ATLANTIS PRESS

Advances in Engineering Research, volume 104

7th International Conference on Mechatronics, Control and Materials (ICMCM 2016)

Document Structure Identification Method Based on Conditional Random Field

Lei Yang[1, a],Tian Yingai[2, b], Li Ning[3,c], Gao Xiaolong[4, d]

[1]School of Computer Science, Beijing information Science and Technology University, Beijing, China

[2]School of Computer and Communication Engineering, University of Science and Technology, Beijing, China

[a]466279552@qq.com, [b]tianyingai@bistu.edu.cn, [c]lining@bistu.edu.cn, [d]773920925@qq.com

Keywords: document structure identification; sequence labeling; CRF

Abstract: On the basis of deep analysis on the structural features and heading features of documents, it has researched the classification method based on templates and the classification method based on statistics as well as the sequence labeling method based on CRF (Conditional Random Field), then proposed to treat document structure identification as sequential data labeling, built CRF training model with feature templates and finally realized document structure identification upon training model with existing way of supervision learning. Experimental results show that identifying paragraph roles from document sequence structure helps to ensure a higher accuracy and it also owns certain fault-tolerant ability. Besides, it is observed that using CRF for many times could further improve the accuracy of identification.

Introduction

No matter whether it is professional researchers' academic papers or the undergraduates' or graduates' thesis papers, ill-formed format is inevitable. When any thesis suffers a nonstandard format, it is the precondition and basis of understanding document structure and checking document format to try to identify the document structure as accurately as possible. CRF is a sequence labeling method based on multi-feature recognition. Even if some features of a document structure is not standard, CRF can correctly identify the document structure from all features as a whole, which shows certain fault-tolerant ability. Thus, when CRF is applied in the identification of document structure, it could further improve the accuracy of document structure identification and then lay a good foundation for document format check.

A document is composed of many paragraphs and each paragraph plays a different role, such as heading, table captions, figure captions and text body. Therefore, each paragraph plays a different role. All paragraph roles form a sequence in order, which shows the structural features of this document. Document headings are in different levels and different paragraphs also show the heading roles in different levels. The relevance between the heading roles in different levels and the relationship between paragraph roles constitute document structure, and the above information forms the features and rules of document structure [1]. For example, after the "first-level heading", it can be a "second-level heading" or "text body", but cannot be a "third-level heading". Considering document structure as sequential data and putting document structure into a sequence for identification, then labeling results can be adjusted with the relations between sequential data. In this way, when structure of the document to be tested is not standard, paragraph role can be correctly acquired with the relationship between paragraphs. Such automatic identification refers to adopting an inartificial way to analyze the paragraphs in the document, then to identifying the role of each paragraph and finally get the roles of all paragraphs and the roles of all headings of the entire paper.

This paper mainly contributes to helping machines to understand document structure automatically and correctly through improving the accuracy of identification when document structure is ill-formed. For the identification of document structure, it is often identified in the

method based on templates, namely document structure can be identified through defining templates to describe document structure and paragraph roles. This method suffers an obvious limitation, namely it cannot identify document structure without templates and the structures that are not involved in templates cannot be identified as well. Thus, each type of documents to be identified needs a template, which has a great limitation. To such defects of the identification method based on templates, a method based on statistics is adopted to identify document structure in this paper, i.e. a data model containing document structures will be used to identify document structure. The method based on statistics are generally divided into two categories[2]: one is based on classification, such as Bayesian Model, SVM (Support Vector Machine) and Maximum Entropy; the other is based on sequence labeling, such as HMM (Hidden Markov Model), MEMM (Maximum Entropy Markov Model) and CRF (Conditional Random Field). During identification, document structure can be treated as a sequence of paragraph roles in certain order. Therefore, a method based on sequence labeling, i.e. CRF is adopted for the identification of document structure in this paper.

Part II will briefly introduce the research work related to identification of document structure; Part III will offer the method to identify document structure with CRF; Part IV will analyze the results of document structure identification based on CRF; and financially, it comes to conclusion.

