Research Article

Information Extraction and Spatial Pattern Analysis of Sports Events Based on Deep Learning

Quansheng Zheng

Physical Education Department, Guizhou University of Commerce, Guiyang 550014, China

Correspondence should be addressed to Quansheng Zheng; zhengqsh@yeah.net

Received 10 March 2022; Revised 7 April 2022; Accepted 11 April 2022; Published 22 April 2022

Academic Editor: Sheng Bin

Copyright © 2022 Quansheng Zheng. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The analysis of sports events information greatly promotes the formulation of scientific development strategy for sports events, the optimization of the spatial distribution of these events, and the popularization of sports culture. Most of the previous studies focused on the positive effects on sports events, but few quantified the effects. Besides, the broad environment of new media information communication was not considered. Therefore, this study carries out information extraction and spatial pattern analysis of sports events based on deep learning. Firstly, the texts of sports events information were classified by convolutional neural network (CNN), and the most valuable news or cases of sports events were screened. On this basis, a named entity recognition model was constructed to extract the most effective information from the data of sports events information. Next, a spatial pattern analysis approach was provided for the diffusion of sports event information flow. Finally, experiments were carried out to demonstrate the effectiveness of the proposed model and provide the spatial pattern analysis results on the diffusion of sports events information flow.

1. Introduction

Since the 2008 Beijing Olympiad, sports events have become an important part of social life in China. More and more people take part in sports, enjoy sports, and pay much attention to information of sports events [1–4]. All sorts of sports event information are delivered via information communication media, such as sports newspapers, sports television channels, sports websites, and sports new media [5–10]. The analysis of sports event information can disclose the public’s concern over different types of sports events, reflect the spatial pattern of sports event information, and guide the organization of such events [11–16]. The information extraction and spatial pattern analysis of sports events greatly promote the formulation of scientific development strategies for sports events, the optimization of the spatial distribution of sports events, and the dissemination of sports culture.

With the dawn of the age of shallow reading and the proliferation of smartphones, data visualization has evolved into an ideal way to present news. As large sports events become increasingly influential and eye-catching, visual communication wins the recognition of viewers thanks to its vivid display of data and news. Through a questionnaire survey, Xia [17] evaluated the influence of the propagation of large sports events against the backdrop of the big data. Facing the rapid growth of sports video data, it is critical to identify the information of interest. Chen and Du [18] analyzed the hidden Markov model, presented a semantic analysis method for sports videos, and clarified the analysis flow for the contents being played and not being played. Then, a few segments were selected for experiments. The experimental results show that the model achieved an accuracy of around 90%. Cannavò et al. [19] investigated the feasibility of the spatiotemporal position dataset collected from athletes and sports equipment during the games and utilized machine learning to automatically identify small sports events in datasets related to basketball. Choudhury and Breslin [20] adopted various approaches to test the named entities and important microevents in tweets published during the live broadcast of sports events, combined
linguistic features with background knowledge, and achieved high-precision detection on different datasets. Despite the fruitful research contents, most of the previous studies are too shallow and concentrated on the positive effects of sports events. However, very few scholars have quantified these effects. With the rapid development of information technology, the information communication method of the traditional method is gradually phased out by that of new media. In the context of new media, the Internet plays an important role in the acquisition of sports events information. This study targets the texts of sports events information on online media and mainly examines the information extraction and spatial pattern analysis of sports events. The main contents are as follows. Based on a convolutional neural network (CNN), Section 2 classifies the texts of sports events information and screens the most valuable news or cases of sports events. Next, a named entity recognition model was constructed to extract the most effective information from the data of sports events information. Section 3 provides a spatial pattern analysis approach for the diffusion of sports events information flow. Finally, experiments were carried out to demonstrate the effectiveness of the proposed model and provide the spatial pattern analysis results on the diffusion of sports events information flow.

2. Deep Learning-Based Information Extraction of Sports Events

The audience’s attention to sports events is mainly manifested as the rating of live broadcasts on traditional media and the attendance in the venue. With the development of the times and information technology, more and more people prefer watching sports events online. To increase the visibility of a sports event, the organizer must communicate the event information in the most effective way. The common communication media for sports events can be divided into three categories: interpersonal communication, organizational communication, and public communication. With the rapid development of information technology, the sports events communication can be achieved in the new media. This new communication model boasts low cost of loop playback, fast updates, large information volume, rich contents, strong interactivity, all-weather operation, and full coverage. Figure 1 shows the communication channels of sports events information.

