IEICE TRANS. INF. & SYST., VOL.E102–D, NO.2 FEBRUARY 2019

LETTER

Personal Data Retrieval and Disambiguation in Web Person Search

Yuliang WEI[†], ***Member***, **Guodong XIN**[†], **Wei WANG**[†], **Fang LV**[†], ***and*** **Bailing WANG**[†a)], ***Nonmembers***

SUMMARY Web person search often return web pages related to several distinct namesakes. This paper proposes a new web page model for template-free person data extraction, and uses Dirichlet Process Mixture model to solve name disambiguation. The results show that our method works best on web pages with complex structure.

key words: *sequential block model, deep learning, web extraction, name disambiguation*

1. Introduction

In web search, 15-21% of the queries contain person names [1], while the returned result often contains web pages related to several distinct namesakes. The task of finding the web pages related to the specific person of interest still needs to be filtered by the user. In order to optimize name search results, the Web Person Search (WePS) task is proposed [2]. The task is to cluster search results for a name according to the different people that share the same one. Meanwhile, the WePS-1, WePS-2, and WePS-3 test sets have become the standard benchmark [3], the latest test set WePS-3 is proposed in 2010. Since 2012, Web2.0 technology has developed rapidly. The web page format of the WePS-3 is greatly different from the current. Compared to the web page in WePS-*, current web page has complex structure which contains a lot of irrelevant information, such as ads, recommendations, dynamic components, etc. While the algorithms using WebPS-* as the test sets do not consider the influence of irrelevant information in the current web page. Webpage information extraction usually uses template matching or text extraction. While there is no uniform extraction template for web pages returned by search engines, and according to our statistics, only nearly 60% of personal data is in the text area. The WePS method for solving complex structure web pages has not been solved so far.

In this paper, a new web page model (Sequence Block Model, SBM) is proposed for template-free person data extraction, and then Dirichlet Process Mixture Model (DPMM) [4] is used for web pages clustering according to the specific person. SBM is inspired by the Vision-based Page Segmentation (VIPS) model [5], and web pages are segmented into sequence blocks by SBM. Different from VIPS, SBM does not require web page rendering, so it can be applied to large-scale web processing. Then a sequence prediction algorithm based on deep learning is used to extract the personal data in the web page. Finally, the personal data extracted from the web page are clustered by DPMM algorithm. The experimental results show that our algorithm improves the F1 value by 9% in the processing of complex web pages (CWPs).

Manuscript received August 16, 2018.
Manuscript revised September 28, 2018.
Manuscript publicized October 24, 2018.

[†]The authors are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.
a) E-mail: wbl@hit.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2018EDL8172

2. Related Work

WePS is a sub task of Name Disambiguation (ND) which contains Named Entity Linking (NEL), Author Name Disambiguation, Place or Organization Disambiguation, etc. Among all ND tasks, WePS focuses on eliminating the disambiguation of web pages returned by person name search. The algorithms for WePS composes of two main phases [6]: (a) web page representation, where the goal is to select suitable features from the web pages for this problem, and (b) applying a clustering algorithm to group web search results, so that each cluster contains all the web pages of a particular individual.

For the first step, web pages representation mainly uses Vector Space Model (VSM) mode, and the most widely used features are bag of words, named entities, and noun phrases [7]. In addition to the VSM, the latent Dirichlet allocation (LDA) is also used for the representation of web pages [8]. Due to the simple structure of the pages in WebPS-* dataset, existing WePS algorithms do not consider the impact of the useless content of the web page. Without web content filtering, every word in the page will be added to the page representation. While, CWPs contain many disturbing words, such as name entities in the ads area, special words in the recommendations area and so on. Currently, text extraction algorithms are mainly used to filter the useless content in web page, but causes about 40% loss of the valid data.

After obtaining the representation of the web page, a clustering algorithm is used for document aggregation in the second step. Depends on the existing researches, Hierarchical Agglomerative Clustering (HAC) is better than other clustering algorithms like K-Medoids or Fuzzy Ants [9]. Although the HAC algorithm works best, HAC needs to specify a specific threshold, and the results are very sensitive to the value of this parameter. For different person name clustering, the optimal thresholds are usually different. [6] proposes an adaptive threshold adjustment method, while the

page representation still uses all text in the web page. Compared with HAC, DPMM has a low sensitivity clustering parameter, and DPMM automatically builds topic associations between words in multiple documents [4].