Related Researches

At present, there are few researches related to the identification of document structure and the methods to identify document structure mainly include the method based on templates, on VSM and corresponding improvement, n-gram, etc. Wang Shuaiqun et al. adopted the method based on templates in the automatic check system for the papers in .NET [3]. This refers to writing document typesetting into templates and then comparing paragraphs with templates to realize the identification of document structure. A good point of this method is it is very simple, but the templates need to be generated manually, so that this method suffers a poor universality. Song Haosu et al. adopted the method based on VSM, which had improved the automation level of document structure identification [4]. This refers to treating paragraph roles as a spatial vector and treating the paragraph to be identified as another spatial vector, then its paragraph role can be got with the distance between the two vectors, so as to realize the identification of document structure. But in this method, a paragraph cannot compare with many paragraph roles at one time and the weight of the vector of document structure is given upon experimental statistics. As to identification of the document structure in a new format, it needs to make another experiment and new weighting, which still suffers a poor universality from this point of view. Peng Xin et al. adopted the improved method based on VSM to identify paragraph labeling. This method is applicable for solving the vector similarity of different types of variables [5], which allows to compare with many paragraph roles at the same time, and dimensionless similarity measurement is used to identify paragraph role, but it is still not universal enough. Then, after improving this method, Peng Xin et al. proposed a method based on n-gram [6]. It builds a 1-gram dictionary for paragraph labeling format and posting list and then rates the logic labeling paragraph to be identified to judge which logic labeling is this paragraph most likely to be.

However, all the above method based on VSM and corresponding improvement as well as the method based on n-gram only identify paragraph roles with the features of the paragraph itself, but can’t judge current paragraph role in accordance with roles of context paragraphs, namely it cannot analyze paragraph roles from the integral structure. If the role is identified with the paragraph itself alone, then when it has used the wrong paragraph features, it is easy to cause identification error. On the basis of the above researches, it puts forward a sequence labeling method based on CRF in this paper, which adjusts labeling results with sequence structure relationship and then identifies paragraph roles from structure level. In this way, it could avoid the identification error caused for using nonstandard paragraph features, so as to further improve the accuracy.

Document Structure Identification Based On CRF

CRF is a conditional probability model based on sequential data labeling and it's a discriminant model. In this paper, it treats document structure as the sequential data composed of paragraph roles and then labels the paragraph role with CRF. CRF is able to express long distance dependence and overlapping feature, it could well solve the label (classification) bias and all features can be globally normalized to obtain a globally optimal solution [7]. As the relationship between paragraph roles influences document structure, it could label the paragraph role globally with the help of the advantages of CRF and on the basis of identifying paragraph roles with features of the paragraph itself.

Conditional probability formula defined by CRF is shown in formula (1):

$$P(y|x,\lambda) = \frac{1}{Z(x)} \exp\left(\sum_i \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i)\right) \quad (1)$$

In the above formula, Z(x) refers to normalizing factor, λ_j refers to feature weight and f_j refers to state characteristics or transfer characteristics. While observing Sequence x and Model λ, make sequence labeling on the Sequence y to be tested and a most possible labeling sequence can be obtained finally. As to a document, it needs to observe the state characteristics of sequence, composed of textural features, format features and contextual features of paragraphs as well as the transfer characteristics of sequence which is the transfer relationship between state characteristics. Identification based on CRF refers to building a CRF model with feature templates and then identifying the document to be tested with this model.

In document structure identification, if CRF model is well built directly influences the accuracy of identification and the key to modeling lies in the selection of state characteristics and transfer characteristics. Here in this paper, the textural features, such as heading number and number of words, format features, such as font, font size and bold/normal as well as the contextual features, such as four paragraphs before and four paragraphs after are selected as state characteristics; change of the state characteristics between the paragraph, such as font of the paragraph before is in "bold" and that of the paragraph after is "SongTi" is selected as transfer characteristics. In the next part, it will describe selection and use of the state characteristics and transfer characteristics of document structure in details.

Feature Templates. In a document, paragraphs play different roles, which can be identified logically, i.e. with contextual features or identified from the form of manifestation of paragraphs, i.e. with textural features and format features. With deep research of document structure and induction of heading and paragraph features, it adopts the following state features for modeling in this paper, including heading number NUMi, number of words NOWi, font FONi, word size WSZi, bold BOLi/normal NORi, paragraph before BEFi, paragraph after AFTi, paragraph role before PAR-i and paragraph role after PAR+i. The subscript i of each state feature is integer, representing certain state features of the paragraph at corresponding position. For example, NUM0 represents number of current paragraph, NUM-1 represents number of the paragraph before; NUM+1 represents number of the paragraph after. Specific template is shown in the Tab.1.

