Research on Text Extraction and Analysis Based on Social Media

Xiaorui Wang

xiaoruiWang24@163.com

Artificial Intelligence College, Beijing Normal University, Beijing, China

Abstract. The onlineization of social networks is a typical feature of the big data era and one of the important reasons for the emergence of big data. Social media data contains rich information, which is carried by human language and contains a large amount of causal analysis and multi-dimensional description of events. It can provide powerful supplements to traditional information collection methods and has become a source of information for public opinion monitoring in recent years. This article takes Facebook as the main data source and uses Latent Semantic Analysis (LSA) method to automatically extract and express knowledge from large data corpora. Based on this, various methods are combined to conduct in-depth analysis of latent semantic features. The method established in this article can provide information supplementation for traditional public opinion monitoring.

Keywords: Text Extraction; Social media; semantic analysis.

1 Introduction

Social media is a platform based on Web 2.0 internet technology, with powerful information publishing and dissemination functions. Users rely on it to create and share information such as interests, hobbies, and status. It has now developed into the most popular internet application. The total user base of various social media platforms worldwide is 3.196 billion, accounting for approximately 40% of the world's total population. [1-2] Numerous social networking sites such as Facebook and Twitter, as well as domestic platforms such as Sina Weibo, Renren, Tencent, etc., have rapidly emerged. Facebook, founded in 2004, has over 1.3 billion registered users, equivalent to the second most populous country in the world; Twitter, released in 2006, also had over 600 million registered users; Tencent, a domestic company, has over 800 million active users; The latest data released by Sina shows that the number of registered users on Sina Weibo has exceeded 560 million. [3]

According to reports, 16% of US users stay online on Facebook, which exceeds 10% of people using traditional search engines such as Google. Undoubtedly, online social networks have become a bridge connecting the physical social world and the virtual cyberspace. [4]The interaction between online users and information, as well as the interaction between users, has left various footprints on social networks, directly promoting the arrival of the era of big data on the internet. Online social networks store a large amount of user information, social relationships between users, and interactions between users. These massive social data have enormous research value and also have broad application prospects in advertising, recommendation systems, and other fields.

PMIS 2024, March 15-17, Changsha, People's Republic of China

DOI 10.4108/eai.15-3-2024.2346529

Social networks provide users with a platform for interaction and dissemination of information, while also providing a data foundation for research on large-scale social networks. Social media has become an essential social tool in the daily lives of internet users. The massive user base not only promotes the diversified development of social media, but also accumulates a massive amount of user data, which contains valuable information such as personalized user characteristics and social network structures. [5-6] It plays an important guiding role in online recommendations, public opinion analysis, and other fields. Therefore, social media data mining has received widespread attention from the academic community. The existing research on social networks includes network structure topology analysis (such as ER model, small world model, Barabási-Albert model, etc.), network evolution analysis (such as network micro evolution), social relationship and influence analysis (such as link prediction, influence analysis, social bond relationship inference), and user behavior prediction. [7-8]

Since the emergence of a series of pre trained language representation models such as ELMo, GPT, BERT, etc., pre trained models have shown much better performance than traditional models in the vast majority of natural language processing tasks and have received increasing attention. This is one of the biggest breakthroughs in the field of NLP in recent years and the most important progress in the field of natural language processing. [9]

After BERT, many models have emerged that have extended it, including XLM and UDify that are pre trained across languages, models that are pre trained across modalities, ERNIE that integrates knowledge graphs, MASS that integrates language generation tasks such as seq2seq into BERT class models, UniLM, etc. [10]

2 Social media data analysis based on geographic location

Currently, people are gradually becoming interested in using geographic location data for analysis. This analysis has a wide range of applications in various fields, such as marketing, urban planning, and social science research. [11]

Firstly, by analyzing geographical location data, we can understand the user's behavioral habits and interests within a specific area. For example, by analyzing the photos and text content posted by users on social media, we can obtain information about their location, such as the most popular tourist destinations, the most popular restaurants, and cafes. This is very valuable information for the tourism and catering industry, as they can make targeted marketing decisions based on this data to attract more tourists and customers.

Secondly, location-based social media data analysis can also help us better understand the social structure and distribution of cities. By analyzing the social interactions of users in different regions, we can obtain the boundaries, population mobility, and group settlement patterns of different communities in the city. Such data is very valuable for urban planning and social science research, as they can be used to formulate more effective urban development and social policies.

