IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 7, 2020, accepted November 20, 2020, date of publication November 25, 2020,
date of current version December 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040408

SUDIR: An Approach of Sensing Urban Text Data From Internet Resources Based on Deep Learning

CHAORAN ZHOU[1], JIANPING ZHAO[1], AND CHENGHAO REN[2]
[1]School of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130022, China
[2]School of Computer Science and Technology, Jilin University, Changchun 130012, China

Corresponding author: Jianping Zhao (jpzhao@yeah.net)

This work was supported by the Science and Technology Development Program of Jilin Province, China under Grant 20190303133SF.

ABSTRACT Urban data is a imperative resource for urban computing, which can promote the establishment of urban knowledge, the collection of urban information and the construction of smart cities. Seen that the urban data collected through hardware sensors and crowdsourcing has the limitations of uneven information distribution, poor data comprehensiveness and high resource costs, we turn to the Internet resources of real-time updates and extensive information coverage. Therefore, we propose an approach to Sensing Urban text Data from Internet Resources (SUDIR). We put forward innovative work on two key issues: urban data recognition for Chinese context and urban data sensing for multi-source web resources. On one hand, we design a Chinese urban data recognition model based on Whole Word Masking for Bidirectional Encoder Representations from Transformers (BERT-WWM) embedding model and Bidirectional Long-Short Term Memory with a Conditional Random Field (BLSTM-CRF) sequence labeling model. We introduce Chinese Word Segmentaion (CWS) concept in BERT embedding model to make the text embedding effect better represent semantic information on Chinese context. BLSTM-CRF model based on deep learning is used to achieve high-quality coding and prediction. On the other hand, we propose a method of Extracting Urban text data based on Web page features and Clustering operation (EUWC). EUWC is used to correct the false negative samples labeled by BERT-WWM+BLSTM-CRF recognition model and enable SUDIR to sense more accurate and comprehensive city data from multi-source web resources. The experimental results show that our work outperforms the other baseline methods, and it also proves that SUDIR using Internet resources and deep learning technology has the advantages of low-cost, high-quality urban data sensing.

INDEX TERMS Urban data sensing, urban computing, Internet resources, deep learning, Chinese text, web page features.

I. INTRODUCTION

Urban computing is driven by sensing data and serves various fields (such as, transportation planning, energy consumption, environmental monitoring, etc) in urban development [1]. Urban sensing technology of urban computing is used to obtain urban data to serve Location Based Services (LBS). LBS take urban data as the computing resource to provide users the services in travel, entertainment, daily life, etc. As the most widely used LBS in the world, GoogleMaps utilizes global procurement as the main solution to update the urban database and is of high labor cost and low efficiency. Other LBS, such as OpenStreetMap and Wikimapia, use the manually labeled geographic information volunteered by regular users to update the urban database, in which the quality and efficiency of extracting data mainly depends on user behaviors and the accuracy can hardly be guaranteed. In addition, the methods of collecting urban data through hardware sensors are even more expensive and inconvenient for widespread deployment, which usually result in uneven distribution of data samples and poor comprehensiveness. Therefore, how to obtain comprehensive, accurate and low-cost urban data is a critical issue in urban sensing research. Given that Internet senses the massive data resources that are always updated to time, the idea of employing Internet as the sensors to collect data with low cost, high efficiency and comprehensive coverage is emerging and it suggests a feasible means for urban sensing. There are many

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

works that prove the feasibility of using Internet resources for data mining. Many works in different fields have proved the feasibility of using Internet resources for data mining, such as medical field [2] and big data classification [3], etc.

The categories of urban data include Point Of Interest (POI) data, map data, Global Positioning System (GPS) data, passenger flow data, mobile phone data, etc. POI can represent any non-geographically meaningful point on the map. A POI can be a house, a shop, a post box, a bus stop, etc. POI data is the data describing the basic information (POI-name, POI-address, POI-coordinates, POI-type, etc) of each functional unit of city. Urban data transmission and storage forms include text, streaming, pictures, audio, etc. Text data, as one of the most basic forms of urban data, is the most common type of data on the Internet and manifests itself as unstructured. As text data is a kind of unstructured data, it is challenging to implement standardization and understand. It is necessary and urgent to acquire accurate, comprehensive and up-to-date urban data. By reviewing the overview references (cf., [4]–[6]), more and more works for urban text data sensing based on Internet resources have been proposed. Most of the traditional research works focus on the rules-based, external knowledge-based and statistical learning-based methodologies of sensing urban text data. The first category of methods is the rule-based method (cf., [7]–[9]). Reference [7] is based on rules and crawler technology to extract POI data and build an maps/location searching application from Internet resources. In addition, there are some rule-based methods used to extract urban data from specific data sources (cf., [8], [9]). The rules-based methods can achieve high-accuracy data collection, but due to the increasing update of Internet information, crawler tools or rules need to perform a lot of maintenance work. The second category of methods is based on external knowledge [10], [11]. They are based on gazetteers to extract location names from structured and unstructured Internet social media data. These methods can achieve data sensing in specific tasks, but they are still difficult to deal with true Internet scenarios. This is because the data in Internet is complex and diverse, and the quality of the knowledge bases is also different. Such the technologies have limitations in generalization. The third category of main methods is based on statistical learning (cf., [12], [13]) or language models (LDA [14], n-grams [15]). The works based on statistical learning and language models are weaker than the models based on deep neural network in obtaining complex and obscure sample information [16]. The models based on deep learning have the advantages of good portability, high performance limit, good adaptability and strong learning ability, which are suitable for urban data sensing facing Internet resources. In addition, how to sense more comprehensive urban data from multi-source Internet resources is also a key issue.

In this paper, we propose an approach of Sensing Urban text Data from Internet Resources based on deep learning (SUDIR) to provide a more effective solution to collect urban text data. SUDIR has many novel work that superior than other traditional data sensing methodologies. First, SUDIR uses BERT-WWM to improve the semantic expression of Chinese text sequences without introducing external knowledge, especially in Chinese plain text. Whole Word Masking (WWM) strategy can enable the word embedding model to better express Chinese semantics through Chinese Word Segmentation (CWS) concept. Secondly, SUDIR uses deep learning-based Bidirectional Long-Short Term Memory with a Conditional Random Field (BLSTM-CRF) to improve the fitting and generalization ability of the urban data recognition model. Finally, a method for Extracting Urban data based on Web page features and Clustering operation (EUWC) is designed to discover the existence of urban data in multi-source web resources. SUDIR can extract more comprehensive and accurate urban data from Internet resources by combining EUWC with BERT-WWM+BLSTM-CRF recognition model. It is worth mentioning that, in addition to basic computing resources, the process of using SUDIR to sense urban data does not require more labor costs and hardware sensor costs.

The rest of this paper consists of four sections. Section II introduces the related work. The details of SUDIR are elaborated in Section III. Section IV states the experimental verification, in which the research results of this paper are verified with the baseline models. Section V summarizes the research achievement and suggests the future work.

II. RELATED WORK

This section is divided into three parts to introduce related research on Internet urban data sensing, namely: (1) Named entity recognition of urban text data; (2) Word embedding expressing semantic features; (3) Internet data extraction and web clustering.

A. NAMED ENTITY RECOGNITION FOR URBAN DATA

Extracting urban text data from Internet resources is a sequential labeling task, i.e., the application of Name Entity Recognition (NER) in urban data processing. At present, there are mainly the following problems concerning Chinese NER.

1) There is no commonly accepted standard description of name entity, i.e., colloquial expressions, non-recognized abbreviations and multiple variants exist.
2) Entities outside the corpus are difficult to identify.
3) Because Chinese does not have some special characteristics, such as the capitalization of the first character of each word and the spaces to break words/phrases, etc. Chinese NER is more complicated than English NER.

The main solutions employed for NER tasks include: the traditional manual rules-based approach with combining with other knowledge [18], [19], the statistical learning approach [20]–[23] and the approach integrating deep neural network models [24]. Since the deep neural networks do not need to manually construct the feature and learn the fuzzy information from data, the approaches integrating deep neural networks are capable of saving cost and significantly improving the NER performance. In recent years, the NER model based

on deep learning has achieved many positive achievements, such as the character-based BLSTM-CRF [25], the lattice LSTM-CRF [26] that is applicable to Chinese grammar rules, sequence labeling model using transfer learning [27], NER model based on automatic continuous segmentation and deep learning [28], and the Iterated Dilation-Convolutional Neural Network (ID-CNN) [29] that can alleviate the loss of information in the traditional Convolutional Neural Network (CNN) and take into account the computational efficiency and result accuracy.

The end-to-end sequence model based on BLSTM-CRF structure manifests better performance without feature engineering or introduction of any external knowledge base [30], [31]. The model structure using BLSTM encoding and CRF decoding has been applied to many Natural Language Processing (NLP) tasks [32]–[35] and accomplished good performance. In this paper, we employ BLSTM-CRF as the fundamental structure for building up the Chinese urban data recognition model.

B. WORD EMBEDDING

In NLP tasks, the feature representation of text is a critical issue. In recent years, many text vector construction methods have been used to describe text semantics. The most representative work of word vector construction is Word2vec approach proposed by Mikolov in [36], which makes the training of word embedding more efficient by removing the hidden layer. Word2vec has two styles, i.e., CBOW and skip-gram. BERT provides a new direction for word embedding methods and demonstrated better text information description. By adopting Bidirectional Encoder Representation Transformers (BERT) [37] embedding method, the representation ability of word embeddings has been significantly improved. Transformer [38] neural network model extracts the global dependencies from the input as well as the output entirely relies on the attention mechanism. Compared with CNN, RNN and other deep neural network structures, transformer not only saves computing resources but also guarantees the feature extraction ability of sequence, showing outstanding performance. Compared with one-hot, CBOW, skip-gram and other word embedding construction methods, BERT embeddings play better in most of the NLP's downstream tasks.