Tab.1.State Feature Template

State characteristics	Feature template
Text features	NUM-1, NOW-1, NUM0, NOW0, NUM+1, NOW+1
Format feature	FON-1, WSZ-1, BOL-1(NOR-1), BEF-1,AFT-1, FON0, WSZ0, BOL0(NOR0), BEF0, AFT0 FON+1, WSZ+1, BOL+1(NOR+1), BEF+1, AFT+1
Context feature	PAR-4, PAR-3, PAR-2, PAR-1, PAR+1, PAR+2, PAR+3, PAR+4

Features of the paragraph itself also may help to identify paragraph role to a certain extent and the feature change between paragraphs also may have positive effects on identification. Thus, change relation between state features is added at modeling as shift feature. For example, after the first-level heading, it often comes the second-level heading, but not the third-level heading, and from feature, after the feature of the first-leveling heading, it is more likely to have the feature of the second-level heading. Thus, such feature shift between paragraphs is concluded as shift feature template. Specific template is shown in the Tab.2.

Tab.2 Shift Feature Template

State characteristics	Feature template
Text features	NUM_{-1}/NUM_0, NOW_{-1}/NOW_0,
	NUM_0/NUM_{+1}, NOW_0/NOW_{+1},
	$NUM_{-1}/NUM_0/NUM_{+1}$, $NOW_{-1}/NOW_0/NOW_{+1}$
Format feature	FON_{-1}/FON_0, WSZ_{-1}/WSZ_0, $BOL_{-1}(NOR_{-1})/BOL_0(NOR_0)$,$BEF_{-1}/BEF_0$,
	AFT_{-1}/AFT_0,
	FON_0/FON_{+1},WSZ_0/WSZ_{+1},$BOL_0(NOR_0)/BOL_{+1}(NOR_{+1})$,$BEF_0/BEF_{+1}$,
	AFT_0/AFT_{+1},
	$FON_{-1}/FON_0/FON_{+1}$, $WSZ_{-1}/WSZ_0/WSZ_{+1}$,
	$BOL_{-1}(NOR_{-1})/BOL_0(NOR_0)/BOL_{+1}(NOR_{+1})$, $BEF_{-1}/BEF_0/BEF_{+1}$,
	$AFT_{-1}/AFT_0/AFT_{+1}$
Context feature	$PAR\text{-}_1/PAR_{+1}$,
	$PAR\text{-}_2/PAR\text{-}_1/PAR_{+1}/PAR_{+2}$,
	$PAR\text{-}_3/PAR\text{-}_2/PAR\text{-}_1/PAR_{+1}/PAR_{+2}/PAR_{+3}$,
	$PAR\text{-}_4/PAR\text{-}_3/PAR\text{-}_2/PAR\text{-}_1/PAR_{+1}/PAR_{+2}/PAR_{+3}/PAR_{+4}$

Before document identification, it needs to extract these features first, but for a document that hasn't been preprocessed, its contextual features cannot be extracted. Thus, two feature models are made for having contextual features or not. The first model is made for no contextual features. When contextual features cannot be extracted from the initial state of a document, this model identifies the document structure with textural features and format features to get its basic structure. The second model is made for contextual features are available. When contextual features can be extracted from the document, i.e. contextual features are available to identify the basic structure, this model identifies the document structure with textural, format and contextual.

Identify Document Structure. After modeling, use model to identify document structure. For the document to be identified, when the document is identified for the first time, contextual paragraph roles cannot be extracted as contextual features, and change of contextual roles also may influence the identification of the role of current paragraph. Considering the above two factors, identification of document structure is divided into three stages.

The first stage refers to initial identification of document structure. This textural features and format features are used to make paragraph identification to identify paragraph roles of all paragraphs and form an initial structure sequence. In the second stage, the paragraph roles identified in the previous stage are used as contextual features and then used together with textual features and format features to identify paragraphs, namely while identifying the i paragraph, paragraph roles from the i±1 paragraph to the i±4 paragraph are used as the contextual features of the i paragraph to make role identification for the i paragraph together with its textural features and format features. As contextual features are added, it is more likely to change the results received in the first stage. This is to say, the document structure sequence identified in the second stage still has unstable factors. Thus, it plans to make adjustment in the third stage. In this stage, the paragraph roles identified in the second stage are used as contextual features to make identification and the received structure will be the final document structure. Contextual features of a document are local structure features and it can accurately identify current paragraph with local structural features. In other

words, the identification results of current paragraph is the optimal solution on local structure. Then, from integral structure of the document, it aims to receive globally optimal solution with the local optimal one, namely to receive the optimal result of document structure identification. Specific process is shown in Code 1.

