Automatic Web News Extraction Based on DS Theory Considering Content Topics*

Kaihang Zhang[1,2], Chuang Zhang[1] (✉), Xiaojun Chen[1], and Jianlong Tan[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
craigzkh@163.com, zhangchuang@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Abstract. In addition to the news content, most news web pages also contain various noises, such as advertisements, recommendations, and navigation panels. These noises may hamper the studies and applications which require pre-processing to extract the news content accurately. Existing methods of news content extraction mostly rely on non-content features, such as tag path, text layout, and DOM structure. However, without considering topics of the news content, these methods are difficult to recognize noises whose external characteristics are similar to those of the news content. In this paper, we propose a method that combines non-content features and a topic feature based on Dempster-Shafer (DS) theory to increase the recognition accuracy. We use maximal compatibility blocks to generate topics from text nodes and then obtain feature values of topics. Each feature is converted into evidence for the DS theory which can be utilized in the uncertain information fusion. Experimental results on English and Chinese web pages show that combining the topic feature by DS theory can improve the extraction performance obviously.

Keywords: content extraction, Dempster-Shafer theory, maximal compatibility blocks, information fusion

1 Introduction

The Internet has become one of the main accesses to news information, and therefore news websites produce a great number of news contents for users' daily demands. With the fast development of front-end techniques, programmers can use Cascading Style Sheets (CSS) and JavaScript to develop more and more complicated web pages, therefore we will face increasing challenges to extract the main contents from highly heterogeneous web pages. In addition to the news content, a news web page commonly contains lots of irrelevant texts which are known as noises, such as advertisements, navigation panels, comments, etc. The studies and applications, such as news topic detection and tracking, require

* Supported by the National Natural Science Foundation of China (NO. 61602474) and Xinjiang Uygur Autonomous Region Science and Technology Project (NO. 2016A03007-4)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

the news contents which have been processed and stored. Extracting the news content automatically is important for massive news information management, retrieval, analysis, and integration.

There are some online content extraction methods that provide theoretical supports for this paper, such as CEPR [1], CETR [2], and CETD [3]. These methods are efficient and concise, which do not need training and pre-processing. Non-content features (e.g. tag path, text layout, DOM structure, hyperlink) used by these methods are easy to be obtained from an HTML page. However, online extraction methods mentioned above do not pay much attention to topics of news content and only rely on the non-content features. These methods are difficult to recognize the noise whose external characteristics are similar to those of the news content.

In this paper, we present Content Extraction based on DS Theory (CEDST) which is an efficient and accurate news content extraction method. CEDST combines the topic feature and non-content features to improve the extraction performance. The contributions of this method are as follows: (1) Improving the recognition accuracy of news content extraction by introducing the topic feature. (2) Maximal compatibility blocks are used to generate topics from text nodes without linguistic analysis, therefore our method can be easily applied at websites in different languages by replacing a word segmentation method. (3) DS theory has the ability to represent and quantify uncertainties, which combine features in a reasonable way.

2 Related work

Web pages are very heterogeneous and there are no rigid guidelines on how to build HTML pages and how to declare the implicit structure [4]. HTML tags without semantic information bring a lot of difficulties in the content extraction. If we want to develop a precise extraction method which is fully automated, the method should be restricted to a specific domain, such as news extraction [1, 4], data records extraction [5, 6], e-commerce information extraction [7, 8], title extraction [9]. In this paper, we aim at extracting entire news articles from HTML pages automatically and efficiently.

Content extraction for HTML pages has been researched more than a decade. Most of the early studies on the content extraction are rule-based methods. Users write extraction rules with a specific language on the extraction system which assists in generating wrappers quickly, such as TSIMMIS [10] and W4F [11]. In order to reduce manual steps, some semi-automatic methods [12, 13] had been developed. Semi-automatic methods need users to identify regions of interest on web pages, and then use inductive or heuristic algorithms to generate extraction rules. Although rule-based methods extract content accurately, users have to do much hard work manually.