In addition, location-based social media data analysis can also be used to provide personalized services and recommendations. By analyzing the preferences and preferences of users in specific locations, we can make recommendations based on their interests. For example, when users are in a certain area, we can recommend them to the best shopping places in that area, the latest movie release times, and local activities and social groups. This can provide a better user

experience and help users better adapt and integrate into the new social environment. This article uses social media to extract and analyze the latent semantics of location. [12]

3 Research methods

This article associates positional features with semantic features and uses Latent Semantic Analysis (LSA) method to automatically extract and express knowledge from large data corpora. On this basis, various methods of spatial analysis are combined to conduct in-depth analysis of positional latent semantic features. [13-15]

The mathematical foundation of latent semantic analysis is Singular Value Decomposition (SVD), which involves using implicit conceptual structures between words in a document to replace keywords and describe the document. In the processing of LSA, the first step is to select m words, then represent each document as a collection of these words, therefore, a corpus containing n documents can be represented as a matrix of $m \times n$, $A = [\alpha_{ij}]$, where α_{ij} represents the degree of co-occurrence between term i and document j, usually, TF-IDF (Term Frequency-Inverse Document Frequency) or Log entropy models are used for weighting processing. Performing singular value decomposition on a matrix yields:

$$A = U \sum V^T \tag{1}$$

In the formula: U is an orthogonal matrix of $m \times r$ composed of r eigenvectors of word AA^T, $V = (v_1,\ v_2,\ \cdots,\ v_r)$; V is an orthogonal matrix of $n \times r$ composed of r eigenvectors of document AA^T, $V = (v_1,\ v_2,\ \cdots,\ v_r)$; $\sum = diag(\sigma_1,\ \sigma_2,\ \cdots,\ \sigma_r)$ is a diagonal matrix, where $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r$ is the singular value of A; $U\Sigma$ is the term load of a document on the common principal components of words. Take the first k singular values and the first k columns of matrices U and V: U_k and V_k, then we can obtain a dimensionality reduction expression for matrix A:

$$A_k = U_k \Sigma_k V_k^T \tag{2}$$

The operation of dimensionality reduction not only removes the "noise" of the original data during the process of matrix rank reduction, but also improves the computational speed of subsequent analysis; More importantly, it abstracts the information on the co-occurrence relationship at a deeper level, which grasps the potential semantic relationship between words and documents. It maps the original matrix to a latent semantic space that is more in line with human cognitive relationships, thereby better mining the implicit connections between documents. However, it should be noted that the selection of k value is particularly important in the process of dimensionality reduction. An ideal solution is that the selected dimensions are comparable and consistent with the semantic features of word meanings, but it is difficult to obtain such prior knowledge in practical operations. Therefore, scholars have conducted various studies on this topic, but currently there is no universal or widely applied method, mainly because its selection depends on specific corpora and research purposes.

In fact, by organizing text in different ways, different implicit conceptual structures can be discovered through the LSA method. In the context of social media, Facebook text is mostly short text with less than 140 words, and LSA for mining long documents cannot be directly used. Therefore, it is necessary to aggregate the text in a certain form, such as according to time intervals, spatial ranges, or user attributes. This article chooses to reorganize Facebook text using

spatial location, which aggregates all Facebook texts within a specific spatial range according to certain spatial partitioning rules to form a document. The original term document matrix is transformed into a term position matrix. Next, the LSA method can be used to analyze the implicit information in spatial location, which is manifested in the following three aspects.

3.1 Location Theme Extraction

After performing singular value decomposition on the term position matrix, three parts are obtained, $U\Sigma$ represents the word load on the common principal components of a position, $u_{ij}\sigma_j$ represents the correlation coefficient between the principal component j and the term i. If a principal component has significant thematic features, the load of words will be significantly different from a uniform distribution and tend to concentrate under specific topic words, becoming the theme. Therefore, by calculating and sorting the $u_{ij}\sigma_j$ under the principal component j, the keyword with the highest correlation coefficient can be found, and the topic can be named accordingly. For example, after sorting, if the keywords of a certain principal component are "ticket", "scenic spot", "tourist", etc., it can be named "tourism theme". Additionally, based on the duality relationship of SVD, by calculating and sorting $v_{ij}\sigma_j$, obtain the significant heat zone of theme j in the research space. Meanwhile, due to the fact that the themes obtained by this method are orthogonal to each other, it also has a good effect on eliminating semantic intersections in the feature domain.

