ISBN 978-93-84422-80-6
17th IIE International Conference on Computer, Electrical, Electronics and Communication Engineering (CEECE-2017)
Pattaya (Thailand) Dec. 28-29, 2017

Information Extraction of Important International Conference Dates using Rules and Regular Expressions

Jiraphong Sirimueng[1], and Kritsada Sriphaew[1]

[1]College of Information and Communication Technology, Rangsit University, Pathumtani, Thailand
(corresponding author's phone: +66 80 806 6455; e-mail: jiraphong.s58@rsu.ac.th).
(kritsada@rsu.ac.th)

***Abstract*:** *Websites such as website www.conferencealert.com are important sources for managing the dates of international conferences but typically require human input for updating, which can be a barrier to the effective dissemination of current information. One approach to solving this problem is to develop an automatic event calendar update system. The framework for such a system would require an information extraction method. This paper proposes algorithm for extracting important dates from web pages content regarding international conference. We selected 720 web pages from which we extracted data from the HTML structure, using rules and regular expressions.*

Keywords: *Information Extraction, Important Date, Regular Expression, Web Mining.*

1. Introduction

As the World Wide Web hosts an increasing number of international conference websites, information access is becoming increasingly complicated and difficult for users. Under the current model, such information is available on international conference websites such as www.conferencealert.com, which currently requires updating by humans, thus creating barriers to the effective and timely distribution of current information to interested parties.

One concept for solving the problem is to develop a framework for an automatic event calendar update system. This would entail using a web crawler to access international conference web pages and classify the important date content on these web pages through the application of an information extraction method to the HTML structures. Then an updated set of events can be entered into a database featuring a query showing end-users the relevant interface web portals.

2. Literature Review

The theory of information extraction (IE) describe how the process of obtaining data from multiple documents can meet user requirements [1].

In related work, researchers have proposed unstructured IE techniques. In 2005, Karl Michael Schneider applied a CRFs algorithm to 161 instances of word-based extracted data on paper submission deadlines with a precision and recall of 92.0% and 68.9%, respectively, to 168 date-of-conference with a precision and recall of 90.8% and 72.7% [2]. IE has also been implemented in studies of semi-structured data. In 1999, Stephen Soderland proposed the WHISK technique. base on regular expressions to extract the number of beds and room prices from an online apartment rental website [3]. In 2007, Xin Xin, Juanzi Li and Jie Tang extract meta-data such as date-of-conference data from a DBWorld dataset using special features and demonstrated that doing so is preferable to using general features [4]. In 2013, Hidetsugu Nanba et al. used a four-fold cross-validation method

for event travel extraction from 1,022 web pages using the search word “ibento”(event) and other information such as identified manually event from another 264 web pages with a precision of 82.4% and recall of 52.2% [5].

In this study, we extracted information from international conference web pages containing important date content as sources for paper submission deadlines and conference dates.

3. Methodology

Figure 1 shows an information extraction flowchart for the four steps in our process: webpage collection, preprocessing, information extraction and evaluation.

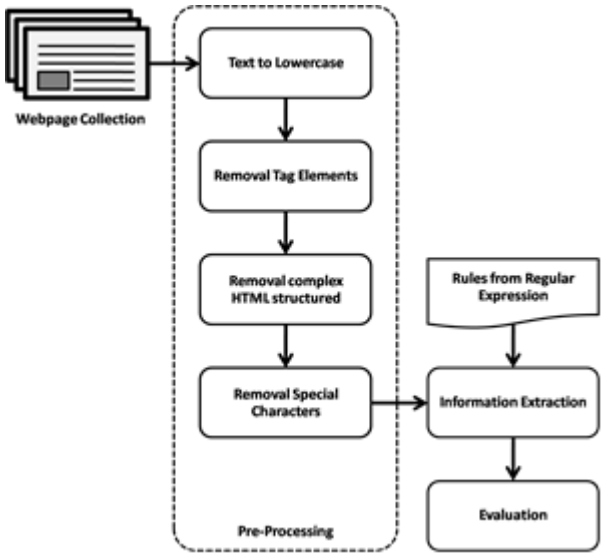

Fig. 1: Flowchart of important date extraction.