3. Personal Data Retrieval and Disambiguation

3.1 Sequence Block Model

VIPS is a kind of web page process model which makes full use of page layout feature: it first extracts all the suitable blocks from html DOM tree, and then it tries to find the separators between these extracted blocks [5]. SBM is an improvement of VIPS which only using text content and structure to segment page into visual blocks. Figure 1 shows an example of SBM. Raw web page is shows in Fig. 1 (a) which is the index page of ChinaDaily, and Fig. 1 (b) is modify VIPS structure generated by our previous work [10]. The page is divided into a series of visual blocks in Fig. 1 (b), and each visual block corresponds to a node in the web page DOM tree, as shown in Fig. 1 (c). Then visual blocks are numbered in depth-first order of DOM tree, and the SBM of this page is the visual blocks with index numbers. SBM is actually turning a web page to a set of text sequences, $\{C_1, C_2, C_3, \ldots, C_n\}$, where $C_i = \{w_1, w_2, \ldots, w_m\}$, w_i is word.

The web page in Fig. 1 (a) does not have content area that means traditional content extraction algorithms can not extract meaningful information from it. In this case, SBM provides a new way for extracting: sequence blocks in SBM are detected by classification algorithm to determine if the contents of blocks need to be extracted. For example, No.1-3 blocks in Fig. 1 (a) are navigation bars which are useless in most extract tasks, and we can distinguish these invalid blocks by classification algorithms [10]. Considering personal data extraction, we find that the content of the previous block may affect the judgment of the current block content. For instance, if the content of No.10 block in Fig. 1 (b) is "Other Recommendations", a secondary title, then No.11 block is not person data more likely. But if the content is "Profile", then No.11 is indeed personal data. Based on this phenomenon we introduce the SBM-based sequence prediction model.

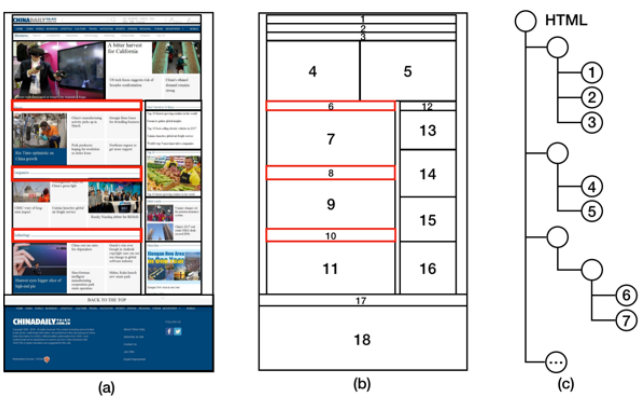


Fig. 1 An example of sequence block model.

3.2 SBM-Based Sequence Prediction Model

Let W be the word set of all pages. According to SBM, a web page is represented by $p_i = \{c_1, \ldots, c_M, s_1, \ldots, s_M\}$, where $c_i = \{w_{i1}, w_{i2}, \ldots, w_{iL}\}, w_{il} \in W$, is the word sequence vector of No.i block content and s_i is other features except text content, like structure features. Then the personal data extraction is expressed as a mapping function of (N, M, L) to $(N, M, 1)$, where N is the number of pages, M is the max blocks number of pages, and L is the max length of block contents. In other words, the extraction algorithm is actually a binary classification algorithm which distinguishes whether each block content is personal data. While traditional classification algorithms only handle (N, M) to (N, M) problems, we build a sequence prediction model based on Long Short-Term Memory (LSTM) [11] for personal data extraction, Double LSTM with Structure (DLS). LSTM is one structure of deep learning which is used for sequence prediction, the usual usages of LSTM are shown in Fig. 2: Fig. 2 (a) is many-to-one which is represented by $Y = LSTM(X)$; Fig. 2 (b) is many-to-many; Fig. 2 (c) is bidirectional LSTM, $Y = BLSTM(X)$ is used to represent that.

