Journal of Visual Language and Computing

journal homepage: www.ksiresearch.org/jvlc/

A Topic Lifecycle Trend Prediction Algorithm on Facebook

Chen Luo[a,*], Jun Shi[a]

[a]CyberAray Network Technology Co.,Ltd, China

ARTICLE INFO

Article History:
Submitted 5.6.2023
Revised 7.18.2023
Second Revision 7.28.2023
Accepted 7.30.2023

Keywords:
Opinions
Facebook
Hot topics
Trend prediction
Similarity
Lifecycle

ABSTRACT

Recently, social media has been widely used for people to discuss public opinions and share their views. Internet public opinions have attracted a lot of attention from the government, enterprises and the general public. How to properly analyze, utilize and guide these online opinions is an extremely important issue that the world is currently faced with, and the prediction of topic lifecycle trends is the key to solving this problem. This paper proposes a topic lifecycle trend prediction algorithm based on Facebook data. The algorithm takes into account the similarity between new topics and historical topics in terms of lifecycle curves and the similarity in terms of text content, finds a curve that can best represent the future lifecycle trend of new topics, and then effectively predicts the trend of new topics. It is helpful and meaningful to use this method in the early warnings and predictions of online public opinions and hot topics.

1. Introduction

Nowadays, social media has become an indispensable part of people's daily lives. According to the Digital 2020 July Global Statshot report released by We Are Social and Hootsuite [1], we know that there are now 3.96 billion social media users worldwide, accounting for about 51% of the world's total population. Therefore, social media is one of the most important ways to propagate and spread information. In those various social media platforms, there are many hot topics generated every day. These hot topics are significant carriers which contain focused information people pay attention to, and they are directly related to the size of the social influence triggered by the events. It is crucial to make good use of the information in hot topics. On one hand, the government can monitor and analyze hot topics to understand the trend of online public opinions and take corresponding measures in time, which is conducive to maintaining the long-term stability of society. On the other hand, enterprises can understand the needs of users through relevant hot topics to make business plans such as personalized marketing to some users. As a result, putting forward a method that can predict the trend and analyze the lifecycle of topics is of great importance.

Facebook, as the world's most popular social media platform with over 2.6 billion monthly active users, has a plenty of topic data. In this paper, we use posts information from Facebook as our datasets, which include date, time, content and some other useful information. We first extract daily hot topics from Facebook daily posts, and then we use Jaccard Similarity algorithm to calculate the similarities between daily hot topics and posts from other days by comparing their keywords in order to find those posts which are related to hot topics. Based on that, we use the number of relevant posts as the value of y-axis, dates as the value of x-axis to plot the lifecycle curve of each hot topic. Topics with similar lifecycle curves are merged into one cluster, which means all topics under the same cluster have similar trends. After that, we extract a centroid curve for each cluster to stand for the trend of the cluster. When a new topic comes, Dynamic Time Warping (DTW) algorithm is used to compute similarities between curves to find curves from all clusters that are most similar to the curve of new topic. From those topics which are similar in curve, we check if their contents are similar to the new topic as well. Considering similarity both in curve and in content, we

*Corresponding author
Email address: linc950412@gmail.com
ORCID: 0000-0002-5747-742X

DOI reference number: 10.18293/JVLC2023-N1-30

can find one curve that best represents the future lifecycle trend of a new topic.

On the basis of description above, this paper proposes a topic lifecycle trend prediction algorithm based on Facebook dataset. There is no doubt that we encountered many difficulties in this process, such as finding posts that are related to hot topics, complicate data cleaning, reducing the time complexity of algorithm operation to increase efficiency and so on. With the efforts of keeping trying and optimizing, we finally propose this method. The main contribution of this paper can be summarized as follows:

- We use the K-Shape algorithm to cluster time series data, making topics with similar lifecycle curves into one cluster, which allows us to effectively observe and analyze the characteristics of different types of topic lifecycle, and make effective trend predictions for new topics that have similar curve characteristics to historical ones.
- We use the DTW algorithm to calculate the similarity between new topics and historical topics on the curve, solving the problem that ordinary Euclidean distance cannot compare the similarity of two unaligned or unequal length sequences.
- Considering the similarity of topics both in lifecycle curve and in text content, we propose a method to predict topic lifecycle trend.