This study focuses on the following communication features of sports events information, such as the diversity of communication subjects, the timeliness of communication, the fragmentation of information, the broadness of communication, and the fission propagation of information. As the most important transmission medium, new media has two basic functions, namely, receiving and publishing sports events information. The speed of information reception and publication directly bears on the visibility enhancement of sports events, which is sensitive to time. After logging onto new media software or apps on their mobile terminals, users can receive sports events information anywhere and anytime and post their reviews. This is the major advantage of the new media in the communication of sports events’ information. With the aid of the new media, it is possible to carry out live broadcasts of sports events and deliver sports events information to more people in a shorter time.

In this study, the sports events information gathered by crawlers is preprocessed through format unification, duplicate removal, and word segmentation. After pre-processing, the texts of sports events information were classified by CNN, and the most valuable news or cases of sports events were selected from the original data. Next, a named entity recognition model was constructed to extract the most effective information from the data of sports events information. The principle and improvement of our methodology, as well as experimental design, are detailed in the following sections.
2.1. Text Classification. The CNN was introduced to classify the texts of sports events information. In the network, the input layer is responsible for formatting the information text series of sports events and converting the texts into word vectors. Figure 2 illustrates the structure of the CNN. Through convolution, the convolutional layer performs the continued multiplication and summation of the array elements in the matrix of the sentences in the texts of sports events information:

\[ E = g(Q \ast C + \psi). \] (1)

Let \( Q \) be the initial weight matrix, \( \psi \) be the offset vector, and \( E \) be the feature matrix of the convolutional layer output. Capable of sparse activation and preventing exploding gradients, the rectified linear unit (ReLU) function can be adopted to map the convolutional results nonlinearly. ReLU has two advantages as the activation function: (1) the gradient of the function is a constant in the positive quadrant, which prevents the diffusion of gradients; (2) the sparse activity ensures that the gradient of the function is zero in the negative quadrant, such that the nodes will not be trained. Thus, we have

\[ \text{Relu}(a) = \begin{cases} a, & \text{if } a > 0, \\ 0, & \text{if } a \leq 0. \end{cases} \] (2)

The pooling operations of the CNN include maximum pooling and average pooling:

\[ UQ = \text{MP}[E + \psi], \]
\[ UQ = \text{AP}[E + \psi]. \] (3)

The output layer adopts softmax logistic regression. Let \( A_i \) be the \( i \)th element in input \( A \). Then, the softmax value of that element can be obtained by

\[ \text{SOF}_i = \frac{e^{a_i}}{\sum_j e^{a_j}}. \] (4)

Formula (4) shows that the softmax value is the proportion of the exponential value of the element in the sum of exponential values of all elements. If \( M \) texts of sports events information wait to be classified, an \( M \)-dimensional vector can be obtained through the softmax layer of the CNN. The \( i \)th value of the vector is the probability of a text of sports events information belonging to the \( i \)th class.

2.2. Named Entity Recognition. In the field of sports events, more than 80% of data are effective information extracted from the texts related to sports events news released on television, newspapers, radio, the Internet, and the new media. The information, including the time and location of each event, the athletes participating in the event, and the live broadcast of the game, greatly supports the judgment of the situation of major sports events, the promotion of sports culture, the feedbacks on event services, and the improvement of service quality. To recognize the entities in sports events, this study constructs a pretrained named entity recognition model embedded with a language representation module.

The model includes an embedded module, a feature extraction module, a conditional random field (CRF) module, and an adversarial training module. In the embedding module, the information text series of sports events, which is of the length \( m \), supports three vector representations: token embedding \( R_{TO}(q) \), segment embedding \( R_{SE}(q) \), and position embedding \( R_{PO}(q) \). The three representations share the same shape \((1, m, 768)\). \( R_{TO}(q) \), \( R_{SE}(q) \), and \( R_{PO}(q) \) represent the character vector, the sentence vector, and the position of the input series, respectively.
The outputs of the two BiLSTM models are spliced and introduced into a linear feedforward neural network layer:

$$f_p^{CRF} = Q(f_p^c \odot f_p^f) + \psi,$$

where $Q$ and $\psi$ are the weight and bias of the linear feedforward neural network layer, respectively.