The structure of BERT is built by the multi-layer bidirectional transformers. BERT embeddings constructed by combining the wordpiece embedding, position embedding and segment embedding. BERT can overcome the difficulty of feature-based approaches [39], [40] and fine-tuning approaches [41], [42] in terms of context information feature extraction.

When training the English embedding models, BERT adopts Masked Language Model (MLM) strategy to conduct the prediction result during the training by randomly inputting some tokens of the mask. The advantage of MLM is that the learned representation could combine the context in both the directions. However, it is not entirely reasonable to use BERT's masking strategy in Chinese context, because there is no interval identifier between Chinese words, which will cause characters instead of words to be randomly masked. To solve this problem, we introduce CWS concept and build the Chinese embedding model based on BERT-WWM [43] proposed by HFL team.

C. DATA EXTRACTION AND WEB CLUSTERING

Web page text has more features than the plain documents, such as URL addresses to locate resources, HTML text to describe page information, HTTP/HTTPS protocol prefix, etc. Therefore, designing methods of extracting data from web page text data needs to consider the characteristics of Internet resources. The selection of data source and the construction of data extraction model are two key issues. Most of the work for extracting urban data from Internet resources (e.g., [12], [21]) usually sets one or more specific websites as the stable data source. However, the use of specific websites as data sources limits the diversity of data collection, making it difficult to obtain data with available comprehensiveness. This paper takes public network containing data from a variety of sources as a perceived resource to address the need to extract comprehensive rosy urban data. The data available on public websites are extensive but extremely noisy and often inaccurate. Given the above problems, many learning models based on web page structure features [44], [45], block tag features [46], [47] and data features [48], [49] were proposed. These research results provide some theoretical basis and reference methods for extracting urban data from Internet resources.

In recent years, some clustering algorithms for identifying and querying information from Internet resources have been proposed [50], [51]. Y Fang *et al.* believe that the essence of the user's task of obtaining information from the WEB is Learning to Query (L2Q), i.e., intelligent search query, which can employ search engines to retrieve a web page containing the entity information of interests [52]. Inspired by the idea of Y Fang, we use clustering algorithms to group web pages into clusters to alleviate the difficulties of data-extraction caused by diverse data sources. The mainstream clustering algorithms can be categorized as classified clustering, hierarchical clustering, density clustering. The summary of mainstream clustering algorithms is shown in Tab.1.

K-means [53] algorithm cannot automatically discard noisy data, and the K values need to be particularly configured. K-means algorithm emerges a little weak in coping with non-spherical clusters and clusters of multiple sizes. Therefore, it is not suitable for handling the work in big data environments, e.g., Internet. Hierarchical clustering [54] is not seldom used to treat big data sets because of its high consumption of computational resources. Taking Agglomerative Nesting (AGNES) and Divisive Analysis (DIANA) as examples, the computational complexity is the number of iterations by the square of the sample points. Density-Based Spatial Clustering of Applications with Noise (DBSCAN [55]) can handle the clusters of different sizes

TABLE 1. Mainstream clustering algorithms.

Representative algorithm	Categorie	Advantages	Disadvantages
K-means	partitional clustering	Low computational complexity,Favors clustering in dense, cluster-like clusters	Sensitive to k-value settings and noisy data,Easy to fall into local optimum
AGNES, DIANA	hierarchical clustering	Distances and rules are easy to define. Discover hierarchical relationships without specifying the number of clusters in advance	Singular values can have a big impact on the results; the algorithm is likely to aggregate into a chain and the computational complexity is high ($t * n^2$)
DBSCAN, OPTICS	density clustering	Insensitive to noise, can find clusters of any shape	Sensitive to the two parameters of cluster spacing and threshold (DBSCAN)

or shapes. DBSCAN algorithm is less affected by noise and outliers and can automatically determine the number of clusters. Such density-based clustering method is suitable for the application of web page clustering [56]. However, DBSCAN also has its limitations. DBSCAN is sensitive to the two parameters of cluster distance and minimum threshold of cluster density. In terms of parameter selection, Ordering Points To Identify the Clustering Structure (OPTICS) [17] algorithm has superior flexibility than DBSCAN. OPTICS can process a set of cluster distance parameter values at once. Therefore, we adopt OPTICS as the core algorithm of Internet web page clustering.

In view of the analysis of the above-mentioned related work, we propose SUDIR to sense urban data on Internet resources. SUDIR uses the advantages of deep learning-based NER technology in feature learning capabilities and model generalization to build urban data recognition model. To enrich the semantic information representation of Chinese plain text, the data recognition model uses BERT combined with whole word mask strategy to realize the construction of word embeddings that introduce the concept of Chinese word segmentation. In addition, based on Web mining technology and OPTICS clustering algorithm, we propose a method called EUWC to make SUDIR more accurate and comprehensive in the Internet application scenarios to sense urban data in multi-source web pages.

III. THE URBAN DATA SENSING APPROACH

In this section, we describe the urban data sensing approach, SUDIR. First, we describe the general framework of SUDIR. After an overview of SUDIR, we introduced the Chinese urban text data recognition model, which consists of BERT-WWM embedding layer, BLSTM encoding layer and CRF prediction layer. Finally, we introduced how EUWC senses more comprehensive and accurate urban data from multi-source web resources.

A. GENERAL FRAMEWORK

SUDIR is a hierarchical approach framework that uses the Internet as sensors and realizes the sensing of urban data. First, SUDIR's data acquisition layer is used to collect target data from Internet resources. Second, the processing layer of SUDIR is used to preprocess the collected data, recognize urban data and extract urban data extraction. Finally, SUDIR's data result layer is used to return valuable urban data. SUDIR's structure as shown in Fig. 1, and the layers of SUDIR are listed as follows:

1) **Data acquisition layer**: According to the content to be retrieved, the transfer request is sent to the Internet resources, and the returned web pages are transferred to web page text data acquisition module. Data acquisition module collects the web pages to be processed based on crawler technology and transfer the collected web page data to handing layer.
2) **Handing layer**: First, text preprocessing module performs data preprocessing on the received web page data, such as text denoising, deduplication and text segmentation. Then, urban data recognition module employs urban data recognition model (BERT-WWM+BLSTM-CRF) to label the data categories in the web page HTML text. Next, urban data-extraction module extracts the data labeled as urban data based on EUWC method and forwards them to data result layer.
3) **Data result layer**: The extracted urban data is built into a data set to provide computing resources for the applications and support urban sensing.

To implement a high-performance urban data recognition function, we propose an urban data recognition model based on deep neural network BERT-WWM+BLSTM-CRF model. The urban data recognition model uses BERT-WWM to embed text sequence and takes BLSTM-CRF as the sequence labeling model. EUWC method based on web page features and OPTICS clustering algorithm is used to realize the function of urban data extraction module from multi-source Internet resources. By combining BERT-WWM+BLSTM-CRF and EUWC, SUDIR can more accurately and comprehensively sense urban text data in Internet scenarios.

B. RECOGNIZING URBAN DATA BASED ON BERT-WWM AND BLSTM-CRF

SUDIR's urban data recognition model uses the WWM strategy of the CWS concept to construct word embeddings to enrich semantic representation, and uses bi-directional deep neural network model to obtain excellent performance for extracting urban data from Chinese plain text. The deep neural network structure of BERT-WWM+BLSTM-CRF urban data recognition model proposed for SUDIR's urban data recognition module includes three layers, i.e., BERT-WWM

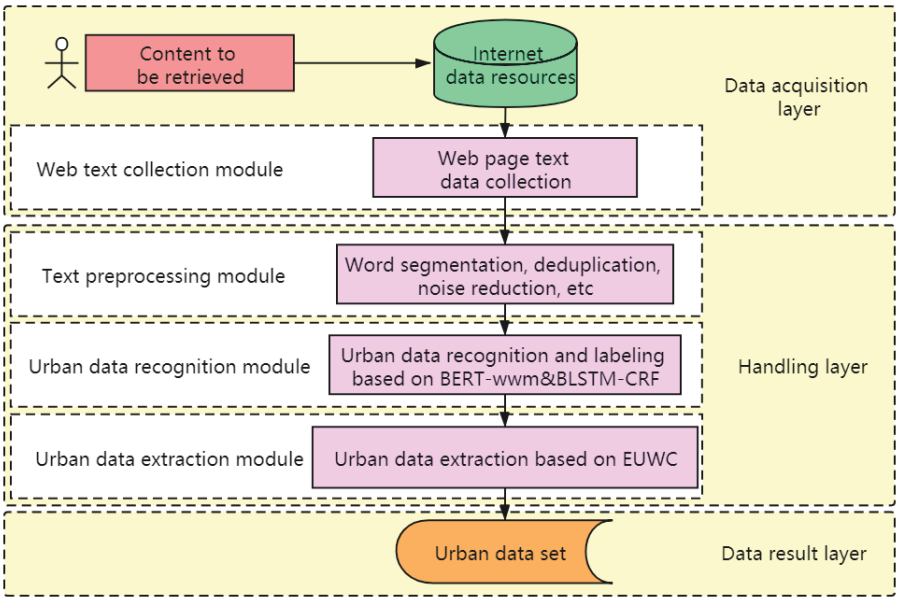

FIGURE 1. Framework of SUDIR. SUDIR is consisted as Data acquistion layer, Handing layer and Data result layer.

word embedding layer, BLSTM encoder layer and CRF prediction layer. This model utilizes BERT model that introduces the whole word mask strategy to enrich the semantic representation capabilities of Chinese text. The input embeddings of BERT-WWM are built up with token embedding, segment embedding and position embedding that respectively represent token, segment and position information in the input sequence. The embeddings are then sent to the Bi-directional LSTM encoder layer that can obtain the hidden information from both forward and backward directions. CRF is finally invoked as the structural unit of the prediction layer to carry out the probability calculation on the hidden information, during which the maximum probability sequence obtained by prediction is the output result.