Numerous methods for automatic content extraction have been presented, but each method has its drawbacks. The template-based [4, 6] methods assume that web pages in the same cluster share the same template. These methods can

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

filter out noises from web pages automatically, but any change of websites may lead to templates' failure, therefore the templates need to be trained again. Cai et al. [14] proposed a classical vision-based method named VIPS, which segments a web page into visually grouped blocks based on the DOM tree. Song et al. [15] proposed a method to rank blocks based on VIPS. The biggest disadvantage of the vision-based methods is high consumption of computer resources and time, because these methods need to render the HTML page and retrieve all the CSS files which relate to the page. Najlah et al. [9] proposed a linguistics-based method which can extract titles from web pages, but this method cannot be put into general use, because it requires manual effort and domain knowledge to build the part-of-speech (POS) pattern trees. CEPR [1], CETR [2], and CETD [3] are the online Web content extraction methods without training and preprocessing, inspired by these methods, we propose a novel method named CEDST for news content extraction.

3 News Content Extraction Method

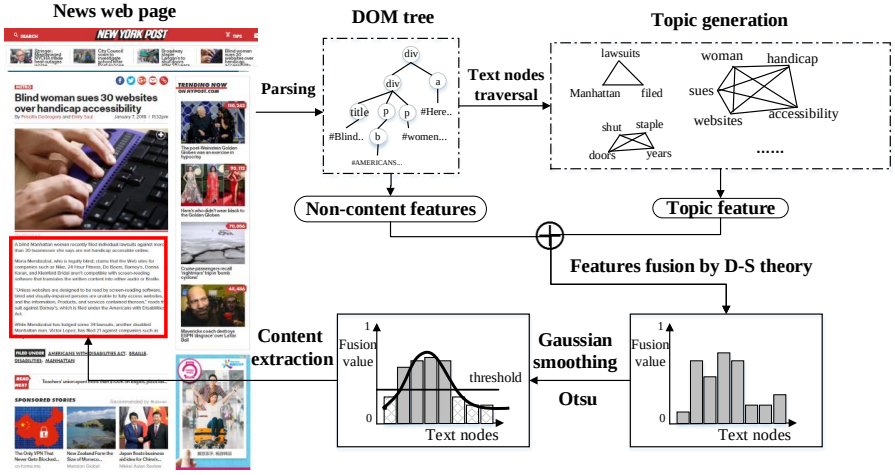


Fig. 1. Process of extracting the news content in the red solid frame.

The HTML page can be parsed into a Document Object Model (DOM) tree. We define the leaf node that contains text as a text node. The process of our news content extraction method is shown in Fig. 1, which consists of following steps: (1) **Parsing the HTML page into a DOM tree**: computing values of non-content features for each text node of DOM tree. (2) **Text nodes traversal**: traversing text nodes of DOM tree in a pre-order manner. (3) **Topic generation**: using maximal compatibility blocks to generate topics from text nodes, and then computing the topic feature of each text node. (4) **Features fusion by DS theory**: combining non-content features and a topic feature by DS theory for obtaining the fusion value of each text node. (5) **Content extraction**: after

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

smoothing the fusion values, we compute a threshold to distinguish the news content from noises by using Otsu algorithm.

3.1 Features Fusion by DS Theory

DS theory [16] is a method for uncertainty. The ability to represent and quantify uncertainties is a key advantage of DS theory. We use DS theory to combine features of a text node to calculate the probability that the text node belongs to news content. We convert features into pieces of evidence by basic mass assignment (BMA) [17]. In the news extraction domain, the frame of discernment for a text node is $\Theta = \{news, \sim news\}$. The BMA function is m:$2^{\Theta} \rightarrow [0, 1]$, where 2^{Θ} is the power set which can be denoted as $2^{\Theta} = \{\emptyset, \{news\}, \{\sim news\}, \Theta\}$. The BMA function should satisfy two conditions: $m(\emptyset) = 0$ and $\sum_{U \subseteq 2^{\Theta}}(U) = 1$. We use the labels "*news*" and "$\sim$ *news*" to denote positive and negative status respectively. If the feature f_i is positive, which supports the text node belonging to the news content, the feature can only assign the probability to label "*news*". The BMA formula is as follows:

$$\left.\begin{array}{c} m_{f_i}(\{news\}) = \alpha_{f_i} \times h_{f_i} \\ m_{f_i}(\{\sim news\}) = 0 \\ m_{f_i}(\Theta) = 1 - m_{f_i}(news) \end{array}\right\} \tag{1}$$