3.1. Webpage Collection

As first step, we collected data from August to December of 2017. We examined the HTML structures of 720 web pages with important date content for international conference and obtained table structures from of 94.72% of the pages (see for example Figure 2) and complex structures from 5.28% of the pages (see example in Figure 3).

```
<table border="1" cellspacing="0" cellpadding="10"
width="90%">
  <tr>
    <td><p>Paper Submission (Full Paper)</p></td>
    <td width="446"><p>Before 28th October, 2017</p>
</td>
  </tr>
  <tr>
    <td><p>Notification of Acceptance</p></td>
    <td><p>Before 1st November, 2017</p></td>
  </tr>
  <tr>
    <td><p>Final Authors' Registration</p></td>
    <td><p>Before 5th November, 2017</p></td>
  </tr>
  <tr>
    <td width="476"><p>Conference Dates</p></td>
    <td><p>5th-6th December, 2017</p></td>
  </tr>
</table>
```

Fig. 2 HTML table structure

```
<p style="text-align: center;"><strong><span
style="color: #141694;">Important dates:</span></strong></p>
<p style="text-align: center;"><span style="color:
#141694;">Paper submission:<span style="color: #ff6600;">
<strong> February 15, 2018</strong></span></span></p>
<p style="text-align: center;"><span style="color:
#141694;">Notification of acceptance: <span style="color:
#ff6600;"><strong>March 20, 2018</strong></span></span>
</p>
<p style="text-align: center;"><span style="color:
#141694;">Camera-ready submission: </span><span style="color
#ff6600;"><strong>March 31, 2018</strong></span></p>
<p style="text-align: center;"><span style="color:
#141694;">Conference: </span><span style="color: #ff6600;">
<strong>June 24 &amp;#8211; 27, 2018</strong></span></p>
```

Fig. 3: Complex HTML structure.

3.2. Preprocessing

After the multiple webpage HTML structure collection process in the preceding step, we normalized the collected text in a preprocessing step comprising four sub-steps:

Changing text to lowercase: This removes ambiguity over whether text is uppercase or lowercase.

Removal of tag elements: To simplify the data, we remove the elements from HTML tags.

Removal of complex HTML structures: Some web documents have content such as table or complex structures, with HTML patterns that use tags such as "<p>", " ", "<li>" etc. As such complex structures have multiple pattern, we remove some of them from the document HTML.

Removal of special characters: To improve information extraction performance, we remove special characters such as " ", ",", "(", ")" etc. from web documents.

For information with date ranges (such as 23-24 or 23rd-24th), we use the format "-\d{1,2}" to remove the latter date in the range to improve date extraction performance. We apply this in cases involving date-of-conference content.

3.3. Information Extraction

After preprocessing, important dates including paper submission deadline and conference dates are extracted from the web documents using regular expressions and rules. The patterns used to extract data from structured table are listed in Table 1.

TABLE I: Example of Regular Expressions for HTML Table Structure

Information	Regular Expression
Deadline for paper submission	<tr><td>[a-z\s]*[a-z]*\s?submission\s?[a-z\s]*[a-z]*<\/td><td>(.*?)<\/td><\/tr>
Date of conference	<tr><td>[a-z\s]*[a-z]*\s?conference\s?[a-z\s]*[a-z]*<\/td><td>(.*?)<\/td><\/tr>

In our collected data, we found that 91.53% had the word "submission" and 99.86% had the word "conference" within sentences; therefore, we looked for expressions such as "[a-z\s]*[a-z]*\s?submission\s?[a-z\s]*[a-z]*" in which the asterisks stand for strings of English characters with or without spaces. To find dates following the "<td>" tag, we used the search expression "(.*?)" to capture any characters in the succeeding tabular data and filtered the date format using extraction rules such as those shown in Table 3.

Complex HTML structures have to be clear of any complex HTML tags, to be considered as unstructured data. For this reason, we define regular expressions to extract dates from unstructured information, which are shown in Table 2.