The rest of the paper is discussed as the following order, Chapter 2 introduces the related work. In Chapter 3, we present the topic lifecycle trend prediction method. Chapter 4 shows the experiments and evaluation. At last, conclusion and future work is discussed in Chapter 5.

2. Related Work

In 2007, Jiang, Yue [2] made a study showing that early lifecycle data can be used to predict the fault-prone modules in a project. In 2012, Shota Ishikawa [3] designed a system detecting hot topics during a certain period of time and a method was proposed to reduce the variation of posted words related to the same topic, which provides a great contribution to AI services. In the same year, Rong Lu and Qing Yang [4] defined a new concept as trend momentum, which are used to predict the trend of news topics. Juanjuan Zhao [5] developed a model of short-term trend prediction of topics based on Sina Weibo dataset while the accuracy still needs to be improved when the trend of topic changes too frequently in 2014. More recently in 2018, Abuhay [6] used NMF topic modeling method to find topics and implemented ARIMA to forecast the trend of research topics. Chaoyang Chen and Zhitao Wang [7] proposed a correlated neural influence model, which can predict the trending research topics among the research evolution of mutually influenced conferences in the same year. In 2021, Yumei Liu and Shuai Zhang [8] researched the use of blockchain technology in the financial field, utilized various kinds of methods like co-word analysis and bi-cluster algorithm to explore hot topics and predict the future development trend. In 2022, a scientific research topic trend prediction model based on multi-LSTM and Graph Convolutional Network was proposed by Mingying Xu and Junping Du [9]. Compared to other baseline models, its experiment results showed an improvement on the prediction precision.

3. Topic Lifecycle Trend Prediction Method

Our goal is to build a topic lifecycle trend prediction algorithm model. Before that there are some necessary work to be done first, including hot topic extraction, finding topic-related posts, drawing historical topic lifecycle graphs, clustering and curve fitting based on lifecycle curve shapes. After all these tasks are completed, we calculate the shape similarity between topic lifecycle curves by using DTW algorithm. In the meantime, we calculate text content similarity based on keyword matching. Considering both shape similarity and text content similarity, we propose a topic lifecycle prediction method to predict the future lifecycle of new topics. The detailed flowchart is shown in Fig 1.

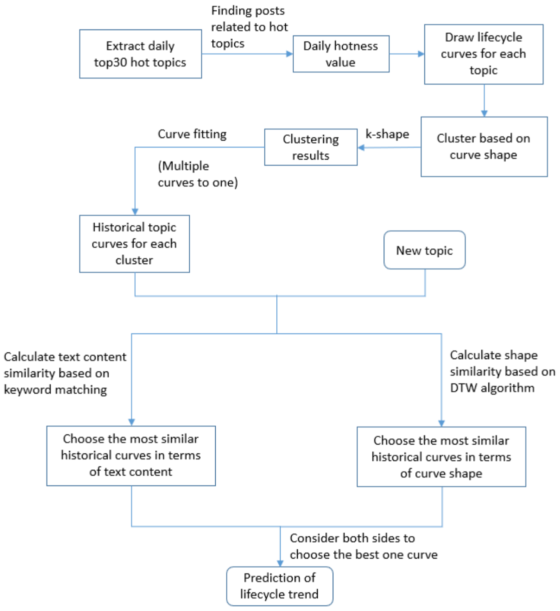

Extract daily top30 hot topics
Finding posts related to hot topics
Daily hotness value
Draw lifecycle curves for each topic
Cluster based on curve shape
k-shape
Clustering results
Curve fitting
(Multiple curves to one)
Historical topic curves for each cluster
New topic
Calculate text content similarity based on keyword matching
Calculate shape similarity based on DTW algorithm
Choose the most similar historical curves in terms of text content
Choose the most similar historical curves in terms of curve shape
Consider both sides to choose the best one curve
Prediction of lifecycle trend

Figure 1: Flow chart of topic trend prediction method.

3.1 Hot Topic Extraction

In this paper, the data we used came from crawling the Facebook platform. We crawled some of the posts' information of the Facebook platform from May 2020 to April 2021 as needed, a total of 100,535,793 posts with non-empty content.