Otherwise, the feature is negative, which can only assign the probability to the label "$\sim$ *news*". The BMA formula is as follows:

$$\left.\begin{array}{c} m_{f_i}(\{news\}) = 0 \\ m_{f_i}(\{\sim news\}) = \beta_{f_i} \times h_{f_i} \\ m_{f_i}(\Theta) = 1 - m_{f_i}(\sim news) \end{array}\right\} \tag{2}$$

h_{f_i} is the feature value normalized between 0 and 1. α_{f_i} and β_{f_i} are the weights assigned between 0 and 1 for features, but we set the feature weights near to 1 in order to avoid the normalizing parameter K of Eq.(5) appearing zero. All the features obtained from BMA can be combined together by using Eq.(3). Given BMA functions m_{f_1},m_{f_2}...m_{f_n} which are reasoned by features of a text node, the fusion function which is denoted as $\oplus$ in Fig. 1 as follows:

$$(m_{f_1} \oplus m_{f_2} \cdots \oplus m_{f_n})(A) = \frac{1}{K} \sum_{\bigcap_{i=1}^{n} A_i = A} \prod_{j=1}^{n} m_{f_j}(A_j) \qquad when A \neq \emptyset \tag{3}$$

$$(m_1 \oplus m_2 \cdots \oplus m_n)(\emptyset) = 0 \tag{4}$$

$$Where: \quad K = 1 - \sum_{\bigcap_{i=1}^{n} A_i = \emptyset} \prod_{j=1}^{n} m_{f_j}(A_j) \tag{5}$$

$(m_{f_1} \oplus m_{f_2} \cdots \oplus m_{f_n})(\{news\})$ is the fusion value which captures the probability that a text node belongs to news content. All the features f_i need to be converted

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

into pieces of evidence which assign the probability to $\{\{news\}, \{\sim news\}, \Theta\}$ in the BMA formula m_{f_i}. Note that the positive and negative features assign the BMA in different manners.

3.2 Topic Generation

Fig. 2 shows the process of topics generation. Firstly, each text node is transformed into a set of keywords, the steps are as follows: word segmentation, deleting stop words and keywords generation by textRank [18]. Maximal compatibility blocks are utilized to generate topics from keywords sets. Compatibility relation is a binary relation which satisfies reflexive and symmetry, so the relation can be denoted as a lower triangular matrix in Fig. 3. Compatibility relation $\boldsymbol{R}$ on keywords is a pair-wise relation between any two words which occur more than γ times in the same text nodes. We set $\gamma = 2$ which will be discussed in section 4.3.

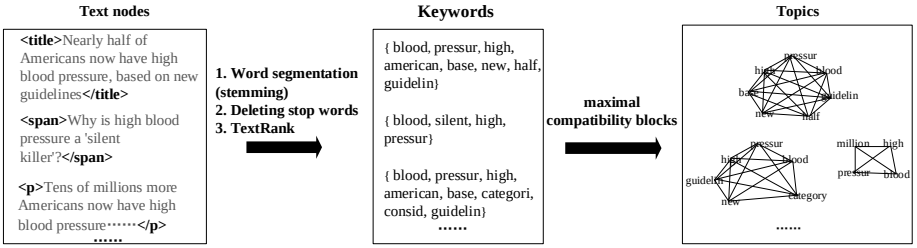


Fig. 2. Process of topic generation

Definition 1. *(**maximal compatibility block**) Let $\boldsymbol{U}$ be the set of all the keywords, if $\boldsymbol{W} \subseteq \boldsymbol{U}$, where any $x, y \in \boldsymbol{W}$ has the relation $x\boldsymbol{R}y$, $\boldsymbol{W}$ is a compatibility block. If there doesn't exist $k \in \boldsymbol{U} - \boldsymbol{W}$ which can be added to the compatibility block $\boldsymbol{W}$, the $\boldsymbol{W}$ is a maximal compatibility block.*

