News Article Extraction with Template-Independent Wrapper

Junfeng Wang[1], Xiaofei He[1], Can Wang[1], Jian Pei[2], Jiajun Bu[1], Chun Chen[1], Ziyu Guan[1], Lu Gang[3]

[1] College of Computer Science,
Zhejiang University, Hangzhou, China
{wangjunfeng@, xiaofeihe@cad., wcan@,
bjj@, chenc@, guanzh@}zju.edu.cn

[2] School of Comp. Sci., Simon Fraser University, Canada
jpei@cs.sfu.ca

[3] College of Information, Zhejiang University of Finance and Economics, China
hz_lugang@163.com

ABSTRACT

We consider the problem of template-independent news extraction. The state-of-the-art news extraction method is based on template-level wrapper induction, which has two serious limitations. 1) It cannot correctly extract pages belonging to an unseen template until the wrapper for that template has been generated. 2) It is costly to maintain up-to-date wrappers for hundreds of websites, because any change of a template may lead to the invalidation of the corresponding wrapper. In this paper we formalize news extraction as a machine learning problem and learn a template-independent wrapper using a very small number of labeled news pages from a single site. Novel features dedicated to news titles and bodies are developed respectively. Correlations between the news title and the news body are exploited. Our template-independent wrapper can extract news pages from different sites regardless of templates. In experiments, a wrapper is learned from 40 pages from a single news site. It achieved 98.1% accuracy over 3,973 news pages from 12 news sites.

Categories and Subject Descriptors

H.3.m [**Information Storage and Retrieval**]: Miscellaneous - *Data Extraction, Web*

General Terms

Algorithms, Experimentation.

Keywords

Data extraction, Web mining, Web news, classification.

1. INTRODUCTION

As most news pages today are generated from some underlying structured sources, it is intuitive to assume the existence of templates in news pages from a specific news portal. Some previous template-dependent approaches like Tree Edit Distance (TED) exploit structure similarities of Web pages [1]. However, the generated wrappers only work properly for news pages belonging to previously seen templates. Any subtle changes in the underlying templates are likely to invalidate the wrappers. To eliminate the template dependency of wrappers, Zheng et al. proposed a news extraction technique by exploiting the visual features of news pages [2]. The news title and body in a news page are identified as many visual blocks and extracted accordingly. However, it may fail to maintain the integrity of an extracted news article as it does not treat a news article as an inseparable unit. This approach is likely to leave out the news title or fragments of the news body.

Copyright is held by the author/owner(s).
WWW 2009, April 20-24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

However, human beings can always accurately identify the news article from a news page. This is because the news pages are usually designed to adapt to people's reading habits. Moreover, the fact that human can easily differentiate news title from news body [2] suggests that news titles and news bodies come with different features. Note that, there is also correlation between the news title and news body in a news page. For example, the news title and news body are usually very close in terms of vertical distance, and usually overlap largely in horizontal direction. If these features of news pages can be fully exploited, it is possible to learn a template-independent wrapper using only news pages from a single news portal such that it can accurately extract news articles from various news portals.

We develop a template-independent wrapper that is able to accurately extract news articles from various news portals on the Web. Once the wrapper is learned using only a very small number of pages from a single news site, we are able to extract news articles from news pages in various sites.

2. News Article Extraction

2.1 Problem Statement

Let P denote a news page, T_{DOM} denote the DOM tree built from P and T denote a subtree of T_{DOM}.

Definition 1. *T is T_{TITLE} iff the news title is contained in T and not contained in any child subtree of T.*

Definition 2. *T is the T_{BODY} iff the whole news body is contained in T and not contained in any child subtree of T.*

Then the news extraction problem is defined as follows: given any P, identify T_{TITLE} and T_{BODY} from T_{DOM}, respectively.

Using novel features defined in following sections, we built T_{TITLE} and T_{BODY} identification model based on nonlinear SVM with Gaussian RBF kernel respectively.

2.2 Feature Representation of Body Subtree

Texts in news pages are usually carefully arranged in a comfortable format for reading via using formatting elements, which are the kind of HTML elements mainly used for texts formatting, including paragraph (<p>), bold font (<b>), new line (<br>), italic font (<i>) and highlight (<strong>).

Let R denote the root element of T, and FE denote the collection of formatting elements which are present in the child elements of R. Let FEA denote all formatting elements inside the first two screens. The bounding rectangle of T denoted by $Rect$ can be obtained. Features of T include the following six:

{ RectLeft, RectTop, RectWidth, RectHeight, FormattingElementsNum, FormatedContentLen}

1) *RectLeft* and *RectTop* are the coordinates of upper left point of *Rect*. *RectWidth*, *RectHeight* are the width and height of *Rect*. 2) *FormattingElementsNum* denotes the size of *FE*. This feature is normalized by total *FormattingElementNum* of *FEA*. 3) *FormatedContentLen* is the total length of texts contained in *FE*. It is normalized by total *FormatedContentLen* of *FEA*.