The structure of BERT-WWM+BLSTM-CRF model is illustrated as Fig.2, in which W represents the embedding vectors of Chinese text words, L and R represent the feature extraction unit of BLSTM, h represents the hidden vector output by encoder layer. Output labels represent the results of CRF prediction layer.

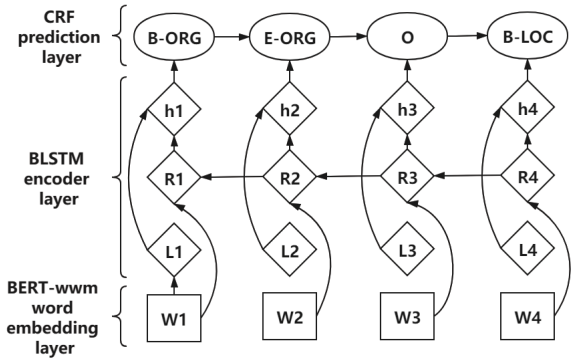

FIGURE 2. The structure of urban data recognition model. The data recognition model is consists of BERT-WWM word embedding layer, BLSTM encoder layer and CRF prediction layer.

1) BERT-WWM EMBEDDING LAYER

English emphasizes grammatical structure, while Chinese emphasizes semantics. Chinese grammar is more flexible. As long as the concept is expressed correctly, many sentences are free of language disorders and ambiguities. English grammar has more fixed collocations, sometimes the meaning is correct but grammatically speaking it is wrong. For example, in some collocations, some prepositions such as to, in, for have no practical meaning, but these prepositions must be reserved in English expressions. However, the prepostitions are not necessary in Chinese expressions. Moreover, there are no spacers similar to the ''space'' in English sentences between words in Chinese sentences. This means that the masking strategy of English word embedding construction method cannot be fully used in the Chinese word embedding construction method. An example of Chinese masking strategy and other details is shown in Fig.3. Directly using the BERT-base project officially released by Google (https://github.com/google-research/bert) to apply character masking as a training strategy has limitations in the construction of Chinese word embeddings. This does not make the embedding model have a better training strategy, resulting in the loss of some semantic information when ignoring CWS concept.

According to WWM strategy, if particular characters of a word are masked, other parts belonging to that word will also be masked. In view of this, we adjusted character masking

[Original Sentence] 使用**BERT**模型来预测下一个出现的单词。 Use **BERT** model to **predict** the next **word**. (In English)
[Original Sentence with Chinese word segmentation] 使用, **BERT**, 模型,预测,下一个,出现,的,单词,。 Use, **BERT**, model, to, **predict**, the, next, **word**. (In English)
[Original BERT masking Input] 使用**B[MASK]RT**模型来预**[MASK]**下一个出现的单**[MASK]**。 Use **B[MASK]RT** model to **pre[MASK]** the next **wo[MASK]**. (In English)
[Whole word masking Input] 使用**[MASK][MASK][MASK][MASK]**模型来**[MASK][MASK]**下一个出现的**[MASK][MASK]**。 Use **[MASK][MASK][MASK][MASK]** model to **[MASK][MASK]** the next **[MASK][MASK]**. (In English)

FIGURE 3. An example of Chinese masking strategy. It shows the example sentence's original description, sentence with Chinese word segmentation and input embeddings with various masking strategies.

[Original Sentence] 用于测试的句子。 A sentence to be used for **testing**. (In English)
[Original Sentence with Chinese word segmentation] 用于,测试,的,句子,。 A, sentence, to, be, used, for, **testing**. (In English)
[Replace the word with [MASK] token] 用于**[MASK][MASK]**的句子。 A sentence to be used for **[MASK][MASK]**. (In English)
[Replace the word with a random word] 用于篮球的句子。 A sentence to be used for **basketball**. (In English)
[Keep the word un-changed] 用于测试的句子。 A sentence to be used for **testing**.(In English)

FIGURE 4. An example of mask setting policy. It shows the example sentence's original description, sentence with Chinese word segmentation and three mask setting policies.

strategy in the training process of BERT model to WWM strategy to improve the representation of text semantic features to make it more suitable for Chinese context. The embedding model generated by applying WWM strategy to BERT embedding model is called BERT-WWM.

The parameters of BERT-WWM's masking setting proposed in literature [37] are adopted in our work. 15% of all words are replaced by *Mask*. Of all the masked words, 80% are replaced by **[Mask]**, and 10% are replaced by random words, and the last 10% are kept as the original words (Fig.4). Further technical details of BERT-WWM are the same as BERT.

2) BLSTM ENCODER LAYER

The pre-trained BERT-WWM model executes the translation of text sequence to embedding vectors. The embedding vectors are invoked as input to BLSTM encoder layer. BLSTM is created through improving LSTM, and the bidirectional strategy utilized in the NLP task can better extract the context

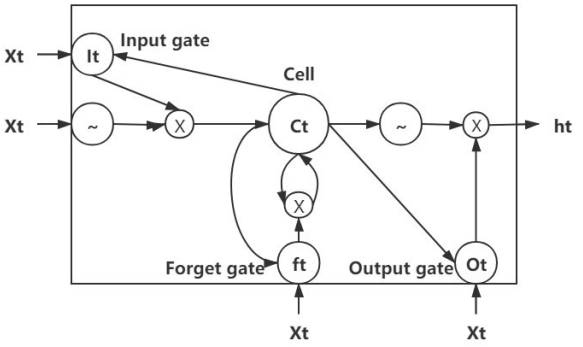

FIGURE 5. LSTM at time *t*.

feature and enrich the feature representation. LSTM is one of the variants of Recurrent Neural Network (RNN). LSTM adds forget gate and cell unit based on the long sequence processing of RNN. Rather than RNN, LSTM can better retain the adequate information and forget some noise information.

Fig.5 gives the illustration of the LSTM deep neural network at time t.

At time t, the calculations of LSTM are stated as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}h_t + W_{hc}c_{t-1} + b_c) \quad (3)$$

in which σ represents the sigmoid function, i_t represents the calculation results of the input unit at time t, f_t represents the calculation results of the forget gate at time t and c_t represents the calculation results of the cell at time t.

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (4)$$

$$h_t = o_t \tan h(c_t) \quad (5)$$

in which o_t represents the calculation of the output unit at time t, and h_t represents the calculation of the hidden unit at time t. Where W and b are weight matrix and bias vectors.

The word embedding sequence obtained through the embedding layer is denoted as $S = (c_1, c_2, \ldots, c_n)$, in which $c_i(1 <= i <= m)$ represents the embedding vector of the $i-th$ content in s sequence. The output of the forward hidden unit and the reverse hidden unit of BLSTM are represented as $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ respectively.

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(s_t, \overrightarrow{h_{t-1}}) \quad (6)$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(s_t, \overleftarrow{h_{t-1}}) \quad (7)$$

After concatenating $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$, the hidden feature of BLSTM is represented as $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$.

3) CRF PREDICTION LAYER

The output h_t of BLSTM encoder layer is the input of CRF prediction layer. CRF prediction layer's structure is shown in Fig.6. In CRF prediction layer, we predict the optimal sequence built by CRF and Viterbi algorithm. CRF is a directed graph model that implements the global probability

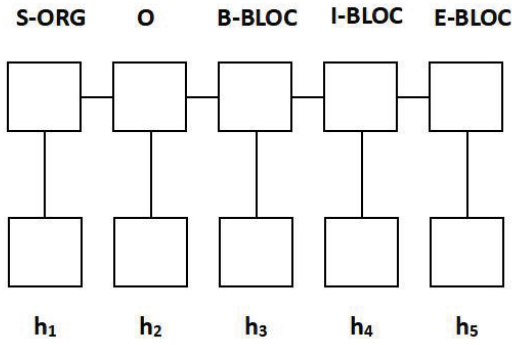

FIGURE 6. Structure of conditional random fields. h_x represents the hidden information calculated by BLSTM encoder layer. Outputs are the data recognition results of CRF prediction layer.

statistics. When performing normalization, CRF considers the global distribution of the data, not just the local normalization, so that it could decode the sequences better.

The definition of the prediction algorithm is stated as follows: the random field model feature is denoted as $F(h, l)$, the weight vector is represents by w, the input sequence (i.e., the observation sequence) is denoted as $h(h_1, h_2, \ldots, h_n)$. The algorithm predicts the sequence of observations to obtain the maximum condition probability output sequence, i.e., y^*. Viterbi prediction algorithm based on CRF model is specified as in Alg.1. The maximum probability output sequence $y^* = (y_1^*, y_2^*, \ldots, y_n^*)$ is the prediction sequence of data categories.

