*Computers, Materials & Continua* CMC, vol.64, no.3, pp.2057-2073, 2020

# Information Classification and Extraction on Official Web Pages of Organizations

**Jinlin Wang[1], Xing Wang[1, *], Hongli Zhang[1], Binxing Fang[1], Yuchen Yang[1] and Jianan Liu[2]**

**Abstract:** As a real-time and authoritative source, the official Web pages of organizations contain a large amount of information. The diversity of Web content and format makes it essential for pre-processing to get the unified attributed data, which has the value of organizational analysis and mining. The existing research on dealing with multiple Web scenarios and accuracy performance is insufficient. This paper aims to propose a method to transform organizational official Web pages into the data with attributes. After locating the active blocks in the Web pages, the structural and content features are proposed to classify information with the specific model. The extraction methods based on trigger lexicon and LSTM (Long Short-Term Memory) are proposed, which efficiently process the classified information and extract data that matches the attributes. Finally, an accurate and efficient method to classify and extract information from organizational official Web pages is formed. Experimental results show that our approach improves the performing indicators and exceeds the level of state of the art on real data set from organizational official Web pages.

**Keywords:** Web pre-process, feature classification, data extraction, trigger lexicon, LSTM.

## 1 Introduction

Authors are encouraged to use the template for Microsoft Word, to prepare the final version of their manuscripts and facilitate typesetting. Authors may elect to submit two versions of their manuscript, one for the printed version of the journal, and the other for the on-line version of the journal. Illustrations in color are allowed only in the on-line version of the journal. Organizations, such as companies and schools, are collectives of people working together to achieve specific goals. As a force to lead social development, most organizations have their official Web pages, which contain numerous personal, departmental, business information and shows the significance and correlations. In the application scenario of supervision and investigation, it is essential to obtain real-time and useful information from organizations. Compared to others, the official Web pages of the organization have the

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150006, China.

[2] China Electronic Equipment System Engineering Company, Beijing, 100039, China

[*] Corresponding Author: Xing Wang. Email: wxhit@hit.edu.cn.

Received: 30 April 2020; Accepted: 15 May 2020.

following characteristics: 1) The content ministers to the public perception. 2) The time of publication is real-time. 3) The information contained is authoritative. Mining the value of information through the essential pre-processing contribute to the works such as analyzing organization relationships and predicting the development trend.

To effectively analyze the information in official Web pages of organizations, preprocessing complex multi-source Web pages should be an essential pre-step. There exist several problems in the current research as follows:

1) Public data sets related to these Web pages are lacking [Pasternack and Roth (2009); Hernández, Rivero, Ruiz et al. (2014)], which requires crawling by researchers.

2) Every organizational Web page has its layout and content [Gautam and Kumar (2013); Saleh, Abulwafa and Al Rahmawy (2017)], which contains much irrelevant information.

3) The organizational information is of varying lengths [Thamviset and Wongthanavasu (2014); Wang, Ma, Zhang et al. (2008)], which makes it challenging to map attributes.

In general, there is still no effective method to deal with the official Web pages of organizations.

In this paper, the information classification and extraction method for the official Web pages of organizations has been proposed, which processes Web pages into unified attributed information. Active information blocks are obtained based on the analysis of valid characters. By combining the structural and content features of Web pages with the specific model, this method accurately completes the information classification. For organizational information of varying lengths, two processing methods based on trigger rule and LSTM are proposed to extract the classified information. The architecture of the classification and extraction method is shown in Fig. 1.

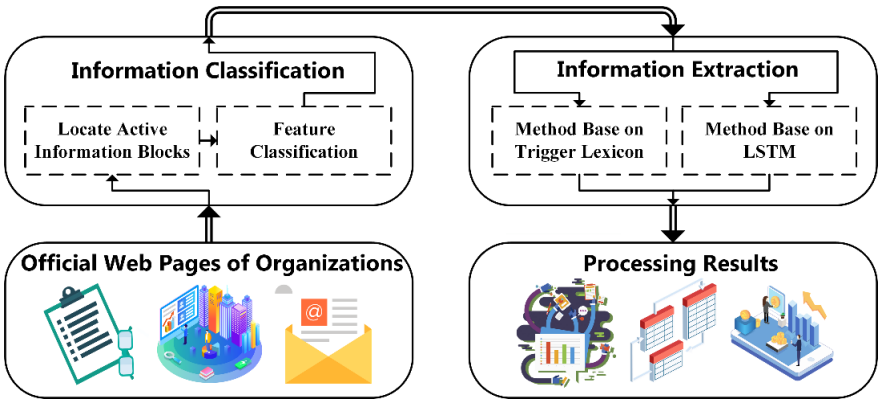


**Figure 1:** Framework of the proposed processing method

The main contributions of the paper are summarized as follows:

1) It is an effective research to design classification and extraction methods specifically on the official Web pages of the organizations.

2) After locating the active information blocks of the official Web pages, the specific structural and content features for classification are proposed and proved effective by the experiments.

3) Innovatively combined with the trigger lexicon and LSTM, methods for extracting

types of information for organizational Web pages are designed, which have better performance than the similar methods.