Since there are too many post contents in different languages, we cannot analyze all of them. Then we intend to study the posts' information of only two languages this time, English and Chinese, by considering the number of language users, language popularity and some other factors. Thus, the first step is to identify Chinese and English posts, and then we perform tokenization. For Chinese lexical analysis, we use the tool called HanLP, and for foreign language lexical analysis, we use spaCy. Both HanLP and spaCy are commonly used natural language processing tools. Subsequently, keyword extraction is performed. A combination of lexical analysis, entity recognition, TF-IDF [10] and TextRank [11] algorithms are used to extract keywords, which can be classified by category as people, places, organizations, etc. Then the single pass algorithm is used to cluster the sentences in the post content based on keyword similarity, and the sentences with similar keywords are clustered into one class, that is, into one topic. Next, we need to give each topic a topic description to stand for the content of this topic. By achieving that, we extract the keywords of this topic, calculate the similarity between the keywords of the topic and the keywords of each sentence in the topic in order, and select the sentence with the highest similarity score and the shortest length as the topic description of this topic. Besides, we need to know each topic's hotness to find hot topics. To reach this goal, the number of posts of each topic is taken as the topic's hotness. We choose the daily Top N topics with highest hotness value as the main topics of this study. Due to the presence of some advertising information in the extracted hot topics, which are useless, it is experimentally concluded that when N=30 is chosen (N is not unique), a sufficient number of valid topics can be guaranteed. In this case, we extract the daily Top 30 hot topics for this study, and there are 9900 topics in total for eleven months. Because there is a possibility that a topic may be a hot topic for several days in a row, we are supposed to de-duplicate these 9900 topics and finally get 8359 unduplicated topics, which are used as our historical topic library (including curves and text contents) for this study. These topics are like “President Donald Trump announced that he and first lady Melania Trump has tested positive for COVID-19”, “Joe Biden just overtook Donald Trump in Pennsylvania, where he’s now leading by 5,594 votes”.

3.2 Finding Topic-Related Posts

In In order to study the lifecycle curve of a topic, we need to know how hot the topic is on a daily basis. Thus, we need an algorithm to find the statistics of the number of posts a topic has on a daily basis, as the daily hotness of the topic.

By achieving this goal, we design a Topic-Finding-Posts algorithm. The algorithm performs keyword extraction from a library of posts to be searched to obtain keywords for each post, and then it calculates the Jaccard similarity coefficient score between the set of post keywords and the set of topic keywords to see if the post is similar to the topic or not. The higher the score, the more similar the two sets. Finally, it outputs the posts related to the topic.

Due to the sheer volume of computing and the limitations of machine, we are unable to find posts related to a topic for an entire year. Given that hot topics are generally not hotter than three months, we set the lifecycle to three months, the month in which the topic is located, the month before and the month after. For example, if a hot topic is on July 3, 2020, we would look for posts related to the topic in June, July and August of 2020, which means it requires us to calculate the Jaccard similarity coefficient score between the set of post keywords in these three months and the set of topic keywords to find topic-related posts. The flow chart of this algorithm is shown in Fig. 2.

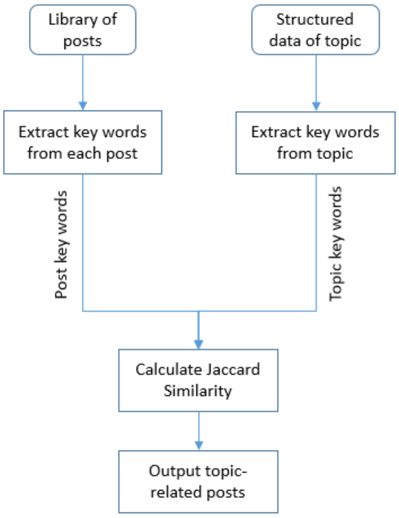

Library of posts
Structured data of topic
Extract key words from each post
Extract key words from topic
Post key words
Topic key words
Calculate Jaccard Similarity
Output topic-related posts