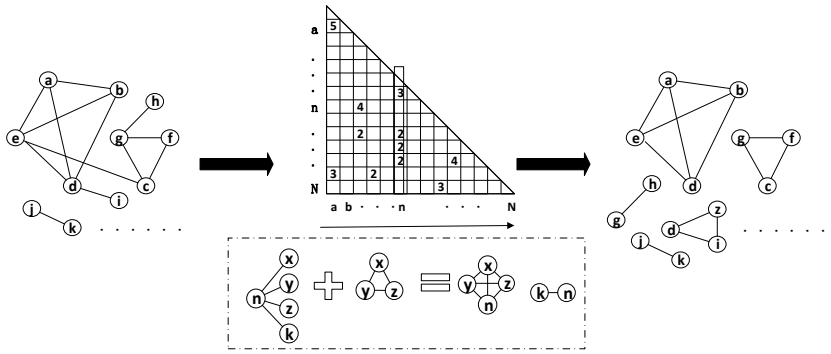


Fig. 3. Process of solving maximum compatibility blocks

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

The process of solving maximum compatibility blocks is shown in Fig. 3. Each node represents a keyword and the line between two nodes represents the compatibility relation $\boldsymbol{R}$. The numbers assigned to the lower triangular matrix are co-occurrence frequencies for any two words which occur more than γ times in the same text nodes. Topics are the maximum compatibility blocks generated from left to right on the lower triangular matrix. As shown in the dotted box below the matrix in Fig. 3, the node n has the relation $\boldsymbol{R}$ with nodes x, y, z, k, and nodes x, y, z belong to an existing compatibility block, so node n and a, b, c can be combined into a bigger compatibility block. After scanning all the nodes on the triangular matrix, all the compatibility blocks that seem like complete polygons can be found out. After removing the compatibility blocks which can be covered by a bigger compatibility block, the rest are the maximal compatibility blocks.The time complexity of topic generation is $O(N * E * R)$, where N, E and R are the number of all the text nodes, existing compatibility blocks and related text nodes respectively. The compatibility relation matrix in Fig. 3 is a sparse matrix, so the maximum compatibility blocks can be generated quickly.

Topics may belong to news content, advertisement, recommendation, etc. A text node has topics with higher weights, which is more likely to belong to news content. The topic weight formula is as follows:

$$tw(topic) = \sum\nolimits_{n \in \{relate(x,y)|x,y \in topic\}} n \tag{6}$$

where $relate(x, y)$ is the co-occurrence frequency of words x, y recorded on the lower triangular matrix in Fig. 3. The value of topic feature is as follows:

$$h_{topic}(text) = \frac{max(\{tw(m)|m \in topic(text)\})}{max(\{tw(n)|n \in topics\})} \tag{7}$$

where $topic(text)$ is a set of topics generated from the $text$, $topics$ is a set of all the topics generated from the web page. The greater the h_{topic}, the more possible the $text$ belongs to news content. The feature value h_{topic} is positive, which can be put into equation.(1) for obtaining the mass function m_{topic}.

3.3 Non-content Features

Non-content features focus on external characteristics of text nodes. For example, the news content commonly contains long texts with a simple format. After observing lots of news websites and considering the features mentioned by [1] and [3], we design the following features to evaluate the probability of a text node belonging to news content.

Text cluster feature. In general, news content consists of continuous text nodes which are commonly appended to several parent nodes. More nodes with long text are appended to the same parent node, these text nodes are more likely to belong to news content, the feature value is as follows:

$$h_{cluster}(text) = \frac{\sum_{t \in sb(text)} size(t)}{max(\{\sum_{w \in sb(n)} size(w)|n \in TextNodes\})} \tag{8}$$

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

where $TextNodes$ denotes a set of all the text nodes in a web page, $sb(text)$ is the set of $text$ and its sibling text nodes. Function $size(\cdot)$ counts the total number of words in a text node. The text cluster feature is positive, so the feature value $h_{cluster}$ should be put into the equation.(1) for obtaining $m_{cluster}$.

Text variance feature. Paragraphs of news content are commonly written with various lengths. But the noise contains brief and neat sentences in general. If the variance of text lengths of a text node and its sibling text nodes is high, the text node is more likely to belong to news content. The feature value is as follows:

$$h_{var}(text) = \frac{var(sb(text))}{max(\{var(sb(n))|n \in TextNodes\})} \tag{9}$$

where $var(sb(text))$ is the variance of text lengths of the $text$ and its siblings. The variance feature h_{var} is positive, which can be put into equation.(1) for obtaining m_{var}.

Hyperlink feature. The noises, such as advertisements and navigation panels, commonly have a high ratio of words nested in hyperlink anchors. Because these noises aim at leading users to other web pages. The feature value is as follows:

$$h_{href}(text) = \frac{size(href(text))}{size(text)} \tag{10}$$

where $href(text)$ is the words nested in hyperlink of the $text$. The more words of a text are nested in hyperlink anchors, the less likely the text belongs to news content. h_{href} is negative, which should be put into equation.(2) for obtaining m_{href}.