The remaining of this paper is organized as follows. Recent studies are described in Section 2. Information classification and extraction method for official Web pages are presented in Sections 3 and 4. Experimental work is given by using real official Web page data set in Section 5. Finally, discussion and conclusion are presented in Section 6.

## 2 Related work

In this section, recent studies are divided into the classification and extraction of Web page information.

### *2.1 Classification on Web page information*

Classification of Web pages plays an essential role in the process of content mining. Hashemi [Hashemi (2020)] surveyed the proposed methodologies in the literature, but also traces their evolution and portrays different perspectives toward this problem. Recent studies mainly focus on the following two aspects: classic text classification and classification based on Web page features.

In the aspect of classic text categorization, Gautam et al. [Gautam and Kumar (2013)] improved the mutual information formula by adding the class correlation balance factor. Then they applied it to the feature weighting algorithm, which significantly enhances the effect of text categorization. Saleh et al. [Saleh, Rahmawy and Abulwafa (2017)] proposed *ONBC*, a novel strategy for vertical Web page classification, which employs several Web mining techniques, and depends mainly on proposed multi-layered domain ontology. Xu et al. [Xu, Yu and Qi (2018)] presented a novel sensitive information classification algorithm and topic tracking algorithm for Tibetan Web pages contents. However, Web pages are semi-structured HTML documents, which have the structural features of layout and rendering in addition to text information. Therefore, this classification method based on the classic text has some limitations in the application of practical problems.

Based on traditional text classification methods, many studies work around Web page features. Pasternack et al. [Pasternack and Roth (2009)] looked for the largest sub sequence of Web page and got the content of Web information by segmentation with the proposed method *EMSS*. Hernández et al. [Hernández, Rivero, Ruiz et al. (2014)] proposed an unsupervised URL-based Web page classification method. By constructing some URLs, it classifies categories of Web pages and matches the classified Web pages with patterns. Saleh et al. [Saleh, Abulwafa and Al Rahmawy (2017)] proposed a new centroid-based model to solve the class imbalance problem, which learns from the training data and weighs each category to indicate the data distribution of the corresponding categories. Wang et al. [Wang and Qu (2017)] proposed a new method of Web text classification based on the improved CNN and SVM, using the CNN model with the five-layer network structure to extract text feature and then classify and predict by using SVM. Onan [Onan (2016)] presented a comparative analysis of four different feature selections and four different ensemble learning methods based on four different base learners. The experimental results

of these methods indicate that feature selection and ensemble learning can enhance the predictive performance of classifiers in Web page classification.

### *2.2 Extraction on Web page information*

According to the extraction methods, recent studies can be divided into three categories: pattern-based, domain ontology-based and machine learning-based methods.

Thamviset et al. [Thamviset and Wongthanavasu (2014)] designed a semi-supervised extraction system and proposed *ERSP*, a method of information extraction based on repetitive patterns. Moreover, this system applies the topic tree clustering algorithm to cluster the target data record and create extraction patterns. Li et al. [Li, Jiang, Xu et al. (2017)] built a Web information retrieval matching and structure extraction model based on search engine, which realized the algorithm of locating and automatically extracting multi-Web news information with regular expression. However, when the structure of Web pages changes, the extraction rules need to be modified.

Compared with the pattern-based approach, the domain ontology-based method *ClusTex* proposed by Ashraf et al. [Ashraf, Özyer and Alhajj (2008)] has its specificity and limitations. Domain ontology is the collective knowledge recognized by people in a specific domain and can also be learned by many Web pages. Moreover, *ClusTex* takes more effort to construct. Zhang et al. [Zhang and Ding (2015)] introduced Web page segmentation into the stage of Web page pretreatment, by analyzing the ontology-based Web information extraction technology. Vigneshwari et al. [Vigneshwari and Aramudhan (2015)] develop a model based on the multiple constructed ontologies from the mutual information, which experimental results shows a healthy improvement for quick access of useful data from a huge information resource like the Internet.

Web page information extraction based on machine learning is to utilize learning models such as conditional random field method [Li (2012)], SVM (support vector machine) method [Wu, Hu and Liang (2014)] and multimodal learning [Gong, Wang and Peng (2017)] to transform Web page information extraction into model optimization. The advantage of this method is that it can better adapt to the change of the structure of the Web page, meanwhile the cost is high.

Therefore, to classify the official Web page information effectively, the combination of location and feature is adopted. In the process of extracting the classified data, the appropriate method is applied according to the classification features, which is a reasonable way to classify and extract the official Web page information.

## 3 Classification on official Web pages

In this section, an information block location method based on valid characters is proposed to complete the information classification work. After that, the structural and content classification features are summarized in Tab. 1 to classify the active blocks in official Web pages.

### *3.1 Location of active information blocks*

The block location is an effective way to identify the active ingredients in Web pages. As

for the official organizational Web pages, valid characters are mostly articles, prepositions and conjunctions, which is the key to express a sentence smoothly. Therefore, it is feasible to utilize the valid character to locate active information blocks in pages. The existence of valid characters indicates that the texts are semantically complete and smooth sentences. The more valid characters the DOM node contains, the more feasible it is to be the active information block.