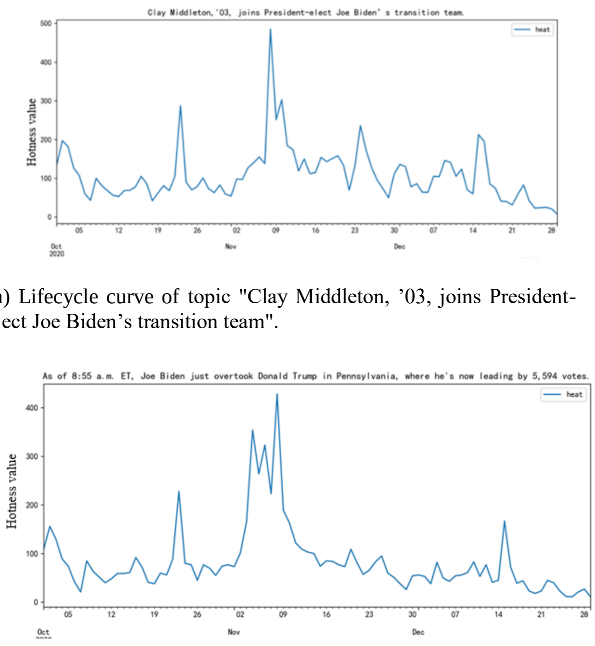
Figure 2: Flow chart of Topic-Finding-Posts algorithm.

3.3 Drawing Historical Topic Lifecycle Graphs

This step is to draw lifecycle graphs of all historical topics to build the historical topic library needed for this study. According to Part 3.1, we know that there are a total of 8359 unique topics. The Topic-Finding-Posts algorithm of Part 3.2 is used to find posts related to these topics over a three-month period and the number is counted as the topic's hotness value. Using the topic's hotness value as the y-axis, the three-month span of time as the x-axis, and the topic description as the title, the lifecycle graphs of these topics are plotted and saved as a historical topic graph library, representing the lifecycle trends of hot topics that emerged during these eleven months.

3.4 Topic Lifecycle Curve Shape Clustering Based on K-Shape

Based on the results of Part 3.3, we can obtain lifecycle graphs of thousands of historical topics. Since several hot topics appearing at the same time may be discussing the same thing, from this perspective they are actually one topic. For example, topic "Clay Middleton, ’03, joins President-elect Joe Biden’s transition team" and topic "Joe Biden just overtook Donald Trump in Pennsylvania, where he’s now leading by 5,594 votes" appeared in November 2020 are both about the US election, and they are similar in terms of their lifecycle curves. These two topics’ lifecycle curves are shown in Fig 3.



(a) Lifecycle curve of topic "Clay Middleton, ’03, joins President-elect Joe Biden’s transition team".

(b) Lifecycle curve of topic "Joe Biden just overtook Donald Trump in Pennsylvania, where he’s now leading by 5,594 votes".

Figure 3: Examples of two topics that have similar lifecycle curves.

In response to this case, we decide to cluster topics with similar lifecycle curves into one class and form a curve that represents this class as the lifecycle curve for this class. There are several advantages of doing this. First of all, it reduces the amount of computation since we only take the curve that represents each cluster into account. Secondly, it may help reduce the errors caused by the Topic-Finding-Posts algorithm on the hotness value of the topic lifecycle graph. Thirdly, it can eliminate some noise. For example, some oddly shaped graphs that appear only once are not representative, indicating that they are not common hot topics and will most likely not appear again in the future, which is not helpful for prediction, and these outlier topics can be found and discarded through clustering.

In this paper, K-Shape algorithm [12] is used for clustering, which is a clustering algorithm specifically for time series data and is concerned with similarity of shape. We use the tslearn package for clustering, which requires that the lengths of the different sequences should be the same. Thus, we cluster the curves by month, which ensures that the time series lengths of the lifecycle curves of topics in the same month are the same. In addition, we need to do feature scaling to bring all hotness values to the same magnitudes. To do this, we standardize the time series data by using z-normalization. Then, we use the normalized dataset to perform K-Shape clustering, where similar curves are clustered in one class and output centroid curves representing the lifecycle curves in this class. For example, the above lifecycle curves in Fig 3 are similar and can be clustered into one class, whose centroid curve is shown in Fig 4.

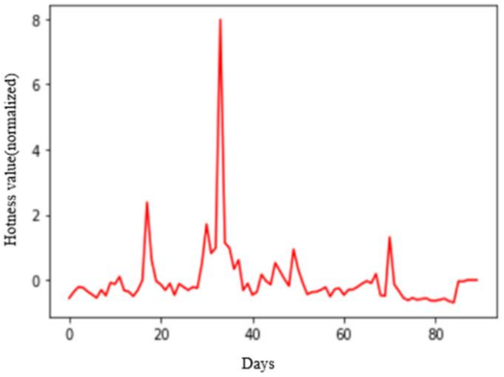


Figure 4: Example of centroid curve of a cluster (z-normalization).