3.2 Location similarity analysis

The structured expression of implicit information in latent semantic space obtained after LSA processing. Because this expression not only reflects word frequency and co-occurrence relationships, but also represents the mining of deep information, eliminating the correlation between words, it helps to better measure positional similarity. Specifically, the existing location i in the study area, corresponding to the column i of the matrix A_k after dimensionality reduction of LSA, therefore, it is possible to directly calculate the vector similarity, such as cosine similarity, with any position within the research scope (any column of matrix A_k) in the latent semantic space. Sorting the similarity results in the distribution of regions closest to position i. In fact, for any text, it can be described as the expression Q on the vector space model (VSM) of the selected word after word segmentation, and then transformed into the same latent semantic space as the research area through equation (3). After standardization, the similarity matrix R with all positions can be calculated using equation (4).

$$Q' = Q^T U_k \Sigma_k^{-1} \tag{3}$$

$$R = (Q')^T A_k \tag{4}$$

Analyze the practical significance of the R matrix. If the input text Q is a description of the location, the obtained similarity can be understood as a measure of semantic similarity between the new location and the location within the study area; If the input text Q is a prior knowledge, the obtained similarity can be understood as a supervised annotation, and the magnitude of the similarity represents the membership degree of a certain region to that prior. For example, when Q is an educational text, it can be used to describe the membership relationship of a region to education in the latent semantic space. On this basis, further analysis of the local spatial autocorrelation of the similarity matrix R can obtain the heat zone of the functional area distribution. Specifically, the commonly used Local Moran'sI index is actually a special case of the Gamma index (equation (5)), used to describe the clustering of specific attributes in spatial

distribution, by defining the positional similarity w_{ij} between position i and position j within the neighborhood range, and attribute similarity a_{ij} to achieve.

$$\Gamma_i = \sum_i^n \sum_j^n w_{ij} a_{ij} \tag{5}$$

3.3 Positional clustering

In traditional positional clustering, positions are clustered based on similarity in feature space. For LSA, the feature space corresponding to position i is the column i of the A_k matrix after SVD and dimensionality reduction, or the load on the position theme (common principal component). Both definitions of feature spaces can achieve clustering of positions in latent semantic space, but the former focuses on data-driven grouping of positions, while the latter focuses on discovering implicit structures between positions. This article uses the former approach for clustering.

4 Results and analysis

4.1 Dataset construction

Selecting users who have posted on Facebook during the Maui Island fire in Hawaii, USA as the experimental subjects, the data is sourced from the Facebook dataset provided by the University of North Texas. The time span of blog posts is from August 10, 2023 to August 15, 2023, with a total of 690755 blog posts as the experimental dataset.

4.2 Data annotation and data preprocessing

When conducting data annotation, this study adopts the following general process.

(1) There are two annotators involved in the annotation process, and the annotation process strictly follows a unified standard. Each annotator annotates all experimental data and conducts cross validation after annotation to ensure accuracy and objectivity. For data with inconsistent annotations, the annotators determine the final label through discussion and analysis of multi-dimensional information in the data.

(2) The data preprocessing stage mainly converts textual data into numerical data as input data for subsequent feature extraction and classification models. This article uses Python programming language to remove symbols, combined with RANKSNL English stop word list to remove stop words, and finally vectorizes the text data.

4.3 Experimental result

Due to the significant impact of the dimension k in LSA on the analysis results, various methods were adopted to explore the k value in the experiment. Based on the research conclusions of Doxas et al., k=8, the results of the Profile Likelihood Test, k=50, and the number of singular values greater than 1, k=150, as candidate parameters, K-MEANS, WARD hierarchical clustering, and spectral clustering were performed respectively(Figure 1-3), and the contour coefficient was used as a measure of clustering results. It was found that in the KMEANS and WARD methods, the clustering effect of k=8 was always much higher than the latter two when the number of clusters was set to 2-20. Considering the applicability of spectral clustering only

when the number of categories was small, the result of k=8 was also significantly better than the latter two. Meanwhile, because the corpus being studied is Facebook check-in texts, mostly short texts with less than 140 words, even if aggregated according to spatial relationships, there is still more interference information compared to traditional long documents. Therefore, choosing to retain a higher dimension to explain the more information in the original data is not meaningful. Therefore, k=8 was ultimately chosen as the dimension for LSA analysis in this experiment.