**Table 1:** Summary of classification features

| Notation | Signification | Range |
|---|---|---|
| MPV | Maximum percentage of valid characters in subtags | 0-1 |
| MDP | Maximum percentage difference of valid characters in subtags | 0-1 |
| NSW | Number of subtags | Integer |
| CPP | Character information character ratio | 0-1 |
| NSP | Maximum number of occurrences of the same name | Integer |
| NBI | Number of business information | Integer |
| CPD | Department information character ratio | 0-1 |
| NSD | Number of occurrences of the same department name | Integer |
| CPB | Business information character ratio | 0-1 |

The valid characters can be regarded as an attribute of the DOM tree node to combine number. As shown in algorithm *CountChars*, an attribute named *validChars* is added. When a node is a leaf of the DOM tree, the text content of the node is judged to contain valid characters. After processing, a Web page file with attribute *validChars* can be obtained, in which DOM node shows the number of valid characters. The number of valid characters represents the possibility where the node locates in the information block.

**Algorithm 1** *CountChars*

**Input:** DOM node $N$
**Output:** The Number of valid characters *validChars*

```
1: Initialize()
2: if N.children ≠ ∅ then
3:     for C ∈ N.children do
4:         CountChars(C)
5:     end for
6:     N.parent.validChars += N.validChars
7: else
8:     if N consists of characters then
9:         N.parent.validChars += GetNonSpaceLength(N)
10:    end if
11: end if
```

**Definition 1.** $i$ represents a node in the DOM tree. $C_i$ represents the number of valid characters of $i$, and $j$ represents the subnode of $i$. The maximum character ratio of the subnode (*MPV*) is defined as follows.

$$MPV = \max_{j \in child(i)} \frac{c_j}{c_i} \tag{1}$$

*MPV* indicates the importance of child nodes in parent nodes. As shown in the block location algorithm *Major*, the DOM tree with attribute *validChars* is added to check all

the subnodes, find the maximum value node, and select the maximum value node *maxNode*. If the value of *MPV* is less than the threshold *k*, *Major* stops and outputs the current node. Otherwise, *Major* is recursively performed on *maxNode* until the Web page information block is determined.

**Algorithm 2** *Major*

**Input:** DOM node $N$, threshold value $k$
**Output:** Subnode $maxNode$ with the largest value of $validChars$

```
maxNode = null
maxValidChars = 0
totalValidChars = 0
for C ∈ N.children do
    totalValidChars = totalValidChars + C.validChars
    if C.validChars >maxValidChars then
        maxValidChars = C.validChars
        maxNode = C
    end if
end for
if maxNode ≠ null then
    if maxValidChars/totalValidChars<k then
        Return maxNode
    else
        Major(maxNode)
    end if
end if
```

### *3.2 Structural features of official Web pages*

After analyzing the organization's official Web pages with the location of active information blocks, the information that needs to be classified is obtained. These active information blocks have features, which mainly reflected in the structure and content of Web pages. Considering the valid characters and sub tags, three structural features are given as follows.

1) The maximum proportion of valid characters in subtags of Web page information (*MPV* is shown in Eq. (1)).
2) The maximum difference in the proportion of valid characters in subtags of Web page information (*MDP*).

$$MDP = \max_{j \in child(i)} \frac{c_j}{c_i} - \min_{j \in child(i)} \frac{c_j}{c_i} \tag{2}$$

As shown in Eq. (2), $i$ represents the node with DOM number. $C_i$ represents the number of valid characters of DOM node $i$, and $j$ represents the child of DOM node $i$.

3) The number of sub tags in Web page information (*NSW*).

$$NSW = \sum_{s \in childnode} 1 \tag{3}$$

In Eq. (3), *childnode* denotes a set of sub tag nodes and $s$ indicates specific result.

As shown in Fig. 2, the cumulative distribution of structural features in the experimental data set proves the validity.

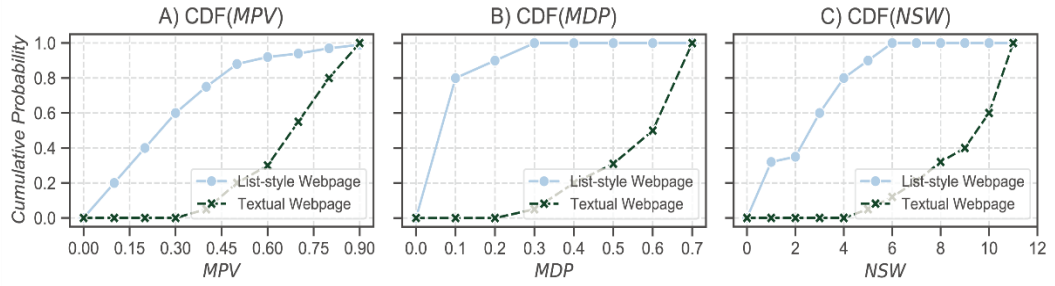


**Figure 2:** Cumulative distribution of structural features

### *3.3 Content features of official Web pages*

In the research of this paper, the classification is mainly oriented to the three types of information as follows. 1) Personal information comes from the employees of the organization, such as name, age, and school. 2) Department information comes from the organizational departments, such as name, e-mail, and functions. 3) News information comes from organizational news, such as name, description, and knowledge. Considering the proportion of different types of information determines the effect of classification, here list six content features in the official organizational Web pages as follows.

