

AUTOMATIC DEVELOPMENT OF ONTOLOGY IN AGRICULTURE DOMAIN

NEHA KAUSHIK



DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2020

©Indian Institute of Technology Delhi (IITD), New Delhi, 2020

AUTOMATIC DEVELOPMENT OF ONTOLOGY IN AGRICULTURE DOMAIN

by

NEHA KAUSHIK

DEPARTMENT OF MATHEMATICS

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2020

CERTIFICATE

This is to certify that the thesis entitled *Automatic Development of Ontology in Agriculture Domain* submitted by *Ms. Neha Kaushik* to the Indian Institute of Technology Delhi, for the award of the Degree of the **Doctor of Philosophy**, is a record of the original bona fide research work carried out by her under my supervision and guidance. The thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

Place:

Prof. Niladri Chatterjee

Date:

Department of Mathematics
Indian Institute of Technology Delhi

ACKNOWLEDGEMENTS

I owe my deepest gratitude and regards to the almighty whose blessings led me to take up this research work.

Many people have supported me during the entire period of this research work. Acknowledging all of them here is not possible. I mention few of them here.

I am grateful to my Ph.D. supervisor **Professor Niladri Chatterjee** for patiently teaching me basics of research. He has supported me through all the ups and downs during the research period. His flexible nature and unfailing support helped me sail through the tides when it seemed impossible to me. At many times, he showed his trust and encouraged me to accomplish difficult tasks.

I thank IIT Delhi for providing me necessary facilities to carry out the research work. I thank Head, Department of Mathematics, DRC Chairperson and my SRC members for their help at various stages of this research work. I am grateful to Prof. Saroj Kaushik, Prof. S.Dharmaraja and Prof. Aparna Mehra for their valuable feedbacks during the progress presentations.

I am thankful to Principal, Kasturba Institute of Technology for letting me continue Ph.D. at IIT Delhi. I am specially thankful to Mrs. Vanita Pimparkar, Head, Department of Computer Engineering for lending her support and encouragement when I was busy in completion of this thesis. I am thankful to all the faculty members of Department of Computer Engineering for their constant support during this research work.

My family, without whose support, this thesis would not have seen light of the day needs special mention here. This thesis actually belongs to them. It is their patience because of which I could devote time and focus on this research work. I want to specially thank my parents, Mr. S.K. Swami and Mrs. Laxmi Swami, who made me capable to reach to a stage where I could join Ph.D. programme. I am deeply indebted to my father-in-law, Mr. D.V.

Sharma, who motivated me to join Ph.D. I am very thankful to my mother-in-law Mrs. Rajkumari Kaushik whose unending support and patience was a key ingredient in my devotion towards this research. My husband, Mr. Amit Kaushik, has been by my side throughout this research. I also want to thank my son, Arjun Kaushik, whose innocent smile made all the difficulties go away like a bubble.

I am thankful to all the anonymous reviewers for their reviews and valuable feedbacks for improvement of this thesis.

Place:

Date:



Neha Kaushik

ABSTRACT

India has a huge amount of agricultural data in the form of textual documents, tables and spread sheets. However, the data is often underutilized by different Government and/or other organizations because of lack in application of data processing techniques to agriculture data. Converting this data into efficient knowledge representation is found to be helpful in answering queries of the underlying domain. Ontology is one such knowledge representation technique.

Automatic ontology development focuses on converting the available domain text to machine processable knowledge representation in the form of ontology. Automatic term extraction and automatic relation extraction constitute important steps prior to ontology design. We propose a regular expression and natural language processing based scheme for automatic term extraction. Automatic relation extraction in the context of automatic ontology development involves identification of relations relevant to the domain. It further involves identification of related pairs of terms corresponding to each relation. Two types of relations *viz.* intra-subdomain and inter-subdomain relations are worked upon in this research. We propose a knowledge based scheme by using expert knowledge for framing of rules for identifying related pairs of terms for each relation. The scheme is further improved by pre-processing the input text using co-reference resolution.

Multiple ontologies for the same domain may be developed in a distributed manner. Hence automatic ontology merging also constitutes an important part of incremental automatic ontology development. Development of automatic ontology merging scheme involves identifying equivalence, hierarchical and other semantic relations between concepts and individuals of source ontologies. We use two linguistic measures to identify lexical anchors between source ontologies. We also use a mathematical model, *viz.*, Formal Concept Analysis, for identification of additional anchors.