Figure 3 is the calculation graph and steps of DLS. For convince, we divide the input vector of web page p_i into two input vectors (p_{ci}, p_{si}) which are content vectors and structure vectors of web page p_i respectively. From calculation step, we can see that there are three parameters in network, word embedding length L_e, LSTM output dimension L_s and BLSTM output dimension L_b.

The output of DLS is shown in Fig. 4. The blocks No. 6-11 and 14 are classified as personal data, and other blocks are not. After DLS extraction, the final output is the contents of No.6-11 and 14 blocks.

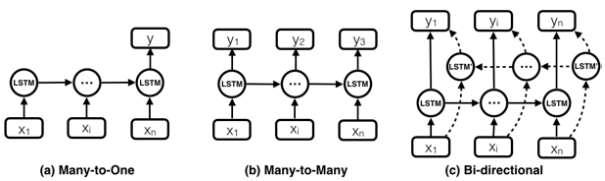


Fig. 2 Multiple uses of long short-term memory.

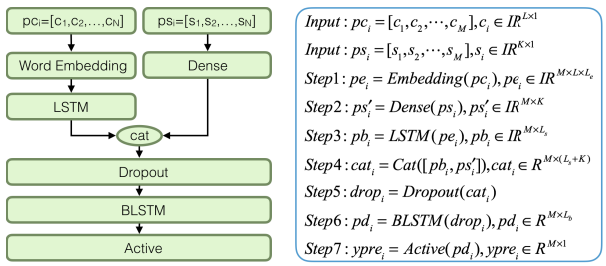


Fig. 3 The calculation graph and steps of double LSTM with structure.

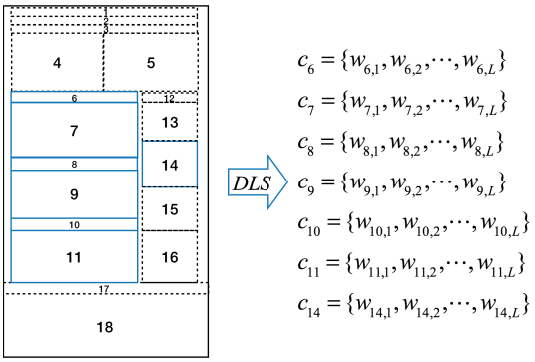


Fig. 4 An example of the output of double LSTM with structure.

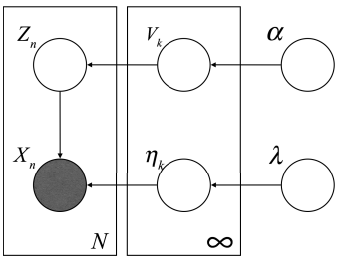


Fig. 5 Graphical model representation of DPMM.

3.3 Name Disambiguation

DLS extracts personal data from raw web pages into texts, and then we need to distinguish which texts describe the same person. The data in DPMM can be generated from the following process [4]:

1. Draw $V_i|\alpha \sim Beta(1, \alpha), i = \{1, 2, \ldots\}$
2. Draw $\eta_i^*|G_0 \sim G_0, i = \{1, 2, \ldots\}$
3. For the nth data point:
   a. Draw $Z_n|v \sim Mult(\pi(v))$
   b. Draw $X_n|z_n \sim p(.|\eta_{z_n})$

The graphical model representation of DPMM is shown in Fig. 5.

A document is presented as Bag-of-Words (BOW) in DPMM, and the words of document are sampled from DPMM. x_i is the BOW vector of the ith document. DPMM assume that the mixture components obeys multinomial distribution with the parameter θ_t, and the base distribution G_0 obeys the Dirichlet distribution with the parameter λ on $L - dimensional$ simplex. The natural parameters of the multinomial is $\eta_t = log\theta_t$ which obeys Dirichlet distribution with the parameter τ_t. After these model parameters are determined, we can use the variational inference [12] to solve the parameters and get the clustering sets.

4. Experiments

Test data contains 5000 web pages retrieved from search engines by 50 Chinese names and all of the pages are Chinese.