1) The character proportion of personal information (*CPP*)

$$CPP = \frac{\sum_{s \in P} \sum_{r \in s} r}{SUM} \tag{4}$$

As shown in Eq. (4), $P$ denotes the result set of regular matching of personal information class and $s$ represents each specific result. $r$ denotes the characters in $s$, and $SUM$ represents the total number of valid characters in the Web page information block.

2) The number of the same personal name (*NSP*)

$$NSP = \sum_{s \in S} 1 \tag{5}$$

In Eq. (5), $S$ represents the set of identical names in the main body information block and, $s$ represents each specific result. The most significant difference between the list type single-person and the list type multi-person information pages is that there will be multiple same personal names.

3) The number of business information (*NBI*)

$$NNI = \sum_{s \in N} 1 \tag{6}$$

As shown in Eq. (6), $N$ denotes the regular matching result set of news and $s$ indicates each concrete result.

4) The character proportion of departmental information (*CPD*)

$$CPD = \frac{\sum_{s \in D} \sum_{r \in s} r}{SUM} \tag{7}$$

In Eq. (7), $D$ represents the regular matching result set of the departmental classes, $s$ denotes each specific result and $r$ represents the characters in $s$. $SUM$ denotes the total number of valid characters in the Web page information block.

5) The number of the same departmental name ($NSD$)

$$NSD = \sum_{s \in K} 1 \tag{8}$$

As shown in Eq. (8), $K$ denotes the same set of department names in the main body information block, and $s$ indicates each specific result. The most significant difference between the list-style single-department and the list-style multi-department information pages is that there will be more same departmental names.

6) The character proportion of business information ($CPB$)

$$CPB = \frac{\sum_{s \in C} \sum_{r \in s} r}{SUM} \tag{9}$$

In Eq. (9), $C$ denotes the regular matching result set of business information, $s$ represents each specific result and $r$ denotes the characters in $s$. $SUM$ represents the total number of valid characters in the Web page information block.

The cumulative distribution of the experimental data set shown in Fig. 3 proves the validity of content features.

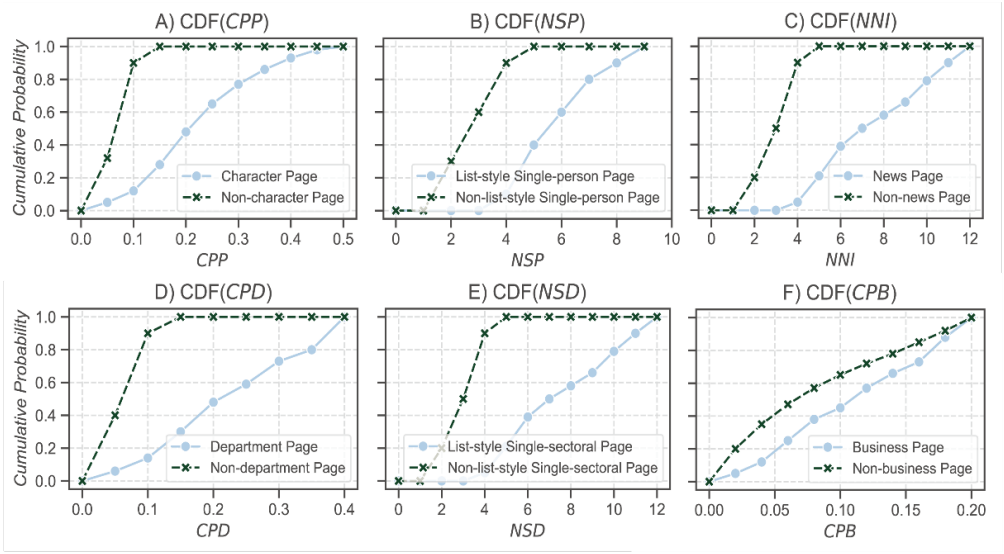


**Figure 3:** Cumulative distribution of content features

## 4 Extraction on official Web pages

After locating the active blocks and presenting the features, the information of the organization's official Web page is classified. Information with a short length can be extracted by rules, while others require experience in the extraction process. Two methods based on trigger lexicon and LSTM are proposed to extract attributes of the classified organizational information as follows.

### *4.1 Extraction method based on trigger lexicon*

It has difficulty to identify personal and departmental information in official Web pages. *TRIE*, a method of information extraction based on trigger lexicon is proposed. The trigger lexicon of target extraction information is established, and the process of rule matching is applied to extract the trigger lexicon.

**Algorithm 3** $TRIE$

**Input:** Text $S$ to be extracted, triggering lexicon $Q$, rule base $W$
**Output:** Structured information text $R$

```
textSpilt = Spilt(S)
R = ∅
for i ∈ textSpilt do
    while j ∈ i and q ∈ Q do
        if RecordMatch(j, q) then
            for w ∈ W[j] do
                if RecordMatch(i, w) then
                    R.append(i)
                    break
                end if
            end for
        end if
    end while
end for
```

In the information that needed to be extracted, there always exist some trigger words such as verbs and nouns. Constructed by expert knowledge, the trigger lexicon to extract information is shown in Tab. 2. It is essential to locate the sentences where the information to be obtained. The corresponding rules are matched by searching the rule base, which is a collection of regular expressions for an attribute that has excellent extraction effects for types of known classifications.

