherence measures the degree of semantic similarity between keywords within the same topic. We use C_{BERT} [21] by calculating the average cosine similarity of the embedding of keywords in a topic. Meanwhile, topic diversity (TD) measures the degree of lexical difference between topics. To evaluate the diversity of topics, we compute the percentage of unique keywords among all keywords from all topics. For fair evaluation, we choose the top 20 keywords by TF-IDF for only using the GLIPCA method, GLIPCA+Modified BERT+AP method, and c-TF-IDF for BERTopic.

Model	C_{BERT}	TD
BERTopic	0.450	0.522
GLIPCA	0.400	1.000
GLIPCA+Modified BERT+AP	**0.454**	**1.000**

Table 2 Comparative result between proposed method and GLIPCA, BERTopic in topic coherence

The experimental result demonstrates that a combination of trend and semantic features achieves a topic coherence score 13.515% higher than only trend features. Moreover, compared

with BERTopic, the topic coherence and topic diversity of our method outperform with C_{BERT} = 0.454 and TD = 1.000. From that, we can conclude that the introduction of semantic clustering into the trend clustering model (GLIPCA) improves the topic coherence of the model.

In addition, the strength of our method lies not only in its ability to represent trends but also in its ability to eliminate indirect (noise) correlations between topics, which often exist with information on the Internet, by using GLIPCA. We use AMI, ARI, and F_{score} to evaluate the quality of the trend and semantic clustering. For fair evaluation, we applied the moving average filter with a length of 2, and the number of clusters is 3 for k-means and k-shape. The result in the Table.3 shows that our methods achieve the highest AMI, F_{score} and second highest ARI among the compared methods. Besides that, the result also shows that inputting the documents for embeddings (Modified BERT) is better than inputting only keywords (BERT).

Model	F_{score}	AMI	ARI
k-means+ Modified BERT +AP	0.406	0.161	0.047
k-shape+ Modified BERT +AP	0.486	0.237	**0.094**
GLIPCA+ BERT +AP	0.525	0.110	0.025
GLIPCA+ Modified BERT +AP	**0.585**	**0.245**	0.087

Table 3 Comparative result on different types of trend and semantic clustering methods

In addition, in trend analysis, we believe that sharp peaks will show the trend more clearly and easily understood by users. Fig. 4 shows that GLIPCA makes the trend easier to distinguish events by removing trends that are not sharp peaks. This result reinforces that the generated topics have higher accuracy by removing indirect correlations than traditional methods.

4 Conclusion

By a combination of trend and semantic feature, our method succeeds in trend-topic extraction that enables exploring the public interest in social events. In experimental results, we have clarified three main points: First, our method allows generating topic trends more intuitively and understandably. Second, by combining trend features and semantic features, our method helps improve the topic coherence of topics. Finally, removing indirect correlations shown to be effective in improving the accuracy and intuitiveness by retaining only sharp peaks.

The result also provides an idea of a quantitative approach in social research in psychology, sociology, politics, etc. Moreover, it is possible to develop information retrieval because of the interpretation of topics.

Acknowledgement

This research is partially supported by JSPS KAKENHI Grant Numbers JP23K28193 and SPRIX Inc.

References

[1] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Information Systems*, vol. 112, p. 102131, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437922001090

[2] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[3] R. Harakawa and M. Iwahashi, "Ranking of importance measures of tweet communities: Application to keyword extraction from COVID-19 tweets in Japan," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 1030–1041, 2021.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, Mar. 2003.

[5] R. Harakawa, T. Ito, and M. Iwahashi, "Trend clustering from covid-19 tweets using graphical lasso-guided iterative principal component analysis," *Scientific Reports*, vol. 12, no. 1, p. 5709, 2022.

[6] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, pp. 129–136, 1982.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 5. University of California press, 1967, pp. 281–298.

[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[9] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015, pp. 1855–1870.

[10] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1037–1042.

[11] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[12] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[13] Binhvq, "Github - binhvq/news-corpus: Corpus tiếng việt." [Online]. Available: https://github.com/binhvq/news-corpus

[14] F. Martin and M. Johnson, "More efficient topic modelling through a noun only approach," in *Proceedings of the Australasian Language Technology Association Workshop 2015*, B. Hachey and K. Webster, Eds., Parramatta, Australia, Dec. 2015, pp. 111–115. [Online]. Available: https://aclanthology.org/U15-1013/

[15] G. Salton and M. McGill, "Introduction to modern information retrieval," 1983. [Online]. Available: https://api.semanticscholar.org/CorpusID:43685115

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[17] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020. [Online]. Available: https://arxiv.org/abs/1802.03426

[18] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information Retrieval*, vol. 12, no. 4, p. 461–486, 2008.

[19] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, no. 95, p. 2837–2854, 2010.

[20] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, p. 193–218, Dec. 1985.

[21] D. Angelov and D. Inkpen, "Topic modeling: Contextual token embeddings are all you need," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.