2.3 Feature Representation of Title Subtree

Here features of *T* include the following ten:

{ RectLeft, RectTop, RectWidth, RectHeight,Overlap, Dist, Flat, FontSize, EndWithFullStop, WordNum }

1) *Overlap* describes the horizontal overlap between *T* and T_{BODY}. News title and body are usually close in horizontal direction. 2) *Dist* describes the vertical distance between *T* and T_{BODY}. The news title is not likely to be far away from the news body in terms of vertical distance. 3) *Flat* describes the shape of *T*. Let *min* be the shorter edge of the bounding rectangle of *T*, and *max* be the longer one. *Flat* is defined to be the ratio *min*/*max*. This is based on the observation that news titles are always bounded by long narrow rectangles. 4) *FontSize* is the font size of the root element of *T*. It is based on the observation that news titles are always displayed in font sizes larger enough to be distinguished from news bodies. 5) *EndWithFullStop* describes whether the text contained in *T* ends up with a full stop. By convention, a title hardly ends up with a full stop. This helps distinguish news titles from sentences. 6) *WordNum* describes the number of words in the text contained in *T*. This feature exploits the fact that news titles would not contain paragraphs of texts.

3. Experiments

We use a dataset of 4,073 news pages crawled from 12 online news sites.

1) In first experiment, we label 40 pages in each news site, among which 10, 20, 40 training examples are randomly chosen to train 3 wrappers (W10, W20, W40) respectively. And in each site, we randomly choose 40 pages for testing. The 480 (40 pages×12 sites) testing pages are different from training pages. Then each of the 36 learned wrappers is applied to extract news articles from the 480 testing pages. Finally, 17,280 (36 wrappers×480 testing pages) extracted news articles are evaluated by user study.

Table 1 shows the testing results. We can see that all these 36 wrappers trained from a small number of pages achieved accuracy higher than 96%, which demonstrates the stability of our approach. Also, our approach only requires a very small number of training examples to achieve this performance. All the three wrappers trained from USNEWS achieved the highest accuracy, which implies that USNEWS is a good training site.

2) In second experiment, we aim to explore the performance of our template-independent wrapper over thousands of pages. The USNEWS-W40 wrapper trained in first experiment is used in this experiment. 40 pages from USNEWS are randomly chosen out for training the USNEWS-W40 wrapper and all the remaining 3,973 pages are used for testing. The 3,973 extracted news articles are evaluated by user study.

Table 2 shows the testing results. The wrapper achieved 98.1% accuracy. This demonstrates that our proposed approach can deal with a large variety of news pages. The state-of-the-art method proposed by Reis [1] is based on template-level wrapper induction. It is unable to extract the news articles from any Web news site. Moreover, our approach achieved significantly higher accuracy than Reis' published 87.71% accuracy [1].

Table 1. Testing results of the learned wrappers

Training Site		W10	W20	W40
1.	CNN	98.9%	99.4%	98.7%
2.	BBC	98.9%	98.9%	99.4%
3.	GC.CA	96.2%	97.3%	98.9%
4.	YAHOO	99.2%	98.9%	98.3%
5.	CBC	98.9%	98.9%	98.9%
6.	CBSNEWS	98.7%	98.3%	98.3%
7.	FOXNEWS	99.4%	98.7%	98.7%
8.	NEWSWEEK	98.5%	97.9%	98.7%
9.	SKY	98.7%	99.2%	99.2%
10.	TIME	99.6%	98.9%	99.4%
11.	USATODAY	98.9%	98.5%	98.3%
12.	USNEWS	99.6%	99.6%	99.6%

Table 2. Testing results of the large scale experiment

Testing Site	Title Accuracy	Body Accuracy	News Accuracy	#. of pages
CNN	99.1%	99.6%	98.7%	234
BBC	99.7%	99.2%	98.9%	360
GC.CA	98.2%	98.2%	97.1%	455
YAHOO	100.0%	100.0%	100.0%	318
CBC	99.0%	99.5%	98.5%	408
CBSNEWS	99.3%	98.6%	98.3%	291
FOXNEWS	100.0%	100.0%	100.0%	292
NEWSWEEK	97.8%	100.0%	97.8%	316
SKY	100.0%	100.0%	100.0%	330
TIME	90.4%	100.0%	90.4%	219
USATODAY	99.8%	100.0%	99.4%	437
USNEWS	97.1%	98.1%	95.2%	313
Overall	**98.6%**	**99.4%**	**98.1%**	**3973**

4. Conclusions

In this paper we proposed an effective approach for template-independent news article extraction using a very small number of training pages from a single news site. Over 3,973 testing pages from 12 sites, our wrapper learned from 40 pages in a single news site achieved 98.1% accuracy, which significantly outperforms the previous news extraction methods.

5. Acknowledgements

The project is sponsored by S&T Planning Projects of Zhejiang Province, No.2007C13050.

6. REFERENCES

[1] Reis, D.C., Golgher, P.B., Silva, A.S. and Laender, A.F. Automatic web news extraction using tree edit distance. In *Proc. WWW 2004*, 2004

[2] Zheng, S., Song, R. and Wen, J. Template-Independent News Extraction Based on Visual Consistency. In *Proc. AAAI'07*, pages 1507-1513, 2007