**Table 2:** Example of trigger lexicon

| Category | Trigger Word |
|---|---|
| Name | Mrs., Ms., Mr., Sir |
| Phone | Telephone, Phone, Mobile Phone. |
| School | Graduate, Graduation, Institution, School. |
| Function | Responsible, Bear, Undertake, Support, Provides. |
| Department | Department, System, Unit, Part, Troop. |

### *4.2 Extraction method based on LSTM*

The work scope of organizations is reflected in their business information from official Web pages, which has several features as follows: 1) no apparent triggers; 2) no structural similarity; 3) lives in a long text. Due to the long length of business information, natural language processing is required. RNN (Recurrent Neural Network) can vectorize sentences and find optimal solutions at high speed. However, the gradient disappears after the training time increases in training RNN model using the business information because RNN can only remember part of the data. LSTM [Hochreiter and Schmidhuber (1997)] is a cyclic neural network model with control units such as the input, output, and forget gate. It makes the weight parameters of the cyclic structure continuously change during the learning process and adds long-term dependency based on RNN. LSTM can deal with the long-term dependence in the previous text and judges the next after understanding the past. Due to the features mentioned in the business information, LSTM has advantages over other methods in processing business information.

As shown in Fig. 4, an information extraction method based on LSTM is proposed. Firstly, the text containing a long length of information is selected and divided into clauses. The information, such as business name and description, is identified by expert labeling. The performance of the LSTM model is gradually improved by continuous training. The business information of organizations in different fields labeled by experts is utilized as the training set. The processing model contains word2vector, vector joint, LSTM computing, and softmax selection, which is utilized to perform probability calculation on the business information to obtain the label with the highest matching value. After calculating by the model, business information and their corresponding labels are obtained.

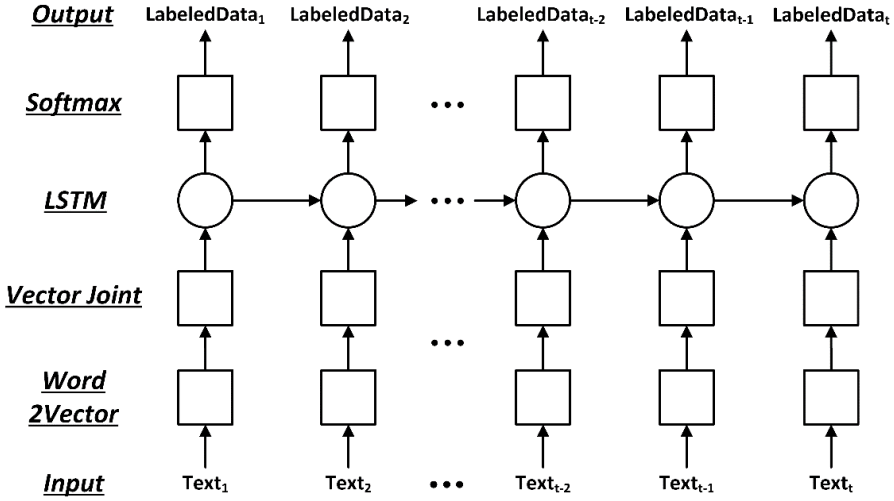


**Figure 4:** Schematic diagram of the LSTM model

This paper proposes *LSTMIE*, a statistics-based algorithm that utilizes the number and the proportion of labels in each statement to extract business information. *LSTMIE* only needs $O(n)$ times to achieve the extraction result.

**Algorithm 4** *LSTMIE*

**Input:** Text $T$ with annotation, threshold value $k$
**Output:** Text $R$ with business information

```
textSpilt = Spilt(T)
R = ∅
for i ∈ textSpilt do
    if IncludeContinuouslyBusinessName(i) or #BusinessProfiles(i)
  / #Words(i) >k then
        R.append(i)
    end if
end for
```

## 5 Experimental study

After introducing the implementation of the above methods, this section will explain the experimental part, including data set, environment and experimental results.

### *5.1 Data set and environment*

Official Web pages of 900 organizations in various industries were collected to test the effectiveness of the above methods. The collection source contains 300 NGOs (non-governmental organizations), companies and schools each. Besides, three categories were obtained through the summary of the Web pages. 20% of each category data set was randomly selected as the experimental data set. The categories and subcategories were labeled by expert knowledge. The summary of the experimental data set is shown in Tab. 3.

**Table 3:** Summary of experimental data set

| Category | Subcategory | Extraction Method | Quantity |
|---|---|---|---|
| Person | List-style multi-person | Based on trigger lexicon | 268 |
| | Textual single-person | Based on LSTM | 1323 |
| | List-style single-person | Based on trigger lexicon | 812 |
| | Institutional profile | Based on trigger lexicon | 154 |
| | List-style institutional profile | Based on trigger lexicon | 882 |
| Department | List-style multi-department | Based on trigger lexicon | 256 |
| | List-style single-department | Based on trigger lexicon | 672 |
| | Textual single-department | Based on LSTM | 4010 |
| News | List-style multi-business | Based on trigger lexicon | 876 |
| | List-style single-business | Based on trigger lexicon | 6700 |
| | Textual single-business | Based on LSTM | 12810 |
| | List-style news | Based on trigger lexicon | 471 |
| | Textual news | Based on LSTM | 6403 |

All experiments were conducted on CentOS 8.0 with an Intel i7 CPU@3.4 GHz, 16 GB of memory, and an SSD hard disk with a capacity of 480 GB.