Algorithm 1 Viterbi Prediction Algorithm Based on CRF Model

Require:
Input feature: $F(h, l)$;
Weight vector: w;
Observation sequence: $h(h1, h2, \ldots hn)$;
Ensure:
Output sequence $y^* = (y_1^*, y_2^*, \ldots, y_n^*)$;
1: Initialize and calculate the denormalized probability of each label $j = 1, 2, \ldots, m$.
$\delta_1(j) = w \bullet F_1(L_0 = start, l_1 = j, h)$
2: **for** i in range(n) **do**
3: Calculate the denormalized maximum of each label $l = 1, 2, \ldots m$ at position t
$\delta_i(l) = \max_{1 \le j \le m} \{\delta_{i-1}(j) + w \bullet F_i(y_{i-1} = j, y_i = l, h)\}$
4: Record the path of the denormalized maximum
$\Psi_i(l) = \arg \max_{1 \le j \le m} \{\delta_{i-1}(j) +$
$w \bullet F_i(y_{i-1} = j, y_i = l, h)\}, l = 1, 2, \ldots, m$
5: **end for**
6: $\max y(w \bullet F(y, h)) = \max_{1 \le j \le m} \delta_n(j)$
7: $u_n^* = \arg \max_{1 \le j \le m} \delta_n(j)$
8: **return** prediction result
$y_i^* = \Psi_{i+1}(y_{i+1}^*), i = n-1, n-2, \ldots 1$
9: Maximum probability output sequence:
$y^* = (y_1^*, y_2^*, \ldots, y_n^*)$

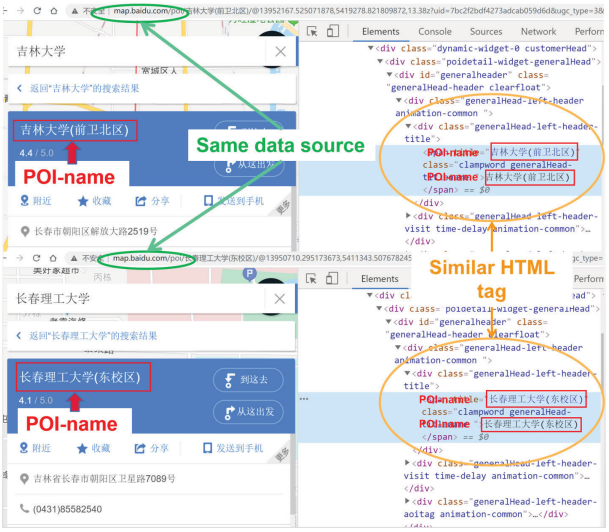

FIGURE 7. An example of the same category of urban data with similar tags in the identical data source.

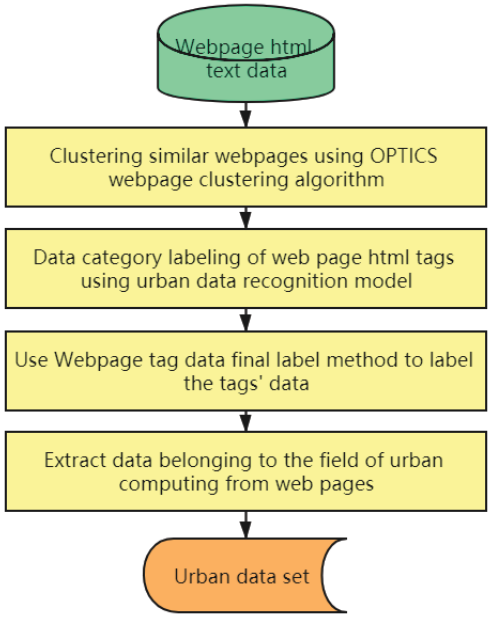

FIGURE 8. Execution steps of EUWC.

C. EUWC:EXTRACTING URBAN DATA BASED ON WEB PAGE FEATURES AND CLUSTERING OPERATION

We found that the data categories in similar HTML tags of the same Internet data source are usually the same, but the data recognition model cannot find and use this information to sense urban data. There is an example as shown in Fig.7. In response to this limitation, we propose a method called EUWC to enable the data recognition model to better sense the urban data in multi-source web resources. EUWC can correct the false negative samples labeled by data recognition model to make the sensed Internet urban data more comprehensive and accurate. The procedures of EUWC are shown in Fig.8. First, EUWC uses OPTICS clustering algorithm

to classify multi-source web pages into clusters and remove the noisy web pages. Then, the deep learning-based BERT-WWM+BLSTM-CRF recognition model is utilized to carry out the initial labeling of the urban data in web pages' tag text. Finally, the data are finally labeled. If most of the data in a specific tag are of the same category, the tag's data will be labeled as the same category. If the tag categories of data in a specific tag are scattered, the tag's data will adopt the initial results. We set the threshold value M to trains as the selection condition.

1) OPTICS WEB PAGE CLUSTERING ALGORITHM BASED ON WEB PAGE FEATURES

We propose a web clustering method based on web page features and OPTICS algorithm to cluster multi-source web pages, as shown in Alg.2. The structure of the web pages' HTML tags is a tree structure. HTML tag is the most basic and paramount unit in web pages. Since HTML data are utilized to describe web page, we use HTML data structure and HTML tag content as the features to implement the similarity calculation of web pages. An HTML structure tree example of a web page is presented in Fig.9. We take the web structure and text as features and propose a web page distance calculation (Eq.8-Eq.11) based on HTML's structure and tag features.

Algorithm 2 Web Clustering Algorithm Based on OPTICS

Require:
Sample collection: $D = d1, d2, \ldots, dn$;
Cluster distance: ε ;
Minimum threshold of cluster density: $MinPts$

Ensure:
$P = \{p_i\}_{i=1}^{n}$

1: Initialize variables: $k = 1$;
$v_i = 0$ //represents whether d_i has been visited;
r_i=UNDEFINED, $i \in 1, 2, \ldots n$ //represents d_i's reachable distance.
2: **while** $D \neq \emptyset$ **do**
3: Get a sample d_i from D, and let $D := D \backslash \{d_i\}$;
4: **if** $v_i = 0$ **then**
5: $v_i = 1, p_k = i, k = k + 1$;
6: **if** $N\epsilon(d_i) >= MinPts$ **then**
7: //Insert the unvisited samples in $N\epsilon(d_i)$ into the queue according to the reachable distance, c_i represents the core distance of d_i
$insertlist(N\epsilon(d_i), \{v_l\}_{l=1}^{n}, \{r_l\}_{l=1}^{n}, c_i, seedlist)$;
8: **while** $seedlist \neq \emptyset$ **do**
9: Get the smallest reachable distance sample d_j from $seedlist$
10: $v_j = 1, p_k = j, k = k + 1$
11: **if** $N\epsilon(d_j) >= MinPts$ **then**
//Insert the unvisited samples in $N\epsilon(d_j)$ into the queue according to the reachable distance
$insertlist(N\epsilon(d_j), \{v_l\}_{l=1}^{n}, \{r_l\}_{l=1}^{n}, c_i, seedlist)$;
12: **end if**
13: **end while**
14: **end if**
15: **end if**
16: **end while**

$$WP_{sim}(A, B) = \frac{\omega^{tag} WP_{sim}^{tag}(A, B) + \omega^{str} WP_{sim}^{str}(A, B)}{2} \quad (8)$$

Eq.8 presents the similarity measurement between web page A and web page B, which states one half times of the sum of tag similarity and structure similarity. In Eq.8, $WP_{sim}^{tag}(A, B)$ and $WP_{sim}^{str}(A, B)$ represent the web pages' HTML tag similarity and structural similarity, respectively, ω^{tag} and ω^{str} represent the two level similarity importance weights, and the sum of ω^{tag} and ω^{str} is 2. The result interval of $WP_{sim}(A, B)$ is [0, 1].

$$WP_{sim}^{tag}(A, B) = \frac{WP_{same}^{tag}(A, B) * 2}{nA + nB} \quad (9)$$

Eq.9 specifies the similarity measurement for the HTML tags between web page A and web page B. WP_{same}^{tag} denotes the number of nodes with the same tag content. nA and nB represent the respective node numbers of the two trees. The result interval of $WP_{sim}^{tag}(A, B)$ is [0, 1].

$$WP_{sim}^{str}(A, B) = \frac{2 * SimTreeMatching(A, B)}{nA + nB} \quad (10)$$

Eq.10 specifies the structural similarity measurement between web page A and web page B, where $SimTreeMatching(A, B)$ represents that the maximum matching node between the two trees in the current layer is returned recursively layer by layer and the return value of each layer is accumulated. The result interval of $Dis(A, B)$ is [0, 1].

$$Dis(A, B) = 1 - WP_{sim}(A, B) \quad (11)$$

Eq.11 calculates the distance of web page based on web pages' similarity. The result interval of $Dis(A, B)$ is [0, 1]. The smaller the $Dis(A, B)$ is, the more similar A and B are, and the closer they are in the mapping space. $Dis(A, B)$ denotes the measurement of web page spacing based on OPTICS web page clustering algorithm. Details of web clustering algorithm based on OPTICS are stated as Alg.2.