Code 1

```
int count = 1;//Computational identification rounds
Document doc;//Get the document to be identified
Document doc_respon;//Document recognition results
 while(count<=3){
        if(count != 1){
            doc = doc_respon;//If it is not the first round of identification, the document to be identified as a result of the previous round of identification
        }
        Extracting text features from doc;
        Extracting format features from doc;
        if(count==2 && count==3){
        Extracting context features from doc;
        }
        Using CRF to identify the structure of doc document;
        The formation of the recognition results document doc_respon;
        count++;
  }
```

Analysis Of Experimental Results

In order to test accuracy and anti-interference of the algorithm proposed in this paper during document structure identification, the thesis papers from undergraduates' thesis library are selected as training and experimental data. In the final experimental results, it will offer the most possible paragraph role of each paragraph and the probability of each paragraph role in accordance with the identification results. Fig.1 shows an example of identification results of a document structure.

In the above case, format of the second-level heading "3.1" is misused as third-level heading. After extracting its document features and making CRF identification, this wrong heading was still correctly identified as a second-level heading and the probability for this heading to be correctly identified as second-level heading is as high as 99.6389%.

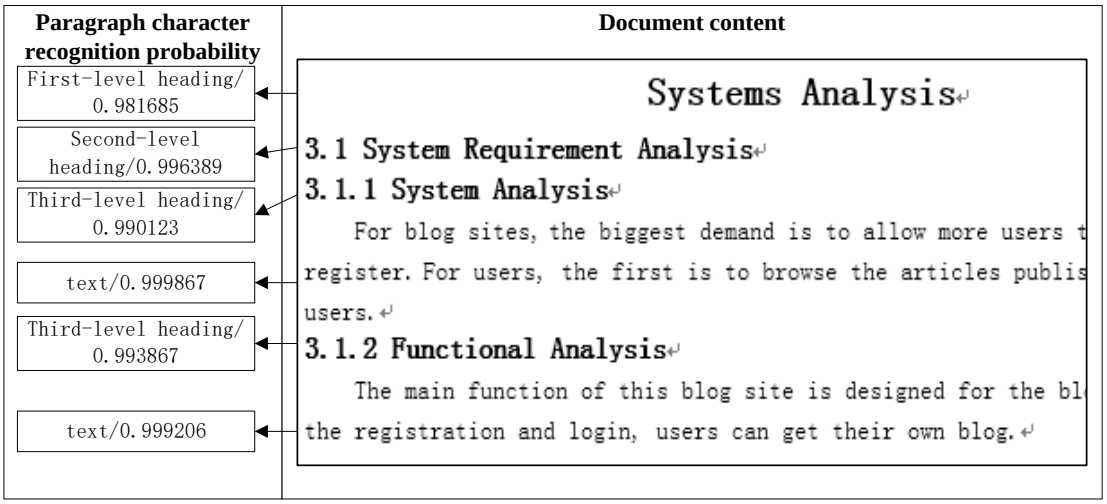

Fig.1 Example of Identification Results

As it requires not that much for undergraduates' thesis papers, nonstandard format often happens. In this experiment, first, select undergraduates' papers as training data to train CRF model. Among these training data, label correct paragraph role of each paragraph manually and it allows nonstandard format. Second, identify about 1,000 headed paragraphs from undergraduates' papers to be tested with the CRF model and the rate of correct identification is 94.84%. Specific experimental data are shown as those received with CRF method in Tab.3. It can be figured out from Tab.3 that the first-level headings owning more obvious features can be better identified and the rate of correct identification of the second-level and third-level headings is relatively low, but it is still able to realize a good effect.

Tab.3 Comparison of Experimental Data

Statistical item name	Conditional random fields method				Based on n-gram method			
	Total item quantity	Correct identification number	Number of error identification	Accuracy	Total item quantity	Correct identification number	Number of error identification	Accuracy
Heading paragraph	1008	956	52	94.84%	1008	890	118	88.29%
First-level heading	125	125	0	100%	125	125	0	100%
Second-level heading	438	413	25	94.29%	438	390	48	89.04%
Third-level heading	445	418	27	93.93%	445	375	70	84.27%

From the above experimental data and compared with n-gram method, corresponding results are shown in Table.3. It indicates that CRF method is better than n-gram method. It is because on the basis of format features and textural features, CRF method also adds contextual features in identification, which has improved the accuracy and showed advantages of the CRF that has integrated multiple features on sequential data labeling.