### *5.2 Information classification on official Web pages*

In this experiment, three structural and six content features of official Web pages were summarized. The training data set was obtained by randomly selecting 80% of the experimental data set, meanwhile the remaining part formed the test set.

#### *5.2.1 Selection of classification models*

The BP neural network model, C4.5 decision tree and SVM algorithm were applied to construct the classification model by using the same training data and the features mentioned above. The comparison results of the classification performance of the three models are shown in Fig. 5. With relatively high accuracy, recall and *F*-measure value for the official Web page classification, the SVM model is chosen to make the classification model according to the comparison analysis.

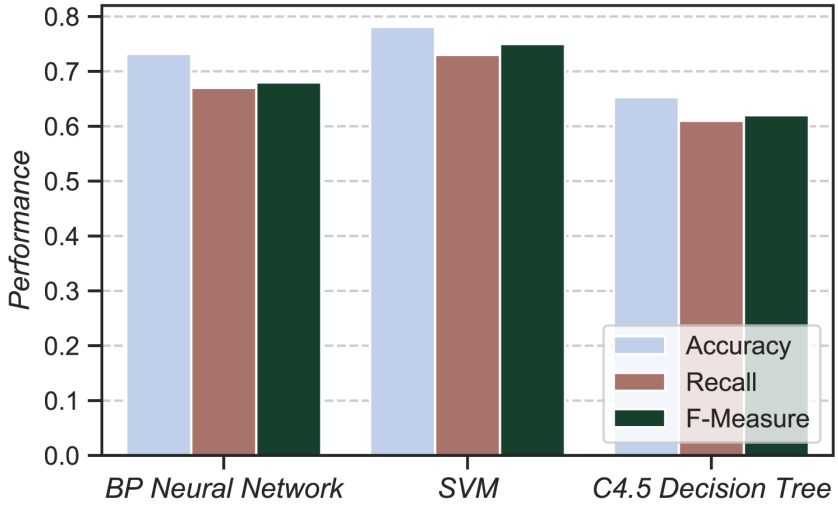


**Figure 5:** Comparative performance of classification models

#### *5.2.2 Comparative analysis of indicator performance*

After selecting SVM as the classification model, the proposed features were applied to process the information. The indicator performance comparison was made for scenarios that consider all proposed features, without any feature, the baseline method *EMSS* and *ONBC*. As shown in Tab. 4, all nine features have positive effects on Web classification, which proves the rationality. Besides, it is noteworthy that the accuracy decreases more when feature *MPV*, *MDP*, *NSW* and *CPB* are missing. The classification effect of Web pages is significantly reduced when the feature *CPB* is lacking. Compared with the baseline methods, the proposed method performs better in terms of parameter performance and time cost.

**Table 4:** Indicator performance with proposed features and baselines

| Feature | TP Rate | FP Rate | Precision | Recall | *F*-Measure | Time Cost per Thousand Pages |
|---|---|---|---|---|---|---|
| *All Features* | **0.831** | **0.062** | **0.821** | **0.832** | **0.826** | **57.986** |
| *EMSS* | 0.736 | 0.066 | 0.711 | 0.729 | 0.716 | 73.379 |
| *ONBC* | 0.782 | 0.065 | 0.764 | 0.752 | 0.776 | 62.233 |
| *Without MPV* | 0.523 | 0.241 | 0.476 | 0.517 | 0.496 | - |
| *Without MDP* | 0.625 | 0.125 | 0.519 | 0.552 | 0.338 | - |
| *Without NSW* | 0.541 | 0.170 | 0.483 | 0.523 | 0.502 | - |
| *Without CPP* | 0.690 | 0.107 | 0.633 | 0.674 | 0.653 | - |
| *Without NSP* | 0.724 | 0.097 | 0.667 | 0.684 | 0.675 | - |
| *Without NNI* | 0.618 | 0.165 | 0.529 | 0.578 | 0.552 | - |
| *Without CPD* | 0.730 | 0.083 | 0.694 | 0.725 | 0.709 | - |
| *Without NSD* | 0.758 | 0.075 | 0.707 | 0.738 | 0.722 | - |
| *Without CPB* | 0.548 | 0.205 | 0.493 | 0.528 | 0.510 | - |

### *5.3 Information extraction on official Web pages*

The official Web pages have been divided into personal, departmental and business information after classification. In this experiment, two proposed methods were applied in extraction according to the text length. We applied *TRIE* to personal and departmental information and *LSTMIE* to business information. The method *ERSP* and *ClusTex* were applied as the baselines.

#### *5.3.1 Analysis of personal information extraction*

As shown in Fig. 6A), *TRIE* performs slightly better than the baseline methods in most categories. However, in terms of the extraction of personal names, the accuracy of *TRIE* performs slightly worse than the baseline methods. The reason is that some official Web pages of organizations do not contain the name in trigger lexicon, which causes a slight loss. Compared with the baseline method *ERSP* and *ClusTex*, *TRIE* performs better in dealing with personal information.