Related definitions in OPTICS:

The definition of core point, if the number of points contained within the radius of a point is not less than the minimum number of points($MinPts$), then the point is core point, and the mathematical description is:

$$N\epsilon(p) >= MinPts \quad (12)$$

The definition of core distance, that is, for the core point, the distance from the $MinPtsth$ closest point to it.

$$cDis(p) = \begin{cases} \text{UNDEFINED, if } N\epsilon(p) <= MinPts, \\ MinPtsth \text{ Distance in } N\epsilon(p), \text{ otherwise.} \end{cases} \quad (13)$$

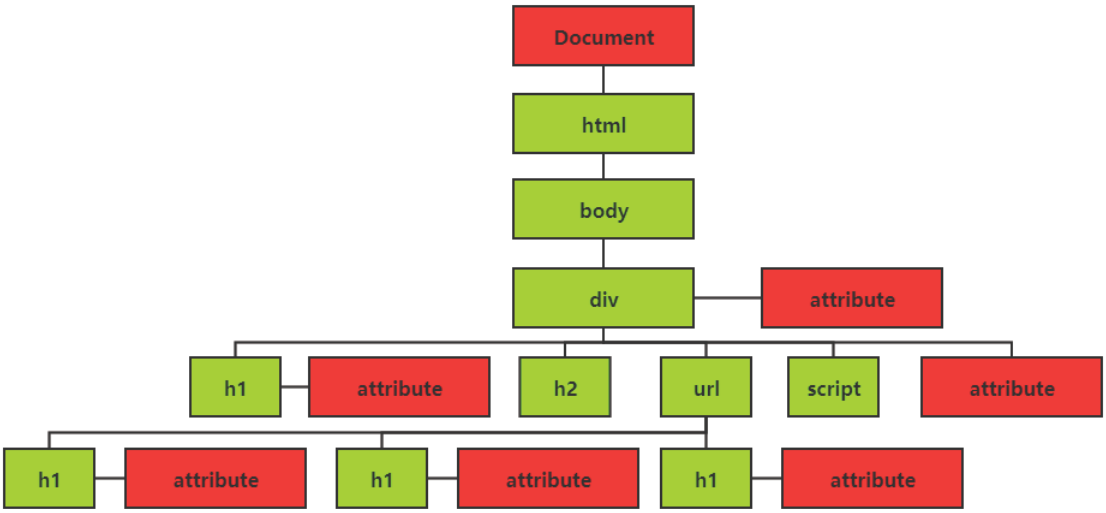

FIGURE 9. An example of html structure tree. The red blocks represent the web page tag categories, and the green blocks represent the attribute text.

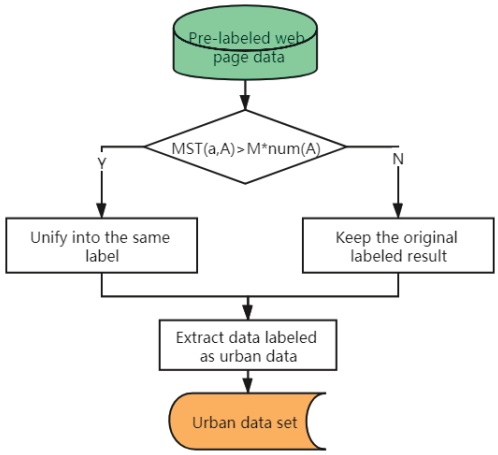

FIGURE 10. Schematic diagram of web page tag data final labeling and extraction method.

The reachable distance, for the core point p, the reachable distance from o to p is defined as the distance from o to p or the core distance of p, which is the equation:

$$rDis(o,p) = \begin{cases} \text{UNDEFINED, if } N\epsilon(p) <= MinPts, \\ max(cDis(p), Dis(o,p)), \text{otherwise.} \end{cases} \quad (14)$$

The direct density from o to p is reachable, that is, p is the core point, and the distance from p to o is less than the radius.

2) FINAL LABELING AND EXTRACTION METHOD OF WEB PAGE TAG DATA

We propose a web tag data final labeling and extraction method to complete the extraction of urban text data in web pages. Fig.10 shows the process of labeling and extracting urban data of a specific clustered web resource. $MST(A, a)$ refers to the Maximum number of Similar data in Tag a of resource A's web pages when BERT-WWM+BLSTM-CRF urban data recognition model is used. M (the interval is [0, 1]) represents the preset threshold value of the judgment condition; $num(A)$ represents the number of web pages belonging to cluster A.

After executing the labeling selection criteria, if the amount of data in cluster A's tag a is greater than $M * num(A)$, all the data in tag a is labeled as of the same category. Otherwise, the data retain the original label category. Finally, the data labeled as urban data are extracted to complete creating the urban dataset.

IV. EXPERIMENTS AND EVALUATION

In this section, a series of experiments are carried on the experimental datasets to show the significant superiority and high performance of our work compared with the existing baseline models. We experimented and analyzed the performance of the Chinese urban data recognition models, the web pages clustering effect of multiple data sources and the Internet urban data extraction effect of SUDIR. First, through BERT-WWM embedding representation based on whole word mask, the NER model of BLSTM encoding and CRF prediction is compared with other baseline models to demonstrate the effect of the proposed model's superiority in Chinese urban data recognition. Second, a set of experiments was carried out on the selection of clustering algorithm and the parameter setting of EUWC method to verify the feasibility of using web page features and density-based clustering algorithm to optimize the recognition effect of the urban data recognition model. Finally, the SUDIR combining BERT-WWM+BLSTM-CRF and EUWC is compared with other urban data extraction methods. The experimental results demonstrate that SUDIR is more accurate and comprehensive in extracting urban text data from different Internet data sources.

A. DATASETS SPECIFICATIONS

We utilize four open-source datasets (Tab.2) to train urban data entity recognition model, i.e., BosonNLP dataset

TABLE 2. Summary of the NER datasets.

Data set	Sentences	Entities	POI-Add	POI-Nam	Others
BosonNLP	2000	12427	4597	2689	5141
MSRA	46364	74703	36517	20571	17615
RMRB	163492	53242	22427	10834	19981
C-NER	27818	45518	22180	12446	10892

TABLE 3. Specifications of the web page dataset.

	Web sources	Valuable data	Noise data	Total data
Number	4	1416	100	1516

TABLE 4. Summary of web page dataset.

Web source	Number	POI-Add	POI-Nam
map.baidu	662	1324	1324
elong	303	2424	3636
city8	102	4896	3264
qy.58	349	13021	1796
Total	1416	21665	10020

(https://bosonnlp.com/), 1998RenMinRiBao dataset (http://www.icl.pku.edu.cn/icl_res/), MSRA dataset (https://www.msra.cn/) and Chinese NER dataset on github (https://github.com/zjy-ucas/ChineseNER). To verify the effectiveness of SUDIR approach in the task of urban data recognition, we keep two entity categories, i.e., the category of POI-address (POI-Add) and POI-name (POI-Nam) and the category of other remaining entities. The dataset uses *BIEOS* (i.e., *B*-Begin, *I*-Inside, *O*-Outside, *E*-End, *S*-Single) tagging mode to tag the named entity data. The data proportions of training set, validation set and test set are 0.7, 0.1 and 0.2 respectively.

To verify SUDIR's urban data sensing ability in multi-source Internet resource scenarios, we use crawler technology to obtain 1416 web page HTML data with geographic information from four web page data sources, and randomly collect 1000 web page data from other data sources as noise data to construct an experimental web page data set. We manually labeled the web page categories attributes of the dataset and utilized the open-source tool Beautifulsoup (https://pypi.org/project/beautifulsoup4/) to reconstruct the tree structure data. The overall specifications of web page data set is shown in Tab.3.

The data of web pages comes from four mainstream LBS in Chinese domain, whose website names (URL) are Baidu map (http://map.baidu.com), elong (http://www.elong.com), city8 (http://www.city8.com) and WUBA (http://www.58.com/) respectively. Tab.4 shows the detail information of web page dataset. The tag category of the urban data is POI-Add and POI-Nam. POI-Add and POI-Nam correspond to the organization and location name entities in NER, accordingly there is no need to repeatedly train the urban data recognition models.

B. EXPERIMENT SETUP

We establish and employ Tensorflow framework (version 1.13.2) to perform all experiments in Python 3.6 environment. The train dataset of word vector models is from Chinese Wikipedia corpus (https://dumps.wikimedia.org/ zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2). We select LTP (http://ltp.ai) contributed by Harbin Institute of Technology as the text segmentation tool. The BERT-WWM training setting is same as BERT [37]. Except for different masking strategies, the experimental parameters are precisely the same. The experimental parameters of BERT-WWM are listed as follows: hidden size = 768, autofocus head = 12, layer number = 12, the total parameter is 110M.

In terms of assessing the performance of the approaches, we employ $Precision(P)$, $Recall(R)$ and $F-measure(F)$ as the performance evaluation indicators.

C. RESULTS AND ANALYSIS

1) COMPARATIVE EXPERIMENT OF URBAN DATA RECOGNITION MODEL

We designed a comparative experiment of SUDIR's urban data recognition model. Skip-gram in Word2vec algorithm proposed by Mikolov *et al.* [36] is selected as the comparison approach of embedding model, and embedding model construction is implemented based on the contribution of Rong X [57] and Shen [58]. Word2vec's dimension is 300. Hidden Markov Model (HMM) [21], CRF [23], IDCNN-CRF [29] and BLSTM-CRF [30] are used as baseline models of BERT-WWM+BLSTM-CRF urban data recognition model.

Comparative experimental results are shown in Tab.5. The results manifest that the performance of BERT-WWM+BLSTM-CRF is better than other baseline models. The condition indicates that the approach is applicable for Name Entity Recognition of urban text data, i.e., POI-Add and POI-Nam. Based on the experimental performance of specific models, the performance of HMM and CRF models based on statistical learning is not as top as models based on deep learning. The performance of CRF model is better than HMM model because CRF does not have the strict independence assumption of HMM. CRF can accommodate any specific information and calculate the conditional probability of the maximum output variable. The fact proves that it is reasonable to use CRF as the prediction layer of the urban data recognition model. Through the experiments, we found that the performance of BLSTM-CRF is slightly higher than IDCNN-CRF. Without considering computing resources, BLSTM-CRF can play better model performance. We utilize BLSTM-CRF as the basic model is rational. In terms of the word embedding vector pre-construction method, the experimental results manifest that BERT's word vector pre-training method is better than Word2vec method, which represents that BERT can better represent the information of text. By comparing BERT-WWM and BERT, we find that BERT-WWM model considers the whole word mask strategy to strengthen the embeddings' performance and proves the positive role of whole word mask strategy in Chinese natural language processing sequence labeling task. Moreover, depending on the results of model experiments on diverse data sets, we find that each model's effect using

TABLE 5. Comparative experiment of urban data recognition model.