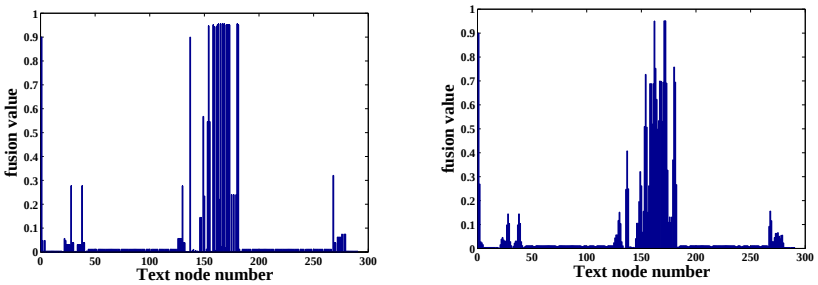
After transforming all the values of features into the evidence construction by using BMA, these features can be combined by using the DS fusion method (equation.(3)). The fusion value is denoted as $m_{topic} \oplus m_{cluster} \oplus m_{var} \oplus m_{href}(\{news\})$, which is the probability of the text node belonging to news content.

3.4 Content Extraction

Gaussian smoothing. After combining all the features by DS theory, we obtain the fusion values of all the text nodes on a web page. The preliminary fusion values might not recognize some special texts of news content, such as short texts without news topics. Although these short texts belong to the news content, their fusion values are low. These texts of news content may be lost without smoothing. We use the one-dimensional Gaussian smoothing mentioned by [1,2] to solve this problem. Fig. 4 shows the fusion values of a news web page of CNN, where the histograms present the fusion values of text nodes before and after smoothing on the web page. The numbers below the abscissa axis represent text

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

nodes traversed in a pre-order manner. Text nodes of news content with low fusion values can be smoothed to higher values. We see that between text nodes 152 and 182, most of the text nodes with high fusion values are identified as news content.



(a) Fusion values before smoothing (b) Fusion values after smoothing

(c) The extraction region of a news web page of CNN

Fig. 4. Fusion values before and after Gaussian smoothing in a news web page of CNN

Threshold segmentation. Based on the fusion values of the histogram after smoothing, we use the Otsu algorithm to compute a threshold which divides the text nodes into two categories. The category with higher fusion values belongs to the news content. Given a threshold t, w_0 and w_1 are the proportions of the two categories separated by the threshold t, u_0 and u_1 are the average fusion values of these two categories.

$$g = w_0(u_0 - \mu)^2 + w_1(u_1 - \mu)^2 \tag{11}$$

$$Where: \quad \mu = w_0 \times u_0 + w_1 \times u_1 \tag{12}$$

μ is the global average and g is the objective function. We calculate the objective function g by using t from 0 to 1 with the step size 0.1. The Otsu algorithm aims to find the threshold t to maximize the objective function g. We use Otsu algorithm to work out that the threshold of Fig. 4(b) is 0.4.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

4 Experimental Result

4.1 Performance Metrics

In this paper, precision, recall and F_1-score are used to evaluate and compare the performance of different content extraction methods. N_e represents the text nodes extracted from a web page and N_l represents the text nodes that are manual labeled results. (Note that in CETR [2], N_e and N_l represent the lines are extracted and hand-labeled respectively). Precision(P), Recall(R) and F_1-score(F_1) are as follows:

$$P = \frac{\sum_{t \in N_e \cap N_l} size(t)}{\sum_{w \in N_e} size(w)}, R = \frac{\sum_{t \in N_e \cap N_l} size(t)}{\sum_{w \in N_l} size(w)}, F_1 = \frac{2PR}{P+R} \tag{13}$$

4.2 Method Evaluation

The experimental data contains two kinds of data sets. (1)**News**: This data set contains news web pages from four Chinese and five English news websites: Xinhuanet, Phoenix News Media (Ifeng), 163 News, People, Freep, CNN, NY Post, Yahoo! News, BBC. Each website contains 100 news web pages which are chosen randomly. (2)**CleanEval**: This corpus contains Chinese and English data sets (ClenaEval-zh and CleanEval-en) from the CleanEval competition mentioned by CEPR [1] and CETR [2]. The CleanEval contains various kinds of web pages, but our goal is to extract entire articles of news web pages, hence we choose the web pages whose structures are similar to news web pages to join the experiment, such as forums and blogs. The manual labeled result of each web page should be restricted to the entire article.