Conclusion

In this paper, through researching CRF and summarizing the features of document structure, it has proposed a document structure identification method based on CRF and proves its strategy to be feasible with experimental data. With CRF method, it not only can identify documents, but also can get rid of the limitations of templates, and different training data will make different models, namely even machine can help to understand multiple document structures automatically, which improves the adaptability of such identification strategy and ensures a high identification accuracy as well.

CRF adopts local structural features to identify document structure. It is applicable for the condition that document structure owns format definition in general and allows the format definition to be inaccurate and nonstandard. It owns certain anti-interference to document structure identification. But if document format is seriously nonstandard, it cannot realize an ideal identification effect at present.

In the future, it plans to make researches and improvements for CRF's insufficient anti-interference to improve CRF's ability of expressing document structure. Besides, hope this idea of using local features to identify document structure could be applied in document structure check. During the identification of document structure, it helps to adjust current paragraph label through identifying structural features to make it meet standard document structures, so as to realize the purpose of document structure check.

Acknowledgement

This work is supported by National Natural Science Foundation of China(No.61672105); the National High-tech R&D Program (863 Program No.2015AA015403) and the General Program of

Science and Technology Development Project of Beijing Municipal Education Commission (No.KM201511232013).

Reference

[1] Li Ning, Liang Qi, Shi Yunmei. The function of format information in document understanding [J]. Journal of Beijing Information Science and Technology University, 2012, 06:1-7.

[2] Yang Jinfen, Yu Qiubin, Guan Yi, Jiang Zhipeng. An Overview of Research on Electronic Medical Record Oriented Named Entity Recognition and Entity Relation Extraction [J]. ACTA AUTOMATICA SINICA, 2014, 08:1537-1562.

[3] Wang Shuaiqun, Xia Bin, Kong Wei. Automatic checking system for paper format based on.NET. Proceedings of the twenty-first National Conference on computer technology and Applications (CACIS 2010) and the second National Conference on security technology and Application [C]. 2010

[4] Song Haosu, Li Ning, Zhang Wei. The application of VSM model in the identification of document structure [J]. Journal of Beijing Information Science and Technology University,2011,26(6):67~69

[5] Peng Xin, Li Ning. Improved VSM algorithm for judging paragraph logic label [J]. Journal of Beijing Information Science and Technology University, 2014, 06:19-24.

[6] Zhao Lin, Li Ning, Peng Xin. Logical structure reconstruction of re-flowable document based on directed graph [J]. COMPUTER ENGINEERING AND DESIGN, 2016, 05:1239-1244.

[7] Yan Tingyi. Chinese Semantic Role Labeling Based on Conditional Random Fields [D]. Beijing University of Posts and Telecommunications, 2010.

[8] Raj, Abhinav. Word level language identification and back-transliteration. ACM International Conference Proceeding Series, v 05-07-Dec-2014, p 74-79, December 5, 2014, FIRE 2014 - Post-Proceedings of the 6th International Workshop of the Forum for Information Retrieval Evaluation.

[9] M. C. Gokul Chittaranjan, Kalika Bali. Word-level language identification using crf: Code-switching shared task report of msr india system. In Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 73-79. Association for Computational Linguistic, 2014.

[10] Zhou Yucan, Hu Qinghua, Liu Jie, Jia Yuan. Combining heterogeneous deep neural networks with conditional random fields for Chinese dialogue act recognition. ELSEVIER SCIENCE BV, v 168, p 408-417, NOV 30 2015.

[11] Bhuyan, M. K.; Kumar, D. Ajay; MacDorman, Karl F. A novel set of features for continuous hand gesture recognition. JOURNAL ON MULTIMODAL USER INTERFACES . v 8-4, p 333-343, OCT 2014.

[12] Klink S, Dengel A, Kieninger T. Document structure analysis based on layout and textual features. In Proc. of International Workshop on Document Analysis Systems [C]. 2000. 99~111

[13] Li Juan. Method for checking document layout format based on template [J]. Beijing: Beijing Information Science and Technology University, 2012

[14] He Yanxiang, Liu Jianbo, Sun Songtang, Wen weidong. Product reviews sentiment classification in Micro-blog based on cascaded conditional random field [J]. Journal of Shandong University (Natural Science), 2015, 11:67-73.

[15] Fang Yan, Zhou Guodong. Word Structure Analysis Based on Cascaded CRFs [J]. Journal of Chinese Information Processing. 2015,04:1-7+24.