Dataset	Model	POI-Add			POI-Nam			Others		
		P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
BosonNLP	Word2vec+HMM	36.07	41.59	38.64	36.87	35.58	36.21	58.22	61.13	59.64
	Word2vec+CRF	66.79	54.13	59.80	46.42	54.41	50.10	68.06	65.55	66.79
	Word2vec+IDCNN-CRF	71.56	70.60	71.07	58.38	61.77	60.03	80.58	73.98	77.13
	Word2vec+BLSTM-CRF	70.13	71.49	70.81	54.03	61.47	57.51	81.98	73.32	77.41
	BERT+BLSTM-CRF	**74.32**	73.45	73.88	60.32	61.23	60.77	84.35	74.89	79.34
	BERT-WWM+BLSTM-CRF	73.97	**75.78**	**74.86**	**65.31**	**63.12**	**64.20**	**85.31**	**77.32**	**81.12**
MSRA	Word2vec+HMM	53.75	49.93	51.77	68.10	60.04	63.82	56.29	54.73	55.50
	Word2vec+CRF	86.13	72.34	78.64	89.08	69.55	78.11	87.08	68.39	76.61
	Word2vec+IDCNN-CRF	84.51	82.06	83.27	83.15	77.72	80.34	87.40	81.12	84.15
	Word2vec+BLSTM-CRF	84.77	81.30	83.00	82.61	78.06	80.27	86.97	81.84	84.33
	BERT+BLSTM-CRF	86.34	**83.96**	85.13	84.24	80.23	82.19	88.23	84.32	86.23
	BERT-WWM+BLSTM-CRF	**88.77**	83.76	**86.19**	**85.54**	**82.97**	**84.24**	**91.72**	**85.29**	**87.89**
RMRB	Word2vec+HMM	66.17	69.97	68.02	48.14	62.12	54.24	77.30	74.15	75.69
	Word2vec+CRF	89.00	83.52	86.17	84.57	77.39	80.82	93.78	81.80	87.38
	Word2vec+IDCNN-CRF	91.23	91.29	91.26	87.75	87.09	87.42	94.33	93.88	94.10
	Word2vec+BLSTM-CRF	92.49	92.84	92.66	90.36	91.71	91.03	95.48	95.07	95.27
	BERT+BLSTM-CRF	93.12	94.23	93.67	92.13	92.41	92.27	96.24	96.35	96.29
	BERT-WWM+BLSTM-CRF	**93.89**	**95.34**	**94.61**	**94.77**	**94.03**	**94.40**	**97.13**	**97.03**	**97.08**
Chinese-NER	Word2vec+HMM	67.06	63.40	65.18	50.19	49.06	49.62	73.39	71.62	72.50
	Word2vec+CRF	89.89	77.91	83.47	81.70	67.60	73.97	88.89	67.87	76.97
	Word2vec+IDCNN-CRF	89.17	92.48	90.79	88.70	79.41	83.80	92.64	91.09	91.86
	Word2vec+BLSTM-CRF	91.72	92.07	91.90	83.41	86.32	84.84	94.66	92.33	93.48
	BERT+BLSTM-CRF	92.06	93.36	92.70	84.34	88.04	86.15	96.11	**96.32**	96.21
	BERT-WWM+BLSTM-CRF	**93.04**	**94.02**	**93.53**	**84.61**	**89.10**	**86.79**	**97.30**	95.99	**96.64**

TABLE 6. Levene's statistical significance test.

Hypothetical content	P-value	Significant difference
Kernel-based models have similar performance	0.4475	No
Deep learning-based models have similar performance	0.9863	No
Different method-based models have similar performance	$1.265 * 10^{-8}$	Yes

BosonNLP data set has not yet reached the ideal state. The reason should be that samples of BosonNLP data set is the smallest. The experimental results manifest that the amount of training data required for data recognition model should reach a certain standard.

We use Levene's test [59] to determine whether there are significant differences between the recognition models based on different technologies. Levene's test is used to test whether the variances of two or more independent samples are equal. Levene's test requires samples to be random and independent of each other. We conducted three sets of statistical significance tests. The variables of levene's are the F1 score and model category. The experimental results of the hypothetical content and p-value are shown in Tab.6. Levene's test judges a significant difference by observing p-value. If the p-value of the two experimental results is less than 0.05, they are considered to have statistical significance, otherwise they are not considered to have statistical significance.

Tab.6 shows that there is no significant difference in performance between the models based on the kernel function, and there is no significant difference in performance between the models based on the deep learning. Compared with the models based on kernel function, the models based on deep learning are significantly different in the urban text data recognition task, and the performance of models based on deep learning is superior to other models. It proves that deep learning has a positive effect on model performance. The main work of the experiments is to verify the effectiveness of BERT-WWM+BLSTM-CRF urban data recognition model.

2) EUWC OPTIMIZATION EXPERIMENT

To observe the impact of different clustering algorithms on EUWC, we designed a comparative experiment to analyze the applicability of OPTICS and DBSCAN. To facilitate the analysis of algorithms are effective and observe the effect of parameter changes on the clustering effect of web pages, we set ω^{tag} and ω^{str} to be 1. In addition, we set *MinPts* to 15. The characteristic of the DBSACN web page clustering algorithm is to determine the minimum clustering interval ε and the minimum clustering density threshold *MinPt*. If the *Minpts* value is excessively big, class-like cluster optimization cannot be accomplished. If the *Minpts* value is excessively small, a host of invalid cluster classes will be generated. The value of ε determines the similarity between the clusters. If ε is excessively big, the data in the same cluster will be unduly different. OPTICS is mainly aimed at improving the sensitivity of the input parameter ε. OPTICS and DBSCAN have the same input parameters (ε and *MinPts*). OPTICS algorithm is not sensitive to ε input, which can be observed through the decision graph clusters generated by different ε.

We map the samples in the web page dataset into a two-dimensional space according to the similarity of web

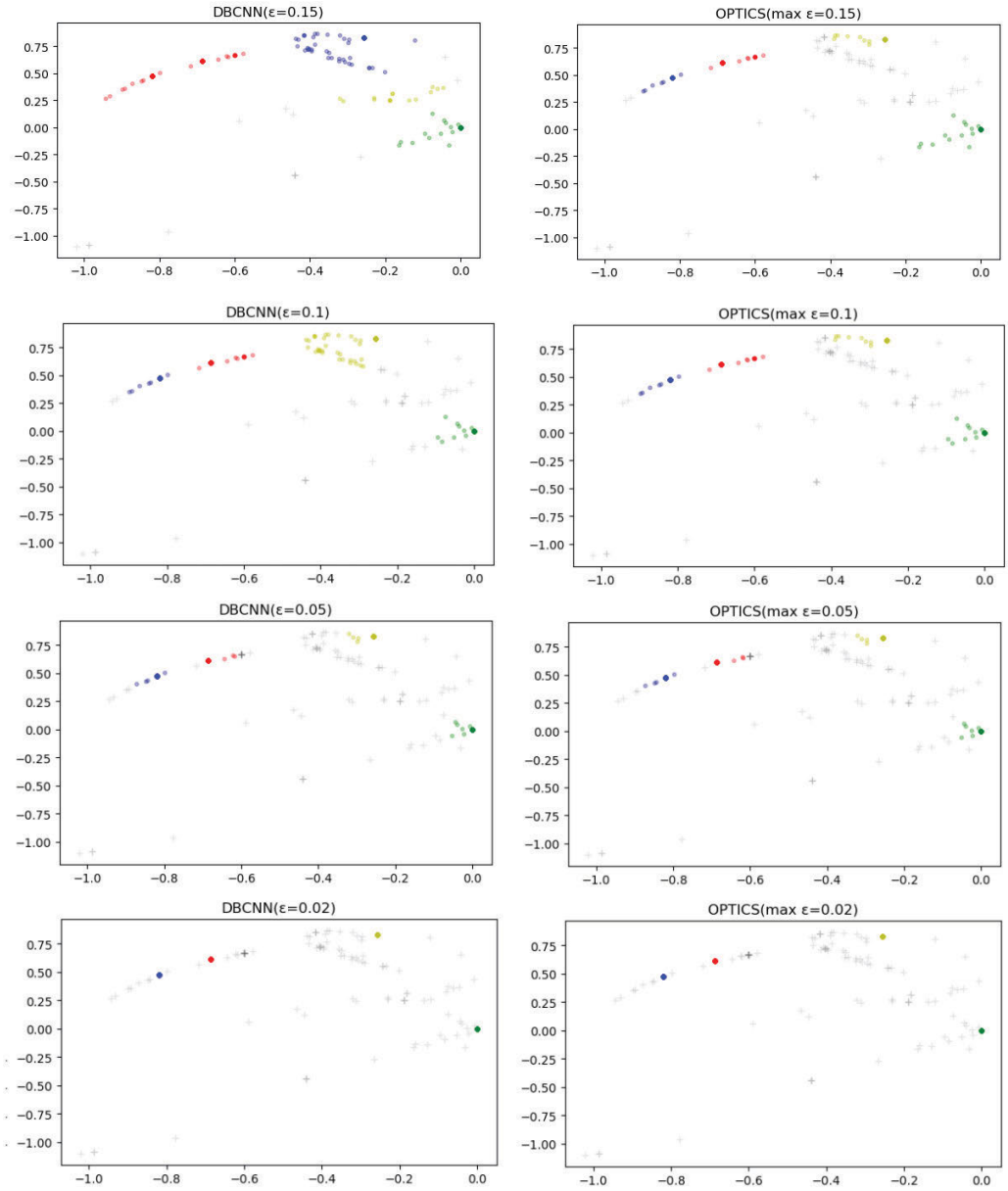

FIGURE 11. Effect of web page clustering algorithms with different parameters.

page tags and the similarity of web page structure. Fig.11 shows the visual clustering effects of web page samples when the DBSCAN and OPTICS algorithms are set with different parameters (ε). We can observe the web page clustering performance of the two algorithms by Tab.7 and Fig.11. Tab.7 shows Correct Noise Recognition Ratio (CNRR) and Clustering Accuracy (CA) results of the two web page clustering algorithms with different parameters. When ε is 0.15, DBSCAN (CNRR=13%) introduce more noise data than OPTICS (CNRR=48%), and DBSCAN's CA (62.66%) is also lower than OPTICSs (96.61%). By optimizing ε, the clustering effect of DBSCAN and OPTICS is improved, and the best clustering effect can be achieved in our experimental environment ($\varepsilon = 0.02$). It can be found that OPTICS has better performance than DBSCAN in web page clustering task, because OPTICS effectively alleviates the limitation of DBSCAN in terms of parameter sensitivity. The experimental results demonstrate that the web page clustering algorithm based on OPTICS can improve the clustering effect after optimizing parameters, and verify the feasibility and effectiveness of combining web page features and density clustering to integrate multi-source web resources.