Table 1 and Table 2 show the news extraction results of our method and comparison methods on different data sets. CEDST is the method proposed by this paper. CEDST-NC uses the fusion value $m_{cluster} \oplus m_{var} \oplus m_{href}(\{news\})$ to denote that combining all the non-content features mentioned in this paper. CEDST-TF only uses the topic feature $m_{topic}(\{news\})$ to recognize the news content. CEPR and CETR are the content extraction methods which have similar ideas to our method. CETR is a classical online content extraction method which offers the framework to extract content without training and preprocessing. CEPR is an excellence method which aims at extracting news content automatically and accurately.

Table 1 shows that CEDST is 1.19% higher than CEPR and 7.15% higher than CETR on average F_1-score. It represents that our method outperforms the comparison methods in most cases. CETR performs best on average recall, but it performs worst on average precision. Taking the web page of Xinhuanet news website as an example, it has long abstracts of recommended articles under the news content. CETR can not distinguish these long text noises from news content, because it relies on the tag ratios which focus on the text length and the number of tags. CEPR is a brief and efficient method, which performs stably on all websites. Although CEPR makes full use of non-content features,

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

CEDST outperforms CEPR on the precision with considering topics of news content. Especially on Xinhuanet, People and Freep, CEDST distinguishes the

Table 1. Recall(R), Precision(P), F_1-score(F_1) of each method on different news websites, the highest values are in bold.

	datasets / methods	CEDST	CEDST-NC	CEDST-TF	CEPR [1]	CETR [2]
R(%)	Xinhua net	96.73	98.23	80.62	94.31	**99.42**
	Ifeng	97.72	97.25	88.21	98.22	**99.37**
	163 News	89.62	96.54	86.54	**97.69**	92.00
	People	**94.42**	93.91	83.42	93.33	93.62
	Freep	92.73	**93.55**	76.43	81.23	92.55
	CNN	**97.74**	97.55	89.23	96.28	97.42
	NY post	89.27	83.23	60.72	90.45	**92.32**
	Yahoo! News	94.32	93.24	82.52	92.42	**96.48**
	BBC	96.52	95.11	82.32	96.34	**98.07**
	Average	94.34	94.29	81.11	93.36	**95.69**
P(%)	Xinhua net	**94.41**	76.40	86.26	87.23	71.32
	Ifeng	**96.96**	87.52	95.27	90.75	76.75
	163 News	85.34	94.73	82.96	**95.84**	78.63
	People	90.32	88.32	**91.03**	89.54	78.58
	Freep	82.36	80.33	**91.72**	72.47	62.25
	CNN	**96.14**	87.21	87.55	92.39	73.32
	NY post	71.42	73.54	64.22	**80.23**	79.56
	Yahoo! News	91.65	86.73	**93.21**	92.24	83.25
	BBC	**94.68**	92.51	94.35	90.22	78.42
	Average	**89.25**	85.25	87.40	87.88	75.79
F_1(%)	Xinhua net	**95.56**	85.95	83.34	90.63	83.06
	Ifeng	**97.34**	92.13	91.60	94.34	86.61
	163 News	87.43	95.63	84.71	**96.76**	84.79
	People	**92.32**	91.03	87.06	91.40	85.44
	Freep	**87.24**	86.44	83.38	76.60	74.43
	CNN	**96.93**	92.09	88.38	94.29	83.67
	NY post	79.35	78.09	62.42	85.03	**85.47**
	Yahoo! News	**92.97**	89.87	87.54	92.33	89.38
	BBC	**95.59**	93.79	87.93	93.18	87.15
	Average	**91.73**	89.54	84.14	90.54	84.58

long text noise with the topic feature, while CEPR depends on fusion values after smoothing. Without relating to the main topics of news content, some long text noises are easy to be identified as news content by these comparison methods. CEDST outperforms the CEDST-NC, which means that introducing topic feature can improve the precision obviously, because the non-content fea-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

tures are weak in describing the noise formatted like the news content. Although CEDST-TM performs high precision, the recall is low, because this method only extracts key paragraphs of news content and losses some paragraphs without the main topics. Therefore, CEDST balance the non-content features and the topic feature, which performs best on extraction performance.