The BiLSTM fails to consider the sequential dependence of text series. To obtain the final labels of the series, this study adds a standard CRF layer atop the named entity recognition model. Let $Q^{CRF}$ be the model parameter, $U_{b_{p-1}, bp}$ be the transfer probability matrix from label $b_{p-1}$ to label $b_p$ and $m$ be the length of the input sentence. Then, the definitions related to the score prediction can be given by

$$E(A, b) = \sum_{p=1}^{m} \left( v_p^{CRF} + U_{b_{p-1}, bp} \right),$$

$$v_p^{CRF} = Q^{CRF} f_p^{CRF}.$$

Let $\hat{b}$ be the optimal text series label. For the character vector $A = \{a_1, a_2, \ldots, a_m\}$ of the input text series, the probability of outputting $\hat{b}$ can be calculated by

$$GR(\hat{b}|A) = \frac{r^a(A, \hat{b})}{\sum_{b \in B_d} r^a(A, b)}.$$  

The CRF model is trained with the following loss function:

$$K_{CRF} = -\log GR(\hat{b}|A).$$

Inspired by the adversarial network, this study extracts the boundary information from the shared words in sports events information and generates the final features by the CNN for recognizing the effective information, aiming to optimize the named entity recognition task based on sports events’ information sharing. The hidden state of private features for the Chinese word segmentation task can be expressed as

$$f_p^{CWS} = \Omega^+(f_{p-1}^{CWS}, a_p).$$

To segment Chinese words and recognize named entities, the feature series generated by the BiLSTM responsible for Chinese word segmentation can be spliced with $f_p^{adv}$:

$$f_p^{adv} = f_p^{CWS} \odot f_p^{common}.$$  

Based on the CNN, the final features are shared by both tasks (Chinese word segmentation and named entity recognition); the discriminator of the adversarial training performs max pooling on the convoluted series using a convolutional operator:

$$f_p^{adv} = \max pooling(f_1^{adv}, f_2^{adv}, f_m^{adv}).$$

Finally, the probability of an actual task $l$ can be calculated by a fully connected layer:

$$f_p^{adv} = \tanh(Q(f_p^{adv}, f_{p-1}^{adv}, f_{p+1}^{adv}, f_{p-2}^{adv}, f_{p+2}^{adv})).$$

$$f^{adv} = \max pooling(f_1^{adv}, f_2^{adv}, f_m^{adv}).$$
3. Spatial Pattern Analysis on Diffusion of Sports Events’ Information Flow

The diffusion strength index of information flow can characterize the spatial diffusion strength of sports events information flows based on the new media:

\[
\alpha = \frac{P_i}{P_y}
\]

(17)

The above index measures the regional spatial diffusion strength of sports events information. Let \(P_i\) be the volume of information forwarded to the \(i\)th region based on the new media and \(P_y\) be the total volume of information forwarded based on the new media. The greater the \(\alpha_i\), the stronger the diffusion of sports events information flow towards the \(i\)th region.

The distribution law of regional size is often measured by primacy. There are three common indices of regional primacy: two-region index \(E_2\), four-region index \(E_4\), and 11-region index \(E_{11}\). Among them, \(E_2\) refers to the ratio of the largest subregion in the region to the second largest subregion:

\[
E_2 = \frac{T_1}{T_2}
\]

(18)

Let \(\varepsilon_1\) be a threshold. If \(E_2 > \varepsilon_1\), then the largest subregion has a strong monopoly, and the subregions cluster moderately in the region. \(E_4\) refers to the ratio of the largest subregion to the sum of the second to fourth largest subregions:

\[
E_4 = \frac{T_1}{(T_2 + T_3 + T_4)}
\]

(19)

\(E_{11}\) refers to the ratio of the largest subregion to the sum of the second to 11th largest subregions:

\[
E_{11} = \frac{2T_1}{(T_2 + T_3 + \ldots + T_{11})}
\]

(20)

Both \(E_4\) and \(E_{11}\) use \(\varepsilon_2\) as the threshold. If \(E_4\) or \(E_{11}\) is greater than \(\varepsilon_2\), then the subregions are over clustered in the region; if \(E_4\) or \(E_{11}\) is smaller than \(\varepsilon_2\), then the subregions cluster moderately in the region.