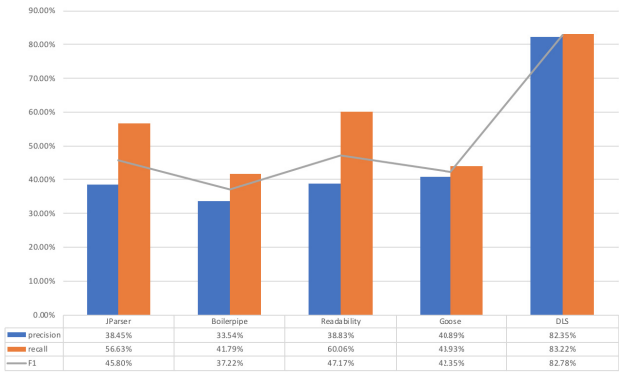


Fig. 6 Comparison of DLS and typical content extraction algorithms.

We only download the HTML source files and convert them into SBMs. Then we labeled blocks in SBMs whether the contents of blocks are personal data. For every record, at least two persons label the record separately and if the labels from them are different, the third person remark that record. After that, we combine the texts which describe same persons and these text sets are used for name disambiguation.

Currently web page extraction without wrapper mainly adopts content extraction algorithms, therefore we choose 4 mainstream content extraction algorithms for comparison. Due to DLS uses neural network and needs labeled samples for training, 3000 web pages are used for training and 2000 for testing. The parameters of DLS is (L_e, L_s, L_b) =(32, 32, 4) after parameters tuning.

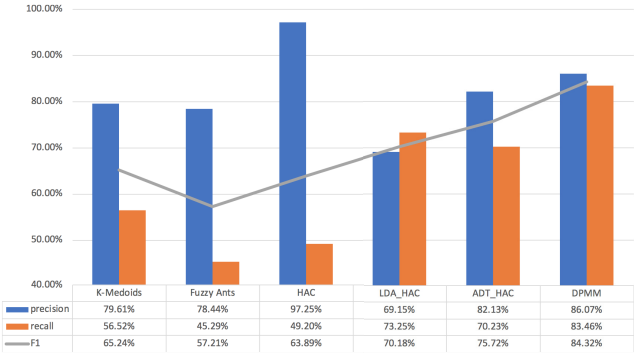
Let P be the whole person text set in 5000 web pages, and C is the text set extracting by algorithm, and $|C|$ is the text length of each algorithm extracting from web pages. For DLS, C is the extracted blocks' text. Then the *recall* is $|P \cap C|/|P|$ and *precision* is $|P \cap C|/|C|$. $F1 = 2 * recall * precision/(recall + precision)$. The experimental results are shown in Fig. 6. From results we can see that $F1$ value of DLS F1 value increases by 30% over content extraction algorithms, and DLS has a 23% higher recall rate than Readablity which is the best content extraction algorithm. The reason is that some person data is not only in the content area but also appears in the sidebar of web pages, and these data can only be extracted by wrapper patterns without SBM.

The inputs of name disambiguation algorithm are contents of one person name from DLS, and the output is a series of textual collections of which one collection corresponds to a real person. The B-Cubed metric [13] is adopted to evaluate the experimental results:

$$precision = \frac{\sum_{S_i \in S} \sum_{d \in S_i} max_{R_j \in R; d \in R_j} \frac{S_i \cap R_i}{S_i}}{\sum_{S_i \in S} |S_i|} \quad (1)$$

$$recall = \frac{\sum_{R_i \in R} \sum_{d \in R_i} max_{S_j \in S; d \in S_j} \frac{R_i \cap S_i}{R_i}}{\sum_{R_i \in R} |R_i|} \quad (2)$$

The $S = \{S_1, \ldots, S_n\}$ is the output of DPMM, and $R = \{R_1, \ldots, R_n\}$ is a real clustering of name disambiguation. Contrast experiments include common clustering algorithms and adaptive algorithms (ADT_HAC) [6], and the



	K-Medoids	Fuzzy Ants	HAC	LDA_HAC	ADT_HAC	DPMM
precision	79.61%	78.44%	97.25%	69.15%	82.13%	86.07%
recall	56.52%	45.29%	49.20%	73.25%	70.23%	83.46%
F1	65.24%	57.21%	63.89%	70.18%	75.72%	84.32%

Fig. 7 Comparison of DPMM and other clustering algorithms.

input of contrast experiments is original web pages. The results are shown in Fig. 7.