Table 2 shows the extraction result on CleanEval. Our method is 2.95% higher than CEPR and 4.74% higher than CETR on average F_1-score, which represents that our method is more suitable to extract entire articles from web pages. CEDST achieves great progress in extracting Chinese web page, because the news topics in Chinese web pages are easier to be captured. Beside the news web page, CleanEval contains many different types of web pages, such as blogs and forums. The result indicates that our method is robust and can be widely used in various web pages with main articles.

Table 2. Recall(R), Precision(P), F_1-score(F_1) of each method on CleanEval, the highest values are in bold.

	datasets / methods	CleanEval-en	ClenaEval-zh	Average
R(%)	CEDST	86.62	**90.73**	88.68
	CEPR [1]	87.42	85.68	86.55
	CETR [2]	**91.33**	89.42	**90.38**
P(%)	CEDST	**90.41**	**92.87**	**91.64**
	CEPR [1]	89.32	86.34	87.83
	CETR [2]	83.25	78.64	80.95
F_1(%)	CEDST	**88.47**	**91.79**	**90.13**
	CEPR [1]	88.36	86.01	87.18
	CETR [2]	87.10	83.68	85.39

Despite many advantages of our algorithm, there are some weaknesses in dealing with extreme circumstances. For example, CEDST performs worst in NY post obviously, because a recommendation may appear many times with identical texts in a web page of NY post. The topic generation method may assign high weights to topics of such recommendations, therefore the noises of these recommendations may be identified as news contents with high values of topic feature.

Table 3 shows that our method is slower than CEPR and CETR on execution time, because the time complexity of topic generation is approximately $O(n^3)$, which is the most time-consuming portion of our method. Although we need more time to construct the topic feature for extracting content from a news web page, our method is the most accurate method. CETR extracts content slower than CEPR, because CETR uses the K-means clustering to distinguish the news content from the noises, the time complexity of K-means clustering

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

is approximately $O(n^3)$, while CEPR calculates a simple threshold to segment news content with time complexity $O(1)$. CEPR compresses execution time of processes of content extraction, but sacrificing accuracy for speed and simplicity. For example, CEPR calculates fusion values of text nodes with the same tag path simultaneously, but losing accuracy when the tag path of noise is the same as news content.

Table 3. Average execution time for each method to extract a news web page

	CEDST	CEPR [1]	CETR [2]
Average execution time	5.39s	1.21s	4.07s

4.3 Parameter Setting

The threshold γ mentioned in section 3.2 is used to adjust compatibility relation among keywords. If the co-occurrence frequency of two keywords is higher than threshold γ, the relation can be recorded in the lower triangle matrix shown in Fig. 3. Threshold γ determines the generation of maximal compatibility blocks, which greatly impacts performance of CEDST. The Fig. 5 shows the tradeoff between recall and precision. Observing trends with the threshold increasing, when $\gamma < 2$, the recall is high, but the precision is low. When $\gamma > 2$, we can see that with the γ increasing, the recall decreases. When $\gamma > 3$, the precision decreases too. $\gamma = 2$ is the best selection for news content extraction, which achieves the highest F_1.