In section III-C2, the threshold M is the selection condition for the labeling method of data on the web pages, and its value will affect the final labeling result of data. We compared the effects of different thresholds M on the recognition

TABLE 7. Effect of ε on clustering performance.

Algorithm	DBSCAN		OPTICS	
ε	CNRR	CA	CNRR	CA
0.15	13%	62.66%	48%	96.61%
0.1	39%	94.91%	63%	97.58%
0.08	65%	97.69%	69%	97.95%
0.06	73%	98.21%	75%	98.34%
0.04	96%	99.74%	96%	99.74%
0.02	100%	100.00%	100%	100%
0.01	100%	100.00%	100%	100%

TABLE 8. Performance of SUDIR with different threshold *M*.

Threshold M	POI-Add			POI-Nam		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
0.6	89.51	93.02	91.23	92.31	95.48	93.87
0.7	89.73	95.32	92.44	92.38	96.22	94.57
0.8	89.89	96.55	93.10	92.55	98.67	95.51
0.9	**90.18**	**98.36**	**93.60**	**92.78**	**99.32**	**95.94**
1.0	89.43	91.70	90.55	92.31	93.54	92.92

performance of urban text data in web pages. The experimental results are presented in Tab.8.

The results demonstrate that when the threshold M is selected reasonably, P, R and F of urban data labeling can be enhanced. Among them, the improvement of P is small, while the improvement of R is large. This is because after the introduction of EUWC, some of the False Negative (FN) samples can be corrected and converted into True Positive (TP) samples. In terms of M setting, we found that 0.9 was the best choice in the current experimental environment, and the F of POI-Nam and POI-Add reached 93.60% and 95.94% respectively. In actual Internet scenario, swarm intelligence algorithms can be used to obtain the optimal parameters more efficiently and conveniently.

3) SUDIR PERFORMANCE

In this section, a comparative experiment is performed to evaluate the performance of SUDIR and other baseline models. SUDIR uses BERT-WWM+CRF model to recognize urban data in Chinese text, and uses EUWC to introduce web page features to improve the model's ability of sensing Internet urban data. The parameter settings of EUWC are, $\varepsilon = 0.02$, $MinPts = 15$, $M = 0.9$, respectively.

Tab.9 demonstrates the experimental evaluation results of different urban data-extraction models. In comparison with the baseline methods, deep learning and web page features-based SUDIR achieved the highest P (90.18%, 92.78%), R (98.36%, 99.32%), F (93.60%, 95.94%) and the Number of sensed Urban data (NU=21310, 9856) for POI-name and POI-address recognition. After introducing EUWC into BERT-WWM+BLSTM-CRF model, both R and NU have an increase of about 6%. Since R is the true positive divided by the actual positive. Therefore, the improvement of R and NU indicates that EUWC has corrected the urban data recognized as false negatives by BERT-WWM+BLSTM-CRF model. SUDIR uses EUWC method to correct false negative samples into true positive samples to make the extracted urban data are more accurate and comprehensive. In addition, we use Levene's test to judge the significant difference between SUDIR and deep learning-based BERT-WWM+BLSTM-CRF on the sensing of urban text data. The p-value of the experimental result is 0.0146. P-value is less than 0.05, indicating that the two models are statistically different. Experiments show that combining BERT-WWM+BLSTM-CRF model with EUWC can make SUDIR achieve better results.

Tab.10 manifests an example of using SUDIR to sense urban text data from clustered web pages. It shows that BERT-WWM+BLSTM-CRF model successfully identified the POI-name urban data in the first three HTML texts, while the last sample failed. "Left eye and right eye" is a short text with ambiguous semantics, so the urban data recognition model cannot directly recognize it's category correctly. In the example, we found that SUDIR (BERT-WWM+BLSTM-CRF + EUWC) successfully sensed the

TABLE 9. Performance comparison of urban data-extraction models.

Model	POI-Add				POI-Nam			
	P(%)	R(%)	F(%)	NU	P(%)	R(%)	F(%)	NU
Word2vec+HMM	62.53	51.23	56.32	11099	65.97	60.34	63.03	6046
Word2vec+CRF	79.13	72.34	75.58	15672	81.08	69.23	74.69	6937
Word2vec+IDCNN-CRF	85.51	86.66	86.08	18775	86.15	87.72	86.93	8790
Word2vec+BLSTM-CRF	85.77	86.53	86.15	18747	87.85	88.06	87.95	8824
BERT+BLSTM-CRF	88.64	88.96	88.80	19273	90.74	92.90	91.81	9309
BERT-WWM+BLSTM-CRF	89.43	91.70	90.55	19867	92.31	93.54	92.92	9373
SUDIR (BERT-WWM+BLSTM-CRF + EUWC)	**90.18**	**98.36**	**93.60**	**21310**	**92.78**	**99.32**	**95.94**	**9856**

TABLE 10. Application examples of urban data-extraction framework. (Chinese has been translated into English).

Web page html data	Urban data	BERT-WWM+BLSTM-CRF	This Work
...Staff canteen...	Staff canteen (POI-name)	Staff canteen (POI-name)	Staff canteen (POI-name)
...Jiuduanshao (rock road store)...	Jiuduanshao (rock road store) (POI-name)	Jiuduanshao (rock road store) (POI-name)	Jiuduanshao (rock road store) (POI-name)
...Free Bar...	Free Bar (POI-name)	Free Bar(POI-name)	Free Bar (POI-name)
...Around the eyes...	Around the eyes (POI-name)	O	Around the eyes (POI-name)

urban data in the HTML tag text(POI-name).

V. CONCLUSION AND PERSPECTIVE

In this paper, we propose an approach of sensing urban text data from Internet resources based on deep learning, i.e., SUDIR. SUDIR uses Internet resources as sensors to obtain real-time, low-cost, large-scale urban data. We propose two novel works, namely the construction of a high-performance Chinese urban data recognition model and a urban data labeling methods for multi-source Internet resources scenario. On one hand, deep learning-based BERT-WWM+BLSTM-CRF urban data recognition model is proposed. This model has a better performance than other similar models in the task of Chinese urban data labeling. On the other hand, EUWC method is proposed to optimize the effect of the urban data recognition model for extracting urban text data from multi-source web resources instead of specific web resources. By correcting the urban text data wrongly labeled as the negative sample on the web pages, EUWC method can help SUDIR to extract more comprehensive and accurate urban data. Experiments verified that SUDIR has better performance than other baseline methods, and manifests the value of SUDIR in the construction of urban sensing technology.

In the following work, we will carry out data mining operations such as data cleaning, entity matching and relationship extraction for the acquired urban text data. Building on valuable urban data, a knowledge map of urban computing will be constructed. The knowledge map is utilized to describe the urban information and knowledge mined from the Internet resources to promote the construction of smart city applications in various fields, such as healthcare, education, transportation, economy, etc.

REFERENCES

[1] Z. Yu, C. Licia, W. Ouri, and Y. Hai, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014.

[2] F. Ali, S. El-Sappagh, S. M. R. Islam, A. Ali, M. Attique, M. Imran, and K.-S. Kwak, "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Future Gener. Comput. Syst.*, vol. 114, pp. 23–43, Jan. 2021.

[3] D. S and I. M, "Hybridization approach to classify big data using social Internet of Things," *Bonfring Int. J. Softw. Eng. Soft Comput.*, vol. 9, no. 2, pp. 31–35, Apr. 2019.

[4] S. E. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris, "Location extraction from social media: Geoparsing, location disambiguation, and geotagging," *ACM Trans. Inf. Syst.*, vol. 36, no. 4, pp. 1–27, 2018.

[5] P. Nesi, G. Pantaleo, and M. Tenti, "Ge(o)Lo(cator): Geographic information extraction from unstructured text data and Web documents," in *Proc. 9th Int. Workshop Semantic Social Media Adaptation Personalization*, Nov. 2014, pp. 60–65.

[6] K. Stock, "Mining location from social media: A systematic review," *Comput., Environ. Urban Syst.*, vol. 71, pp. 209–240, Sep. 2018.

[7] H.-M. Chuang, C.-H. Chang, T.-Y. Kao, C.-T. Cheng, Y.-Y. Huang, and K.-P. Cheong, "Enabling maps/location searches on mobile devices: Constructing a POI database via focused crawling and information extraction," *Int. J. Geographical Inf. Sci.*, vol. 30, no. 7, pp. 1405–1425, Jul. 2016.