TABLE II: Example of Regular Expressions for Complex HTML Structure

Information	Regular Expression
Deadline for paper submission	submission?\s*[a-z\s]*\s([a-z] {3,9}\s+\d{1,2}\s+\d{4}) submission?\s*[a-z\s]*\s([a-z]{3,9}\s+\d{1,2}[a-z]{2}\s+\d{4}) submission?\s*[a-z\s]*\s(\d {1,2}\s+[a-z]{3,9}\s+\d{4}) submission?\s*[a-z\s]*\s(\d {1,2}[a-z]{2}\s+[a-z]{3,9}\s+\d{4})
Date of conference	conference?\s*[a-z\s]*\s([a-z] {3,9}\s+\d{1,2}\s+\d{4}) conference?\s*[a-z\s]*\s([a-z]{3,9}\s+\d{1,2}[a-z]{2}\s+\d{4}) conference?\s*[a-z\s]*\s(\d {1,2}\s+[a-z]{3,9}\s+\d{4}) conference?\s*[a-z\s]*\s(\d {1,2}[a-z]{2}\s+[a-z]{3,9}\s+\d{4})

Using the expressions in Table 2, the wording relevant to deadlines for paper submission and conference dates based on the "submission" and "conference" font positions date formats. We then created a pattern for standard date formatting to extracting dates from the content.

Table 3 lists the features and rules used for extracting date, month, and year (the dictionary feature uses the prefix is D_.)

TABLE III: Features for Date Extraction

Feature	Rule	Example
Day	1-digit day or 2-digit day	2 or 12th
D_Month	january, february, march, april, may, june, july, august, september, october, november, december, jan, feb, mar, apr, may, jun, jul, aug, sep, sept, oct, nov, dec	
Year	4-digit year	2017

3.4. Evaluation

We conducted evaluations of our proposed method to measure its precision, recall, and accuracy.

4. Experimental Results

To assess our rule and regular expression-based IE methodology, we measure precision (P), recall (R), and accuracy (A). We partitioned the experiment results into two sections: the results of extracting table structured information, which are listed in Table 4, and the results of extracting unstructured information, listed in Table 5.

TABLE IV: Results of Information Extraction for HTML Table Structure

Information	Instance	P	R	A
Deadline for paper submission	682	91.28%	99.71%	91.03%
Date of conference	680	100%	99.85%	99.85%

TABLE V: Results of Information Extraction for Complex HTML Structure

Information	Instance	P	R	A
Deadline for paper submission	38	52.11%	97.37%	51.39%
Date of conference	38	58.73%	97.37%	57.81%

The results indicate good performance in extracting information from table structures since the HTML table tags provide exact structuring in such cases. By contrast, the results of extracting unstructured information are somewhat worse.

5. Conclusion and Future Work

In this study, we successfully extracted information on important date from international conference web pages using rules and regular expressions. In future work we contemplate using machine learning to analyze multiple HTML structures for higher performance extraction of data.

6. References

[1] L. Xiao, D. Wissmann, M. Brown, and S. Jablonski, "Information Extraction from the Web: System and Techniques," *Appl. Intell.*, vol. 21, no. 2, pp. 195–224, 2004.

[2] K.-M. Schneider, "An Evaluation of Layout Features for Information Extraction from Calls for Papers," in *LWA 2005*, 2005, pp. 111–115.

[3] S. Soderland, "Learning Information Extraction Rules for Semi-Structured and Free Text," *Mach. Learn.*, vol. 34, no. 1, pp. 233–272, 1999.

[4] X. Xin, J. Li, and J. Tang, "Enhancing Semantic Web by Semantic Annotation: Experiences in Building an Automatic Conference Calendar," in *Web Intelligence, IEEE/WIC/ACM International Conference on*, 2007, pp. 439–442.

[5] H. Nanba, R. Saito, A. Ishino, and T. Takezawa, "Automatic Extraction of Event Information from Newspaper Articles and Web Pages," in *Digital Libraries: Social Media and Community Networks: 15th International Conference on Asia-Pacific Digital Libraries, ICADL 2013, Bangalore, India, December 9-11, 2013. Proceedings*, S. R. Urs, J.-C. Na, and G. Buchanan, Eds. Cham: Springer International Publishing, 2013, pp. 171–175.