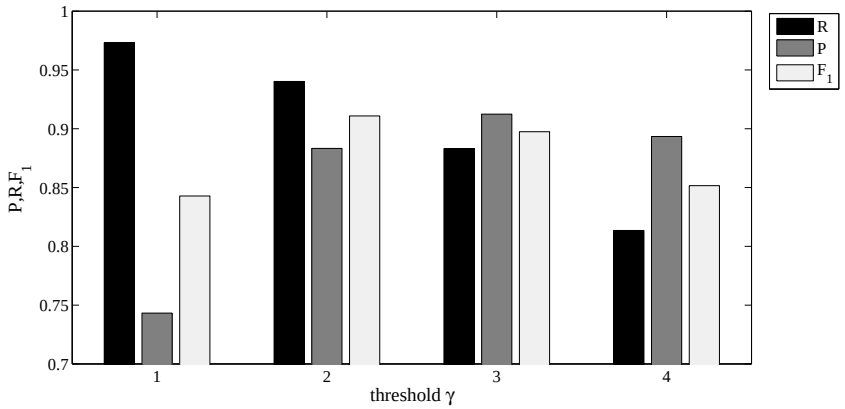


Fig. 5. The extraction performance of CEDST with different γ thresholds

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

5 Conclusion

In this paper, we have proposed a method named CEDST for extracting news contents from news web pages based on DS theory. Considering most of existing online extraction methods that only use the non-content features to recognize the news content, we combine the topic feature and non-content features to achieve the best performance among comparison methods. The proposed method uses maximal compatibility blocks to generate topics from text nodes without complicated linguistics analysis, therefore our method can be applied at websites in different languages easily. DS theory is a method for uncertainty, with the BMA framework, we combine all features to obtain fusion values which are the probabilities of text nodes belonging to new content. The experimental result shows that CEDST outperforms other methods in most cases and performs robustly on various news websites in different languages. CEDST is a concise and efficient method which can extract news content automatically and accurately.

References

1. Wu, G., Li, L., Hu, X., Wu, X.: Web news extraction via path ratios. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 2059–2068. ACM (2013)
2. Weninger, T., Hsu, W.H., Han, J.: Cetr: content extraction via tag ratios. In: Proceedings of the 19th international conference on World wide web. pp. 971–980. ACM (2010)
3. Sun, F., Song, D., Liao, L.: Dom based content extraction via text density. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 245–254. ACM (2011)
4. Reis, D.d.C., Golgher, P.B., Silva, A.S., Laender, A.: Automatic web news extraction using tree edit distance. In: Proceedings of the 13th international conference on World Wide Web. pp. 502–511. ACM (2004)
5. Fang, Y., Xie, X., Zhang, X., Cheng, R., Zhang, Z.: Stem: a suffix tree-based method for web data records extraction. Knowledge and Information Systems pp. 1–27 (2017)
6. Gulhane, P., Madaan, A., Mehta, R., Ramamirtham, J., Rastogi, R., Satpal, S., Sengamedu, S.H., Tengli, A., Tiwari, C.: Web-scale information extraction with vertex. In: Proceedings of the 27th International Conference on Data Engineering (ICDE). pp. 1209–1220. IEEE (2011)
7. Bing, L., Wong, T.L., Lam, W.: Unsupervised extraction of popular product attributes from e-commerce web sites by considering customer reviews. ACM Transactions on Internet Technology (TOIT) 16(2), 1–17 (2016)
8. Charron, B., Hirate, Y., Purcell, D., Rezk, M.: Extracting semantic information for e-commerce. In: Proceedings of the International Semantic Web Conference. pp. 273–290. Springer (2016)
9. Gali, N., Mariescu-Istodor, R., Fränti, P.: Using linguistic features to automatically extract web page title. Expert Systems with Applications 79, 296–312 (2017)
10. Hammer, J., McHugh, J., Garcia-Molina, H.: Semistructured data: the TSIMMIS experience. In: Proceedings of the East-European Conference on Advances in Databases and Information Systems pp. 1–8 (1997)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15

11. Sahuguet, A., Azavant, F.: Building intelligent web applications using lightweight wrappers. Data & Knowledge Engineering 36(3), 283–316 (2001)
12. Ashish, N., Knoblock, C.A.: Semi-automatic wrapper generation for internet information sources. In: Proceedings of the Ifcis International Conference on Cooperative Information Systems. pp. 160–169. IEEE (1997)
13. Liu, L., Pu, C., Han, W.: Xwrap: An xml-enabled wrapper construction system for web information sources. In: Proceedings of the 16th International Conference on Data Engineering. pp. 611–621. IEEE (2000)
14. Deng, C., Shipeng, Y., Jirong, W., Wei-Ying, M.: Vips: a vision-based page segmentation algorithm. Technical Report MSR–TR–2003–79 (2003)
15. Song, R., Liu, H., Wen, J.R., Ma, W.Y.: Learning block importance models for web pages. In: Proceedings of the 13th international conference on World Wide Web. pp. 203–211. ACM (2004)
16. Sentz, K., Ferson, S., et al.: Combination of evidence in Dempster-Shafer theory, vol. 4015. Citeseer (2002)
17. Dong, F., Shatz, S.M., Xu, H.: Reasoning under uncertainty for shill detection in online auctions using dempster–shafer theory. International Journal of Software Engineering and Knowledge Engineering 20(07), 943–973 (2010)
18. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing pp. 404-411 (2004)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_15