HAC algorithm has the highest precision, but the recall rate is low. This is because HAC can only consider fixed features and some features are ignored which is not taken into account in feature engineering. Due to the number of contents taken each time is small (about 50-100), LDA-HAC [14] methods do not work well in our experiments. LDA requires a large number of samples for topic distribution calculations, and if the articles are few, the randomness is too large. DPMM is high in both precision and recall, and F1 is higher than ADT_HAC about 9%.

5. Conclusion

Comparing to content extraction algorithms, DLS improves recall rate by nearly 23%. Due to the DLS does not use feature engineering, it can be seamlessly applied to the extraction of other content, such as company information extraction, commodity information extraction, and so on. The DPMM algorithm is more suitable for our problem than LDA and other clustering algorithms, and the F1 value is increased by 9%. The main problem at the moment is that it takes a lot of time to mark up the data. In the future, we will study how to use semi-supervised methods to reduce marking samples time.

Acknowledgments

This work is partially supported by National Key Research and Development Program of China under Grant No.:2016YFB0800802 and Shandong Key Research and Development Plan under Grant No.:2016ZDJS01A04, 2017CXGC0706.

References

[1] A. Spink, B.J. Jansen, and J. Pedersen, "Searching for people on web search engines," J. Documentation, vol.60, no.3, pp.266–278, 2004.
[2] J. Artiles, J. Gonzalo, and S. Sekine, "The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task," Proc. 4th International Workshop on Semantic Evaluations, pp.64–69, Association for Computational Linguistics, 2007.
[3] H. Ji, R. Grishman, H.T. Dang, K. Griffitt, and J. Ellis, "Overview of the tac 2010 knowledge base population track," Third Text Analysis Conference (TAC 2010), p.3, 2010.
[4] R.M. Neal, "Markov chain sampling methods for dirichlet process mixture models," J. Computational and Graphical Statistics, vol.9, no.2, pp.249–265, 2000.
[5] D. Cai, S. Yu, J.R. Wen, and W.Y. Ma, "Vips: a vision-based page segmentation algorithm," 2003.
[6] A.D. Delgado, R. Martínez, S. Montalvo, and V. Fresno, "Person name disambiguation in the web using adaptive threshold clustering," J. Association for Information Science and Technology, vol.68, no.7, pp.1751–1762, July 2017.
[7] R. Nuray-Turan, D.V. Kalashnikov, and S. Mehrotra, "Exploiting web querying for web people search," ACM Trans. Database Systems (TODS), vol.37, no.1, p.7, Feb. 2012.
[8] Y. Song, J. Huang, I.G. Councill, J. Li, and C.L. Giles, "Efficient topic-based unsupervised name disambiguation," Proc. 7th ACM/IEEE-CS joint conference on Digital libraries, pp.342–351, ACM, 2007.
[9] E. Elmacioglu, Y.F. Tan, S. Yan, M.Y. Kan, and D. Lee, "Psnus: Web people name disambiguation by simple clustering with rich features," Proc. 4th International Workshop on Semantic Evaluations, pp.268–271, Association for Computational Linguistics, 2007.
[10] Y. Wei, B. Wang, Y. Liu, and F. Lv, "Research on webpage similarity computing technology based on visual blocks," Chinese National Conference on Social Media Processing, pp.187–197, Springer, 2014.
[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol.9, no.8, pp.1735–1780, Nov. 1997.
[12] D.M. Blei and M.I. Jordan, "Variational inference for dirichlet process mixtures," Bayesian analysis, vol.1, no.1, pp.121–143, 2006.
[13] A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," The first Int. Conf. Language Resources and Evaluation Workshop on Linguistics Coreference, pp.563–566, Granada, 1998.
[14] B. Huang, Y. Yang, A. Mahmood, and H. Wang, "Microblog topic detection based on lda model and single-pass clustering," Int. Conf. Rough Sets and Current Trends in Computing, pp.166–171, Springer, 2012.