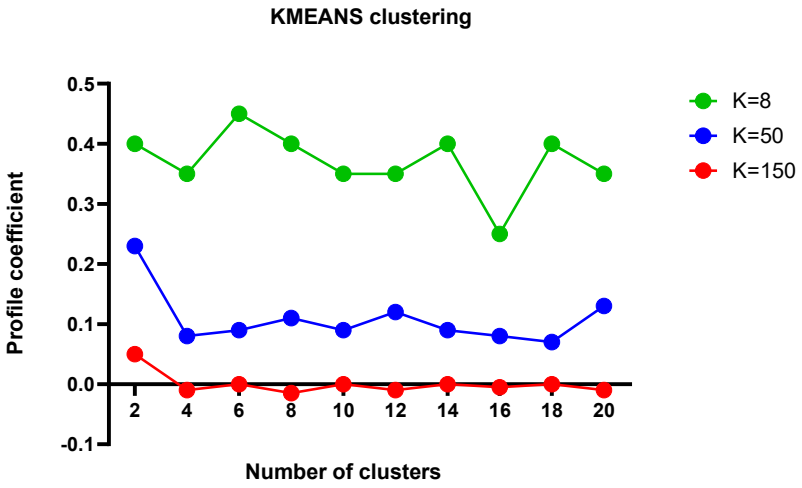

Figure 1. Silhouettes of clustering in different LSA dimensions(KMEANS).

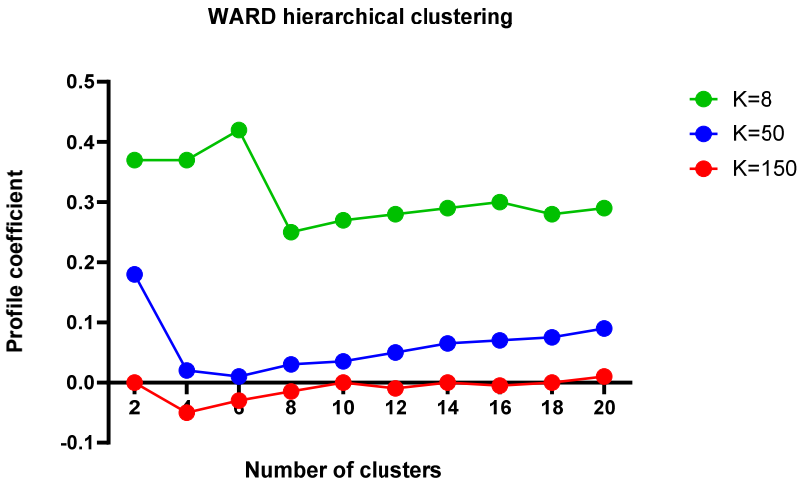

Figure 2. Silhouettes of clustering in different LSA dimensions(WARD).

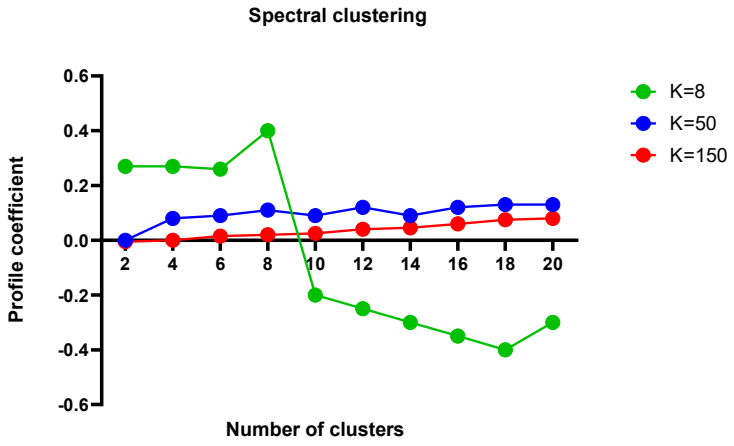

Figure 3. Silhouettes of clustering in different LSA dimensions(Spectral clustering).