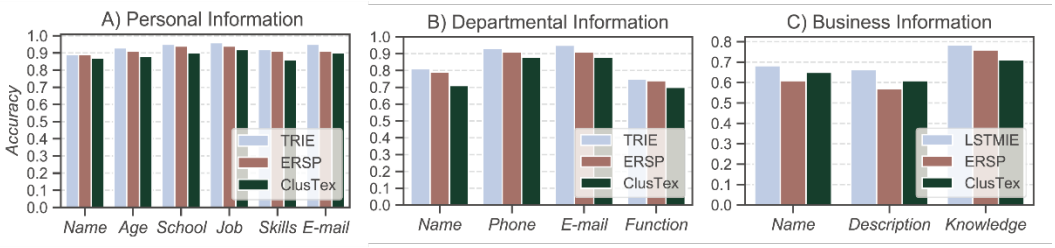


**Figure 6:** Accuracy performance of *TRIE*, *LSTMIE* and the baseline methods

#### *5.3.2 Analysis of departmental information extraction*

As shown in Fig. 6B), the accuracy of the four categories of information extraction performs higher than the baseline methods, which proves the effectiveness of algorithm

*TRIE*. It is noteworthy that the two methods of extracting function and name information perform poorly. Therefore, the reason for the low accuracy of departmental names is that there exist no trigger words in the official Web pages of organizations. Besides, the accuracy performs poor because the information style of the departmental function is flexible. It brings challenges to cover all cases with extraction rules.

*5.3.3 Analysis of business information extraction*

The LSTM model was trained with labeled text containing business information. After obtaining the model, the other non-labeled text was processed, and the business information extraction was completed. The accuracy of the extraction results shown in Fig. 6C) was compared according to three types of information. In the extraction of business information, the accuracy of *LSTMIE* performs higher than that of the baseline methods. Compared with the baseline method, *LSTMIE* has better performance in dealing with the official organizational Web pages information with flexible forms, various changes and low similarity between Web pages.

*5.3.4 Analysis of three types of organizations*

By applying *TRIE* and *LSTMIE*, the information was extracted from the Web pages of schools, NGOs, and companies in the experimental data set. As shown in Fig. 7, the accuracy performance of three types of organizations in different information categories is consistent. The reason why schools perform better is that their organizational structure is relatively fixed. In contrast, NGOs are slightly obscure in introducing their personal and business information.

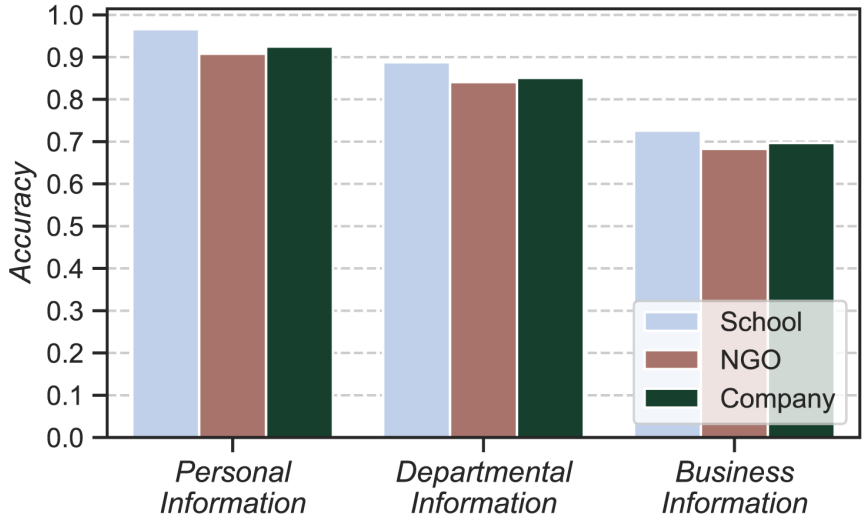


**Figure 7:** Accuracy performance of three types of organizations

*5.3.5 Efficiency of the proposed TRIE and LSTMIE*

*TRIE* and *LSTMIE* have achieved excellent accuracy in the experimental data set. To evaluate the efficiency of our methods, *ERSP* and *ClusTex* were applied in the

experiment. Fig. 8 shows the results of the comparative efficiency experiment. It can be observed that *TRIE* and *LSTMIE* have the advantage of efficiency cost in the processing of data sets of various categories of the same size.

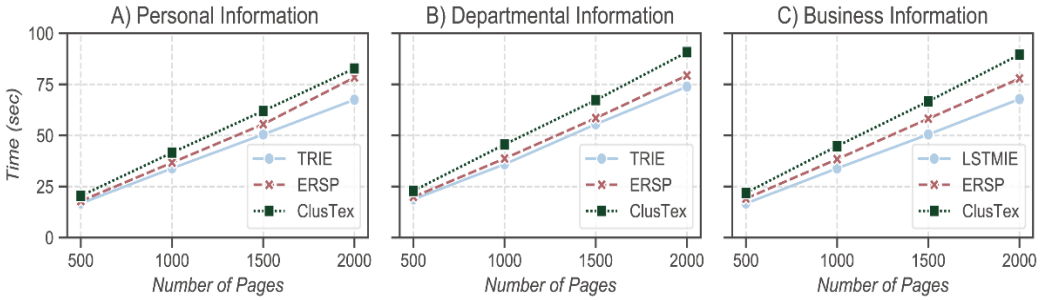


**Figure 8:** Efficiency of *TRIE*, *LSTMIE* and the baseline methods

## 6 Discussion and conclusion

It can be observed that the proposed method shows the characteristics in the experiments as follows:

1) The proposed structural and content features pay more attention to the intuitive representation of the active information blocks, prove the effectiveness, and improve the performance of Web classification.