All centroid curves after clustering are collected as a historical topic graph library. When a new topic emerges, the shape similarity between the new topic and the curves in historical topic graph library can be compared to predict the future trend of the new topic.

3.5 Calculation of Shape Similarity Based on DTW

DTW (Dynamic Time Warping) [13] is a dynamic programming algorithm that calculates the similarity of two time series [14], especially those of different lengths. When a new topic has been around for a while, we use this algorithm to calculate the shape similarity between the lifecycle curve of the new topic at that point and the centroid curves in the historical topic graph library in turn, and rank them to get some historical topic curves that are most similar to the current new topic. This gives an indication of some possible future trend directions for the new topic.

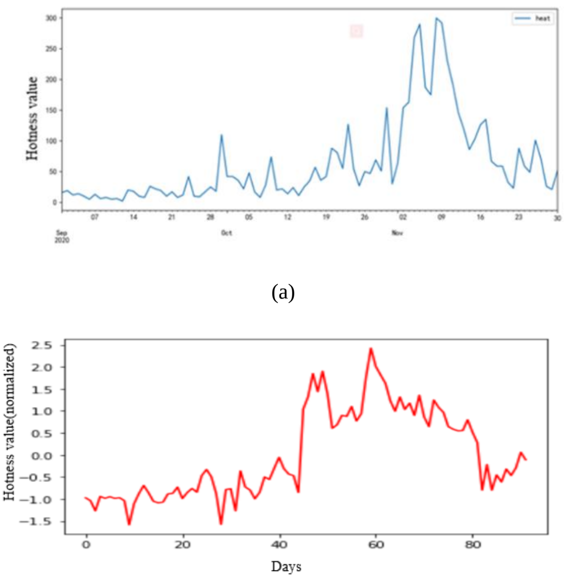
3.6 Calculation of Text Content Similarity Based on Keyword Matching

From Part 3.4, we can see that we have successfully clustered topics with similar lifecycle curves and obtained the centroid curves representing each category of topics. Next, we perform keyword extraction for each category of topics to learn the main textual content. The Jaccard similarity coefficient score between these topic keywords and the new topic keywords are calculated in turn and ranked to find the curves of historical topics that are most similar to the new topic in terms of textual content. This gives an indication of some possible future trend directions for new topics when considering similarity in text content. When a new topic has just come out and there is no obvious curve, text content similarity can be considered to use to solve the cold start problem, but only for reference, as similar text content does not mean that the trend is similar.

