

Contents

1	Introduction	1
1.1	Information Extraction	1
1.2	Information Extraction by Text Segmentation	4
1.3	Main Contributions	5
	References	7
2	Related Work	9
2.1	Web Extraction Methods and Tools	9
2.1.1	Languages for Wrapper Development	9
2.1.2	Wrapper Induction Methods	10
2.1.3	NLP-Based Methods	11
2.1.4	Ontology-Based Methods	11
2.1.5	HTML-Aware Methods	12
2.2	Probabilistic Graph-Based Methods	13
2.2.1	Supervised Probabilistic Graph-Based Methods	13
2.2.2	Unsupervised Probabilistic Graph-Based Methods	14
	References	15
3	Exploiting Pre-Existing Datasets to Support IETS	19
3.1	Overview	19
3.2	Knowledge Bases	21
3.3	Learning Content-Based Features	22
3.3.1	Attribute Vocabulary	22
3.3.2	Attribute Value Range	23
3.3.3	Attribute Value Format	24
3.4	Inducing Structure-Related Features	26
3.5	Automatically Combining Features	28
3.5.1	Combining Content-Based Features	29
3.5.2	Combining Structure-Based and Content-Based Features	29
3.6	Unsupervised Extraction Methods	30
	References	31

4	<i>ONDUX</i>	33
4.1	Overview	33
4.2	Blocking Step	35
4.3	Matching Step	36
4.4	Reinforcement Step	36
4.5	Experimental Evaluation	37
4.5.1	Setup	37
4.5.2	Extraction Evaluation	39
4.5.3	Dependency on Previously Known Data	44
4.5.4	Performance Issues	46
4.5.5	Comparison with Previous Methods	47
4.6	The <i>ONDUX</i> Tool	48
4.6.1	Tool Architecture	48
4.6.2	Graphical User Interface	50
4.6.3	Case Study	50
	References	52
5	<i>JUDIE</i>	53
5.1	The <i>JUDIE</i> Method	53
5.2	Overview	55
5.3	Structure-Free Labeling	56
5.3.1	Processing the Structure-Free Labeling	56
5.3.2	Limitations of the Structure-Free Labeling	57
5.4	Structure Sketching	58
5.5	Structure-Aware Labeling	59
5.6	Structure Refinement	60
5.7	The SD Algorithm	60
5.8	Experimental Evaluation	63
5.8.1	Setup	63
5.8.2	General Quality Results	64
5.8.3	Impact of the Knowledge Base	67
5.8.4	Impact of Structure Diversity	69
5.8.5	Comparison with Previous Work	70
5.8.6	Performance Issues	71
	References	72
6	<i>iForm</i>	75
6.1	The Form-Filling Problem	75
6.2	The <i>iForm</i> Method	77
6.3	Using Content-Based Features	78
6.4	Mapping Segments to Fields	79
6.5	Filling Form-Based Interfaces	80

6.6	Experiments	81
6.6.1	Setup	81
6.6.2	Varying ε	82
6.6.3	Experiments with Multi-Typed Web Forms	83
6.6.4	Number of Previous Submissions	85
6.6.5	Content Overlap	86
6.6.6	Comparison with <i>iCRF</i>	87
	References	89
7	Conclusions and Future Work	91
7.1	Conclusions	91
7.2	Future Work	92
	References	93

Unsupervised Information Extraction by Text
Segmentation

Cortez, E.; da Silva, A.

2013, XV, 94 p. 25 illus., Softcover

ISBN: 978-3-319-02596-4