2) The comparison with the baseline algorithm proves the validity of the selected classification model and features.

3) Collaboration between *TRIE* and *LSTMIE* can be applied to extract attributes of the classified information from types of organizations. Compared with existing works, the proposed method has achieved better classification and extraction performance in official Web pages of organizations.

In this paper, we have proposed the information classification and extraction method for the official Web pages of organizations. After locating the active information blocks of Web pages, the content and structural features were summarized. The specific method was applied to construct the model to classify the Web pages. Two extraction methods were proposed for types of Web information, which are respectively based on trigger lexicon and LSTM. Experiments showed that our method performs better than existing methods in terms of effectiveness and efficiency.

In the future, it is necessary to expand the size of the trigger lexicon and reduce costs with parallelization. The focus will also be placed on further enhancing efficiency and discovering more stable features.

**Funding Statement:** This work was supported by the National Key Research and Development Program of China (Nos. 2016QY03D0501, 2017YFB0803300), the National Natural Science Foundation of China (Nos. 61601146, 61732022) and Sichuan Science and Technology Program (No. 2019YFSY0049).

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

**Ashraf, F.; Özyer, T.; Alhajj, R.** (2008): Employing clustering techniques for automatic information extraction from HTML documents. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 38, no. 5, pp. 660-673.

**Gautam, J.; Kumar, E.** (2013): An integrated and improved approach to terms weighting in text classification. *International Journal of Computer Science Issues*, vol. 10, no. 1, pp. 310-314.

**Gong, D.; Wang, D. Z.; Peng, Y.** (2017): Multimodal learning for Web information extraction. *The 25th ACM International Conference on Multimedia*, pp. 288-296.

**Hashemi, M.** (2020): Web page classification: a survey of perspectives, gaps, and future directions. *Multimedia Tools and Applications*, pp. 1-25.

**Hernández, I.; Rivero, C. R.; Ruiz, D.; Corchuelo, R.** (2014): CALA: an unsupervised URL-based Web page classification system. *Knowledge Based Systems*, vol. 57, pp. 168-180.

**Hochreiter, S.; Schmidhuber, J.** (1997). Long short-term memory. *Neural Computation*, vol. 9, no. 8, pp. 1735-1780.

**Li, C.** (2012): *Adaptive Web Information Extraction Method Research Based on Ontology (PhD Thesis)*. University of Science and Technology of China.

**Li, J.; Jiang, G.; Xu, A.; Wang, Y.** (2017): The automatic extraction of Web information based on regular expression. *Journal of Software*, vol. 12, pp. 180-188.

**Onan, A.** (2016): Classifier and feature set ensembles for Web page classification. *Journal of Information Science*, vol. 42, no. 2, pp. 150-165.

**Pasternack, J.; Roth, D.** (2009): Extracting article text from the Web with maximum subsequence segmentation. *Proceedings of the 18th International Conference on World Wide Web*, pp. 971-980.

**Saleh, A. I.; Abulwafa, A. E.; Al Rahmawy, M. F.** (2017): A Web page distillation strategy for efficient focused crawling based on optimized Naïve Bayes (ONB) classifier. *Applied Soft Computing*, vol. 53, pp. 181-204.

**Saleh, A. I.; Rahmawy, M. F. A.; Abulwafa, A. E.** (2017): A semantic based Web page classification strategy using multi-layered domain ontology. *World Wide Web*, vol. 20, no. 5, pp. 939-993.

**Thamviset, W.; Wongthanavasu, S.** (2014): Information extraction for deep Web using repetitive subject pattern. *World Wide Web*, vol. 17, no. 5, pp. 1109-1139.

**Vigneshwari, S.; Aramudhan, M.** (2015): Web information extraction on multiple ontologies based on concept relationships upon training the user profiles. *International Conference on Artificial Intelligence and Evolutionary Algorithms in Engineering Systems*, pp. 1-8.

**Wang, Z.; Qu, Z.** (2017): Research on Web text classification algorithm based on improved CNN and SVM. *IEEE 17th International Conference on Communication Technology*, pp. 1958-1961.

**Wu, Q.; Hu, L.; Liang, J.** (2014): Web information extraction based on block importance model and 2D conditional random fields. *Journal of Nanjing University*, vol. 50, no. 1, pp. 79-85.

**Xu, G.; Yu, Z.; Qi, Q.** (2018): Efficient sensitive information classification and topic tracking based on Tibetan Web pages. *IEEE Access*, vol. 6, pp. 55643-55652.

**Zhang, J.; Ding, W.** (2015): An improved ontology-based Web information extraction. *International Conference of Educational Innovation through Technology*, pp. 37-41.