4. Experiments and Evaluation

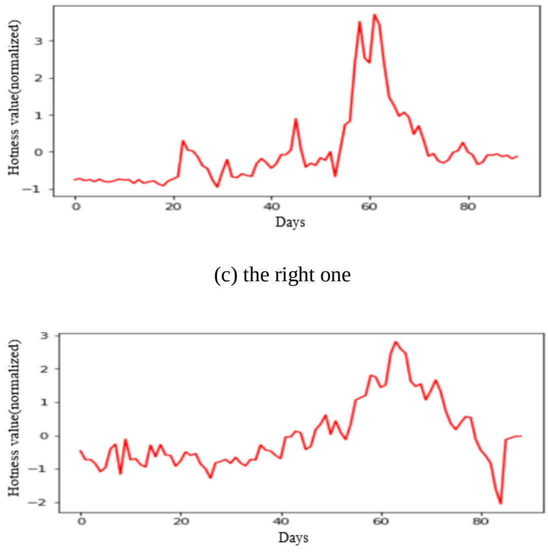
4.1 Topic Trend Prediction

Following the order in Chapter 3, we first extract the daily hot topics from the training set, then use the Topic-Finding-Posts algorithm to find the topic-related posts in a three-month cycle (here we set the similarity threshold of 0.45, if the Jaccard similarity coefficient score is greater than the threshold, the post is considered relevant to the topic). After that, we count the number of daily posts as the hotness to draw the lifecycle curve of the topic, and similar curves are clustered. Next, the DTW-based shape similarity calculation is performed on the clustered centroid curves and the test topic curves, and it is tested that when the similarity distance is less than 3.4, the trends of the two topics are similar. At the same time, the text content similarity calculation based on keyword matching is also performed (here the similarity threshold is also 0.45, and when the similarity is greater than 0.45, the text contents of the two topics are similar). Based on these experiments, historical topics that meet all the above conditions are considered similar to new topics, and their lifecycle trends can be used as a prediction of future trends of new topics. Let’s take an example (see Fig 5), when the test topic “Trump's dream is America's dream, Biden's dream is China's dream, Ivanka says” shows a lifecycle trend (see Fig 5. a, the part of the diagram within the dotted line), we can get three similar curves from historical topic graph library by only considering shape similarity (see Fig 5. b c d). Then we consider text content similarity, only one curve meets all the conditions at the same time, which is the second one of the three predictions (see Fig 5. c). Then we get our predicted trend curve for the test topic.



(a)

(b)



(c) the right one

(d)

Figure 5: Topic trend prediction experiments.

4.2 Clustering Effect Evaluation – Silhouette Coefficient Prediction

We use the Silhouette Coefficient as the effect evaluation index of this K-Shape clustering. The Silhouette Coefficient is a useful metric for evaluating clustering performance. It is computed by using mean distance between data points in the same cluster (cohesion) compared to the mean distance between data points in other clusters (separation) [15]. The calculated score ranges from -1.0 to 1.0. The higher the score, the better the clustering effect. To make the score computable, there have to be at least two clusters.

Assume that the data have been clustered into *k* classes. For data point $x(i) \in K$ (*K* is the cluster containing all the data points $x(i)$), $a_{x(i)}$ is the mean distance between $x(i)$ and every data point in the cluster *K*, $b_{x(i)}$ is the minimum mean distance between $x(i)$ and every data point in other clusters that is not a member of *K*. The calculation [16] of the Silhouette Coefficient of $x(i)$, the Silhouette Coefficient of each cluster, and the Silhouette Coefficient of all clusters can be shown as in (1), (2), and (3), respectively.

$$S_{(x_i)} = \frac{b_{x(i)} - a_{x(i)}}{max(a_{(x(i)}, b_{x(i)})} \tag{1}$$

where

$x(i)$ = data point in the cluster, $i = 1, 2, 3, \ldots, n$,

$a_{x(i)}$ = the mean distance between $x(i)$ and every data point in the cluster *K*, and

$b_{x(i)}$ = the minimum mean distance between $x(i)$ and every data point in other clusters that is not a member of *K*.

$$S_m = \frac{1}{n}\sum_{i=1}^{n} S_{(x_i)} \tag{2}$$

where

m = the number of the cluster, and

n = the number of data points in the same cluster.

$$S_{avg} = \frac{1}{k}\sum_{m=1}^{k} S_m \qquad (3)$$

where

k = number of all clusters.

We take the data of November 2020 as an example and cluster out ten classes as shown below (see Fig 6), where the red line represents the centroid curve of each class with a S_{avg} of 0.5162703634739744. The results tell us that the clustering works well. Data from other months are treated in the same way.

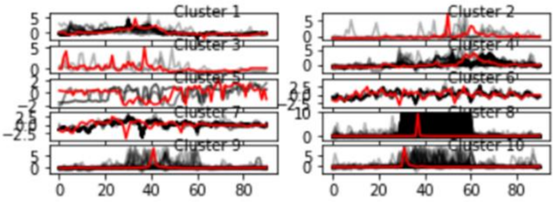


Figure 6: Clustering results of the data of Nov 2020(normalized).

4.3 Prediction Effect Evaluation

According to the clustering results of Part 4.2, we can briefly classify the type of clustering into short lifecycle class topics and long lifecycle class topics. The short ones refer to as above cluster 8, 9, 10, suddenly appeared to reach the peak and then the hotness immediately dropped and disappeared, mostly for some sudden events whose whole duration is just a few days. While the long ones are like cluster 5, 6, 7, whose hotness duration is long enough. They are usually serious events that need to be widely discussed. In this paper, we use the data of test set (120 topics in Jul 2021) as a collection of new topics. This method performs well on the test set, and the accuracy can reach 90%. For those test topics that are incorrectly predicted, we can see that there has not been a curve in the historical topic graph library similar to the curve of these test topics and there are no similar keywords in the text content either. In response to this situation, we add the lifecycle curves of these incorrectly predicted topics into the historical topic graph library to enrich it, so that these unmatched curves can be matched in the future.