The rank-size rule can be applied to explore the distribution features of subregion sizes, according to the correlations between subregion sizes and size ranking. Let \(CO\) be the ranking of subregion size and \(US_{CO}\) be the subregion size. Then, the rank size of subregions can be characterized by \(US_1 \geq US_2 \geq \ldots \geq US_m\). Let \(RS_i\) and \(CO\) be the population and ranking of the \(i\)th subregion, respectively; \(\varsigma\) be the constant. Then, we have

\[
RS_iCO^{-\varsigma} = \varsigma \text{ or } RS_i = \varsigma CO^{-\varsigma}.
\]

(21)

Taking the logarithm of the two sides of the above formula,

\[
\log RS_i = \log \varsigma + \log CO.
\]

(22)
Figure 4: Relationship between the published volume of sports events information and time.

Figure 5: Relationship between the information diffusion level and number of regions.

Table 1: Rank-size regression on sports event information.

<table>
<thead>
<tr>
<th>Time</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression equation</td>
<td>$P_i = 353R_i^{1.374}$</td>
<td>$P_i = 125R_i^{1.250}$</td>
<td>$P_i = 18R_i^{1.718}$</td>
<td>$P_i = 502R_i^{1.258}$</td>
</tr>
<tr>
<td>$\sigma_{yi}$</td>
<td>1.285</td>
<td>1.086</td>
<td>1.174</td>
<td>1.326</td>
</tr>
<tr>
<td>$E_i$</td>
<td>0.728</td>
<td>0.925</td>
<td>0.718</td>
<td>0.682</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.958</td>
<td>0.926</td>
<td>0.847</td>
<td>0.926</td>
</tr>
</tbody>
</table>

|        |       |       |       |       |
| Regression equation | $P_i = 859R_i^{1.495}$ | $P_i = 148R_i^{1.075}$ | $P_i = 35R_i^{1.142}$ | $P_i = 1057R_i^{1.274}$ |
| $\sigma_{yi}$ | 1.485 | 1.062 | 1.148 | 1.317 |
| $E_i$    | 0.924 | 0.846 | 0.851 | 0.715 |
| $R^2$   | 0.915 | 0.947 | 0.936 | 0.918 |
The training accuracy and training loss of the CNN were observed (Figure 6). Initially, the loss curve dropped significantly, while the accuracy curve rose markedly. Later, both curves tended to be stable. Through the CNN training, the two indices changed in opposite directions. This meets the expectation of our model in training. Table 3 shows the text classification results of our model. It can be seen that the proposed text classifier performed well in classifying the texts related to sports events’ information.

Furthermore, the embedding method our model was compared experimentally with another embedding model: word2vec. The experimental results are recorded in Table 4, where 1, 2, and 3 stand for three different training models: the conventional model, the conventional model coupled with the CRF layer, and our model. It is learned that our embedding method is superior to word2vec embedding in named entity recognition, especially after the introduction of the CRF layer and the adversarial training. The superiority of our model demonstrates that the word boundary information of the texts obtained through adversarial training promotes the performance of named entity recognition and helps to solve the adaptation to different fields.

5. Conclusions

Based on deep learning, this study carries out information extraction and spatial pattern analysis of sports events based on deep learning. Initially, CNN was adopted to classify the texts of sports events information and to identify the most valuable news or cases of sports events. Next, the authors established a named entity recognition model to extract the effective information from the data of sports events information and provided a spatial pattern analysis approach for the diffusion of sports events information flow. Through experiments, the authors described the relationship between the information diffusion levels of different new media information on sports events and the number of regions and concluded that the relationship between information diffusion level and the number of regions obeyed a pyramid-shaped distribution. In addition, a log-log regression analysis was performed on the diffusion of sports events information on different new media. The analysis results show that the diffusion of sports events’ information on different new media satisfied the rank-size distribution. Finally, the accuracy and loss of network training were compared, and different embedding models were tested, which confirms the effectiveness of our model.

Many users acquire sports events’ information from new media, such as sports newspapers, sports television channels, sports websites, and sports news. Facing the huge amount of information about sports events, it is an arduous task to collect and analyze the basic data, and it is very difficult to study the information flow of sports events. Future research will further explore the information flow of sports events more comprehensively by expanding the size of the research dataset.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References


[18] Y. B. Chen and G. Du, “Sports events videos-based semantic content analysis and hidden Markov model application,” *Bio-