सार

भारत में पाठ्य सामग्री, टेबल और प्रसार शीट के रूप में कृषि डेटा की एक बड़ी मात्रा है। हालाँकि, डेटा को अक्सर विभिन्न सरकारी और / या अन्य संगठनों द्वारा रेखांकित किया जाता है क्योंकि कृषि डेटा के लिए डेटा प्रसंस्करण तकनीकों के आवेदन में कमी होती है। इस डेटा को कुशल ज्ञान प्रतिनिधित्व में परिवर्तित करना अंतर्निहित डोमेन के प्रश्नों का उत्तर देने में मददगार पाया जाता है। ओन्टोलॉजी एक ऐसी ज्ञान प्रतिनिधित्व तकनीक है।

स्वचालित ओन्टोलॉजी विकास, ओन्टोलॉजी के रूप में उपलब्ध प्रक्रियात्मक पाठ को मशीन प्रक्रियात्मक ज्ञान प्रतिनिधित्व में परिवर्तित करने पर केंद्रित है। स्वचालित शब्द निष्कर्षण और स्वचालित संबंध निष्कर्षण ओन्टोलॉजी डिजाइन से पहले महत्वपूर्ण कदम हैं। हम स्वतः अवधि निष्कर्षण के लिए एक नियमित अभिव्यक्ति और प्राकृतिक भाषा प्रसंस्करण आधारित योजना का प्रस्ताव करते हैं। स्वचालित ओन्टोलॉजी विकास के संदर्भ में स्वचालित संबंध निष्कर्षण में डोमेन से संबंधित संबंधों की पहचान शामिल है। इसमें आगे प्रत्येक संबंध के अनुसार संबंधित युग्मों की पहचान शामिल है। इस शोध में दो प्रकार के संबंधों अर्थात् इंTRA-सबडोमेन और इंTER-सबडोमेन संबंधों पर काम किया जाता है। हम प्रत्येक संबंध के लिए शर्तों के संबंधित जोड़े की पहचान के लिए नियमों के निर्धारण के लिए विशेषज्ञ ज्ञान का उपयोग करके एक ज्ञान आधारित योजना का प्रस्ताव करते हैं। सह-संदर्भ रिज़ॉल्यूशन का उपयोग करके इनपुट टेक्स्ट को प्री-प्रोसेसिंग करके योजना को और बेहतर बनाया गया है।

एक ही डोमेन के लिए एकाधिक ओन्टोलॉजी को वितरित तरीके से विकसित किया जा सकता है। इसलिए स्वचालित ओन्टोलॉजी विलय भी वृद्धिशील स्वचालित ओन्टोलॉजी विकास का एक महत्वपूर्ण

हिस्सा है। स्वचालित ओन्टोलॉजी विलय योजना के विकास में समकालिकता, पदानुक्रमित और अन्य अर्थ संबंधी संबंधों की पहचान और स्रोत सिद्धांतों के व्यक्तियों के बीच संबंध शामिल हैं। स्रोत ओन्टोलॉजी के बीच लेक्सिकल एंकर की पहचान करने के लिए हम दो भाषाई उपायों का उपयोग करते हैं। अतिरिक्त एंकरों की पहचान के लिए हम गणितीय मॉडल का उपयोग करते हैं।

TABLE OF CONTENTS

CERTIFICATE	i
ACKNOWLEDGEMENTS	ii
ABSTRACT	iv
TABLE OF CONTENTS	vii
LIST OF FIGURES	xii
LIST OF TABLES	xv
Chapter 1 INTRODUCTION.....	17
1.1 Ontology Development	18
1.2 Challenges in Ontology Development	20
1.2.1 General Challenges	20
1.2.2 Challenges Specific to Ontology Development in Agriculture Domain.....	21
1.3 Motivation for Automatic Development of Ontology in Agriculture Domain	22
1.4 Objectives of the Thesis	24
1.5 Organization of the Thesis	24
Chapter 2 LITERATURE SURVEY	30
2.1 Ontology.....	31
2.2 Ontology Development Methodologies	31