[8] Y. Gu, Z. Qian, and F. Chen, "From Twitter to detector: real-time traffic incident detection using social media data," *Transp. Res. C, Emerg. Technol.*, vol. 67, pp. 321–342, Jun. 2016.

[9] L. Hennig, P. Thomas, R. Ai, J. Kirschnick, H. Wang, J. Pannier, N. Zimmermann, S. Schmeier, F. Xu, and J. Ostwald, "Real-time discovery and geospatial visualization of mobility and industry events from large-scale, heterogeneous data streams," in *Proc. ACL Syst. Demonstrations*, 2016, pp. 37–42.

[10] H. S. Al-Olimat, K. Thirunarayan, V. Shalin, and A. Sheth, "Location name extraction from targeted text streams using gazetteer-based statistical language models," 2017, *arXiv:1708.03105*. [Online]. Available: http://arxiv.org/abs/1708.03105

[11] J. Gelernter and W. Zhang, "Geocoding location expressions in Twitter messages: A preference learning method," *J. Spatial Inf. Sci.*, no. 9, pp. 37–70, Dec. 2014.

[12] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2014, pp. 43–52.

[13] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza, "On the accuracy of hyper-local geotagging of social media content," in *Proc. 8th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2015, pp. 127–136.

[14] A. Tigunova, J. Lee, and S. Nobari, "Location prediction via social contents and behaviors: location-aware behavioral LDA," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 1131–1135.

[15] K. Ishida, "Estimation of user location and local topics based on geotagged text data on social media," in *Proc. IIAI 4th Int. Congr. Adv. Appl. Informat.*, Jul. 2015, pp. 14–17.

[16] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[17] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM SIGMOD. Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999.

[18] J. Gao, M. Li, C.-N. Huang, and A. Wu, "Chinese word segmentation and named entity recognition: A pragmatic approach," *Comput. Linguistics*, vol. 31, no. 4, pp. 531–574, Dec. 2005.

[19] L. A. Souza, C. A. Davis, Jr., K. A. V. Borges, T. M. Delboni, and A. H. F. Laender, "The role of gazetteers in geographic knowledge discovery on the Web," in *Proc. 3rd Latin Amer. Web Congr. (LA-WEB)*, Nov. 2005, p. 9.

[20] X. Yu, W. Lam, S.-K. Chan, Y. K. Wu, and B. Chen, "Chinese NER using CRFs and logic for the fourth SIGHAN bakeoff," in *Proc. Sixth SIGHAN Workshop Chin. Lang. Process.*, 2008, pp. 1–4.

[21] D. Woodward, J. Witmer, and J. Kalita, "A comparison of approaches for geospatial entity extraction from Wikipedia," in *Proc. IEEE 4th Int. Conf. Semantic Comput.*, Sep. 2010, pp. 402–407.

[22] D. Inkpen, J. Liu, A. Farzindar, F. Kazemi, and D. Ghazi, "Location detection and disambiguation from Twitter messages," *J. Intell. Inf. Syst.*, vol. 49, no. 2, pp. 237–253, Oct. 2017.

[23] S. Sarawagi and W. W. Cohen, "Semi-Markov conditional random fields for information extraction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1185–1192.

[24] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 17, 2020, doi: 10.1109/TKDE.2020.2981314.

[25] C. Dong, J. Zhang, C. Zong, M. Hattori, and H. Di, "Character-based LSTM-CRF with radical-level features for Chinese named entity recognition," in *Natural Language Understanding and Intelligent Applications*. Cham, Switzerland: Springer, 2016, pp. 239–250.

[26] Y. Zhang and J. Yang, "Chinese NER using lattice LSTM," 2018, *arXiv:1805.02023*. [Online]. Available: http://arxiv.org/abs/1805.02023

[27] L. Liu, J. Shang, F. F. Xu, X. Ren, H. Gui, J. Peng, and J. Han, "Empower sequence labeling with task-aware neural language model," 2017, *arXiv:1709.04109*. [Online]. Available: http://arxiv.org/abs/1709.04109

[28] S. Tomori, T. Ninomiya, and S. Mori, "Domain specific named entity recognition referring to the real world by deep neural networks," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2016, p. 1–7.

[29] E. Strubell, P. Verga, D. Belanger, and A. McCallum, "Fast and accurate entity recognition with iterated dilated convolutions," 2017, *arXiv:1702.02098*. [Online]. Available: http://arxiv.org/abs/1702.02098

[30] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*. [Online]. Available: http://arxiv.org/abs/1508.01991

[31] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," 2016, *arXiv:1603.01354*. [Online]. Available: http://arxiv.org/abs/1603.01354
[32] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," 2016, *arXiv:1603.01360*. [Online]. Available: http://arxiv.org/abs/1603.01360
[33] W. Wang, F. Bao, and G. Gao, "Mongolian named entity recognition with bidirectional recurrent neural networks," in *Proc. IEEE 28th Int. Conf. Tools with Artif. Intell. (ICTAI)*, Nov. 2016, pp. 495–500.
[34] J. M. Giorgi and G. D. Bader, "Transfer learning for biomedical named entity recognition with neural networks," *Bioinformatics*, vol. 34, no. 23, pp. 4087–4094, 2018.
[35] M. N. A. Ali, G. Tan, and A. Hussain, "Boosting arabic named-entity recognition with multi-attention layer," *IEEE Access*, vol. 7, pp. 46575–46582, 2019.
[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781
[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[39] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," 2017, *arXiv:1705.00108*. [Online]. Available: http://arxiv.org/abs/1705.00108
[40] L. Logeswaran and H. Lee, "An efficient framework for learning sentence representations," 2018, *arXiv:1803.02893*. [Online]. Available: http://arxiv.org/abs/1803.02893
[41] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*. [Online]. Available: http://arxiv.org/abs/1801.06146
[42] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," OpenAI, San Francisco, CA, USA, Tech. Rep., 2018. [Online]. Available: https://openai.com/blog/language-unsupervised/
[43] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu, "Pre-training with whole word masking for chinese BERT," 2019, *arXiv:1906.08101*. [Online]. Available: http://arxiv.org/abs/1906.08101
[44] E. Michelioudakis, A. Artikis, and G. Paliouras, "Semi-supervised online structure learning for composite event recognition," *Mach. Learn.*, vol. 108, no. 7, pp. 1085–1110, Jul. 2019.
[45] W. Shi, X. Liu, and Q. Yu, "Correlation-aware multi-label active learning for Web service tag recommendation," in *Proc. IEEE Int. Conf. Web Services (ICWS)*, Jun. 2017, pp. 229–236.
[46] R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma, "Learning block importance models for Web pages," in *Proc. 13th Conf. World Wide Web (WWW)*, 2004, pp. 203–211.
[47] J. Feng, P. Haffner, and M. Gilbert, "A learning approach to discovering Web page semantic structures," in *Proc. 8th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2005, pp. 1055–1059.
[48] L. Niu, Q. Tang, A. Veeraraghavan, and A. Sabharwal, "Learning from noisy Web data with category-level supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7689–7698.
[49] J. Gibson, B. Wellner, and S. Lubar, "Adaptive Web-page content identification," in *Proc. 9th Annu. ACM Int. Workshop Web Inf. Data Manage. (WIDM)*, 2007, pp. 105–112.
[50] V. Loia, W. Pedrycz, and S. Senatore, "Semantic Web content analysis: A study in proximity-based collaborative clustering," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 6, pp. 1294–1312, Dec. 2007.
[51] D. Lin and X. Wu, "Phrase clustering for discriminative learning," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. (AFNLP: ACL-IJCNLP)*, vol. 2, 2009, pp. 1030–1038.
[52] Y. Fang, V. W. Zheng, and K. C.-C. Chang, "Learning to query: Focused Web page harvesting for entity aspects," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 1002–1013.
[53] Y. Hong and S. Kwong, "Learning assignment order of instances for the constrained K-Means clustering algorithm," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 39, no. 2, pp. 568–574, Apr. 2009.
[54] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," in *Proc. 11th Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2002, pp. 515–524.
[55] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, 1996, pp. 226–231.
[56] K. Santhisree, A. Damodaram, S. V. Appaji, and D. Nagarjunadevi, "Web usage data clustering using dbscan algorithm and set similarities," in *Proc. Int. Conf. Data Storage Data Eng.*, Feb. 2010, pp. 220–224.
[57] X. Rong, "Word2vec parameter learning explained," 2014, *arXiv:1411.2738*. [Online]. Available: http://arxiv.org/abs/1411.2738
[58] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du, "Analogical reasoning on chinese morphological and semantic relations," in *Proc. 56th Annu. Meeting Comput. Linguistics*, vol. 2, 2018, pp. 138–143.
[59] H. Levene, *Robust Tests for Equality of Variances in Contribution to Probability and Statistics*, 1st ed. Palo Alto, CA, USA: Stanford Univ. Press, 1960.

CHAORAN ZHOU was born in Jilin, China, in 1994. He received the B.S. and M.S. degrees from the Changchun University of Science and Technology, in 2016 and 2019, respectively, where he is currently pursuing the Ph.D. degree. His research interests include urban computing, natural language processing, and data mining.

JIANPING ZHAO was born in Yushu, Jilin, China, in 1964. He received the Ph.D. degree from the Changchun Institute of Optics and Fine Mechanics, in 1986. Since 1986, he has been a Teacher with the Computer Science and Technology College, Changchun University of Science and Technology, where he is currently a Professor. His research interests include data mining and machine learning.

CHENGHAO REN was born in Yanbian, Jilin, China, in 1994. He received the M.S. degree from the Changchun University of Science and Technology, in 2019. He is currently pursuing the Ph.D. degree with Jilin University. His main research interests include cloud computing, stream processing, and wireless sensor networks.