Randomly divide the dataset into 80% training set and 20% testing set. The blog data for the experiment is 690755, and the effective data for noise removal is 385200. Maui Island in Hawaii is divided into 650 basic units of geographic data organization. The training was conducted using a 5-fold cross validation method, with each experiment repeated 10 times to prevent random effects. After training, call the model to predict the test set and compare it with the correct tag. The results showed that out of 77040 test data, 65500 were correctly judged for positional features, with an accuracy rate of over 85% (Table1).

Table 1. Numerical value.

Term	Numerical value
Blog data	690755
Effective data	385200
Test data	77040
Correctly judged for positional features	65500
Accuracy rate	85%

5 Conclusions

Geolocation based social media data analysis has important application value in many fields. By analyzing geographical location data, we can gain insights into user behavior, understand urban structure, and provide personalized services. However, we also need to recognize the challenges and privacy issues that exist and take corresponding measures to address them. Only with full respect for user privacy can we better utilize location-based social media data analysis to promote social development and progress.

Social networks have experienced rapid development in China, accumulating a large amount of user data and bringing enormous opportunities for deep mining and analysis of massive

heterogeneous social networks. Our country's basic research and technological accumulation in the fields of computational sociology, network science, data mining, databases, and machine learning will help us occupy the technological high ground in the era of big data and social networks, and improve the level of network information management and application.

However, there are still certain limitations to the research: (1) the representativeness of the data. The percentage of blog texts with geographic tags is small, the population is biased, and the spatial distribution is uneven. Therefore, although check-in texts can reflect residents' perception of location, they often tend to favor specific types of areas, such as entertainment, travel, tourism, etc. This also makes the extraction of latent semantic features for certain types, such as residential areas, insufficient in research. (2) In the study, only spatial latent semantic features were extracted and analyzed, without considering the temporal differentiation of latent semantics. However, in reality, the time series curve of blog posts and the topic content of different time periods have a significant contribution to the semantic characteristics of location. Therefore, improvements in these two aspects are also directions for future research.

References

[1] Asif, Muhammad, et al. "Sentiment analysis of extremism in social media from textual information." Telematics and Informatics 48 (2020): 101345.

[2] Heidari, Maryam, and James H. Jones. "Using bert to extract topic-independent sentiment features for social media bot detection." 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, 2020.

[3] Ali, Farman, et al. "Traffic accident detection and condition analysis based on social networking data." Accident Analysis & Prevention 151 (2021): 105973.

[4] Nemes, László, and Attila Kiss. "Social media sentiment analysis based on COVID-19." Journal of Information and Telecommunication 5.1 (2021): 1-15.

[5] Chauhan, Priyavrat, Nonita Sharma, and Geeta Sikka. "The emergence of social media data and sentiment analysis in election prediction." Journal of Ambient Intelligence and Humanized Computing 12 (2021): 2601-2627.

[6] Koh, **g Xuan, and Tau Ming Liew. "How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds." Journal of psychiatric research 145 (2022): 317-324.

[7] Han, Xuehua, et al. "Using social media to mine and analyze public opinion related to COVID-19 in China." International journal of environmental research and public health 17.8 (2020): 2788.

[8] Kumar, Sunil, Arpan Kumar Kar, and P. Vigneswara Ilavarasan. "Applications of text mining in services management: A systematic literature review." International Journal of Information Management Data Insights 1.1 (2021): 100008.

[9] Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi. "A BERT-based transfer learning approach for hate speech detection in online social media." Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8. Springer International Publishing, 2020.

[10] Abidin D Z, Nurmaini S, Malik R F, et al. A model of preprocessing for social media data extraction[C]//2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS). IEEE, 2019: 67-72.

[11] Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang. "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach." Multimedia tools and applications 80.8 (2021): 11765-11788.
[12] Balaji, T. K., Chandra Sekhara Rao Annavarapu, and Annushree Bablani. "Machine learning algorithms for social media analysis: A survey." Computer Science Review 40 (2021): 100395.
[13] Valanarasu, R. "Comparative analysis for personality prediction by digital footprints in social media." Journal of Information Technology and Digital World 3.2 (2021): 77-91.
[14] Saroj, Anita, and Sukomal Pal. "Use of social media in crisis management: A survey." International Journal of Disaster Risk Reduction 48 (2020): 101584.
[15] Ali, Farman, et al. "An intelligent healthcare monitoring framework using wearable sensors and social networking data." Future Generation Computer Systems 114 (2021): 23-43.