5. Conclusion and Future Work

This paper proposes a topic lifecycle trend prediction algorithm based on Facebook data, which integrates shape similarity and text content similarity, and the results are more accurate than considering only shape similarity or only text content. The experimental results also show that the method is effective, but there are still some shortcomings that need to be improved.

1. For topics or lifecycle curves that have not appeared in history, it is impossible to make effective predictions, and the possible solution is to continuously expand the historical topic library to cover as many topics and curves as possible.

2. Because of the large amount of data, some topics have a very long lifecycle and the complete curve cannot be obtained. The algorithm can be optimized later to improve the running speed so that the complete lifecycle curve can be reached.

3. For shape similarity, the new topic needs to have a long enough lifecycle curve (to cover local or global features) to determine whether two topics are really similar by shape similarity calculation.

Acknowledgment

I would like to express my great appreciation and thanks to my supervisor Dr. Jun Shi, for her patient guidance and useful suggestions. I would also like to thank all my colleagues, for brainstorming with me and helping me solve the problems I encountered in this work. Finally, special thanks should be given to Shenzhen CyberAray Network Technology Co., Ltd, for supporting the data and resources I need.

References

[1] Kemp S. Digital 2020: July Global Statshot, Datareeportal.

[2] Jiang, Yue, Bojan Cukic, and Tim Menzies. "Fault prediction using early lifecycle data." The 18th IEEE International Symposium on Software Reliability (ISSRE'07). IEEE, 2007.

[3] Ishikawa, Shota, et al. "Hot topic detection in local areas using Twitter and Wikipedia." ARCS 2012. IEEE, 2012.

[4] Lu, Rong, and Qing Yang. "Trend analysis of news topics on twitter." International Journal of Machine Learning and Computing 2.3 (2012): 327.

[5] Zhao, Juanjuan, et al. "A short-term trend prediction model of topic over Sina Weibo dataset." Journal of Combinatorial Optimization 28 (2014): 613-625.

[6] Abuhay, Tesfamariam M., Yemisrach G. Nigatie, and Sergey V. Kovalchuk. "Towards predicting trend of scientific research topics using topic modeling." Procedia Computer Science 136 (2018): 304-310.

[7] Chen, Chengyao, et al. "Modeling scientific influence for research trending topic prediction." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[8] Liu, Yunmei, et al. "The sustainable development of financial topic detection and trend prediction by data mining." Sustainability 13.14 (2021): 7585.

[9] Xu, Mingying, et al. "A scientific research topic trend prediction model based on multi-LSTM and graph convolutional network." International Journal of Intelligent Systems 37.9 (2022): 6331-6353.

[10] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. Vol. 242. No. 1. 2003.

[11] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[12] Paparrizos, John, and Luis Gravano. "k-shape: Efficient and accurate clustering of time series." Proceedings of the 2015 ACM SIGMOD international conference on management of data. 2015.

[13] Myers, Cory, Lawrence Rabiner, and Aaron Rosenberg. "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition." IEEE Transactions on Acoustics, Speech, and Signal Processing 28.6 (1980): 623-635.

[14] Meesrikamolkul, Warissara, Vit Niennattrakul, and Chotirat Ann Ratanamahatana. "Shape-based clustering for time series data." Advances in Knowledge Discovery and Data Mining: 16th Pacific-Asia Conference, PAKDD 2012, Kuala Lumpur, Malaysia, May 29-June 1, 2012, Proceedings, Part I 16. Springer Berlin Heidelberg, 2012.

[15] Kaoungku, Nuntawut, et al. "The silhouette width criterion for clustering and association mining to select image features." International journal of machine learning and computing 8.1 (2018): 69-73.

[16] Aranganayagi, S., and Kuttiyannan Thangavel. "Clustering categorical data using silhouette coefficient as a relocating measure." International conference on computational intelligence and multimedia applications (ICCIMA 2007). Vol. 2. IEEE, 2007.