2.3 Automatic Ontology Development	36
2.4 Ontology Merging.....	39
2.5 State of the Art in Agriculture Domain.....	46
2.6 Concluding Remarks	47
Chapter 3 AUTOMATIC TERM EXTRACTION	48
3.1 Term Extraction Methods	49
3.1.1 Statistical Methods.....	49
3.1.2 Distributional Methods	51
3.1.3 Contextual Methods	52
3.1.4 Linguistic Methods	53
3.1.5 Hybrid Approaches	54
3.2 Term Extraction Tools	56
3.2.1 RAKE.....	56
3.2.2 TermRaider	58
3.2.3 TerMine.....	59
3.3 Need of a Novel Approach for Term Extraction.....	60
3.4 The Proposed Algorithm: RENT	62
3.6 Results and Analysis	66

3.7 Concluding Remarks	71
Chapter 4 AUTOMATIC RELATION EXTRACTION	73
4.1 Introduction	74
4.2 Relation Extraction Methods.....	75
4.3 Intra-subdomain Relation Extraction	78
4.3.1 mOIE	78
4.3.2 RelExOnt: <i>Relation Extraction for Ontology</i>	91
4.4 Inter-subdomain Relation Extraction	101
4.4.1 Baseline Algorithm	102
4.4.2 Improved Algorithm for Inter-subdomain Relation Extraction	108
4.5 Concluding Remarks.....	111
Chapter 5 ONTOLOGY MERGING FOR INCREMENTAL DEVELOPMENT OF ONTOLOGIES	113
5.1 Introduction	114
5.2 Similarity Measures	119
5.2.1 N-gram	119
5.2.2 Edit Distance	119
5.2.3 Li's Similarity	121

5.2.3 Cosine Similarity.....	125
5.3 Word Embeddings.....	127
5.3.1 Word2Vec	127
5.3.2 GloVe	128
5.3.3 FastText.....	129
5.3.4 BERT	129
5.4 Formal Concept Analysis.....	131
5.5 Related Work	136
5.6 The Proposed Method	138
5.6.1 Identification of Lexical Anchors	138
5.6.2 Identification of hierarchical relations	140
5.6.3 Merging of Source Ontologies	142
5.7 Results	143
5.8 Concluding Remarks	149
Chapter 6 CONCLUSION AND FUTURE WORK.....	151
6.1 Summary of the Research Work	152
6.1.1 Automatic Term Extraction.....	152
6.1.2 Automatic Relation Extraction.....	153

6.1.3 Ontology Merging.....	155
6.2 Future Work.....	155
REFERENCES.....	157
APPENDIX I: ONTOLOGY REPRESENTATION AND EVALUATION.....	171
I.1 Ontology Representation	172
I.1.1 Description Logics.....	175
I.1.2 RDF.....	178
I.1.3 OWL	181
I.1.4 Protégé	183
I.1.5 CMap COE	185
I.1.6 Ontology Creation using Owlready	186
I.2 Ontology Evaluation.....	187
I.3 Concluding Remarks.....	191
APPENDIX II: SPARQL QUERIES FOR ONTOLOGY EVALUATION	192
PUBLICATIONS.....	202
BRIEF RESUME OF THE AUTHOR.....	203

LIST OF FIGURES

Figure 2.1 Uschold & King Methodology	32
Figure 2.2 Gruninger and Fox Methodology	32
Figure 2.3 METHONTOLOGY Framework	33
Figure 2.4 OTK Methodology	34
Figure 2.5 IDEF5 Methodology.....	35
Figure 2.6 Ontology Merging	40
Figure 2.7 Ontology Mapping.....	41
Figure 2.8 Ontology Alignment	41
Figure 2.9 Ontology Refinement.....	42
Figure 2.10 Ontology Integration.....	42
Figure 2.11 Three-Tier Architecture for Ontology Merging Approaches	43
Figure 2.12 Screenshot from AGROVOC showing the hierarchical structure	46
Figure 3.1 Overview of RAKE	56
Figure 3.2 Overview of TermRaider	59
Figure 3.3 Overview of TerMine	60
Figure 3.4 Sample Text 1	64
Figure 3.5 Sample Text 1 after POS-Tagging.....	65

Figure 3.6 The RENT Algorihtm	65
Figure 4.1 Sample Text 2	74
Figure 4.2 Example subdomains of agriculture	78
Figure 4.3 Algorithm for computation of similarity between two terms	86
Figure 4.4 RelExOnt Algorithm.....	93
Figure 4.5 Framework for Inter-subdomain Relation Extraction.....	102
Figure 4.6 Baseline algorithm for Inter-subdomain Relation Extraction.....	104
Figure 4.7 Example text for relation between crop and soil	105
Figure 4.8 Example text for relation between soil and region	106
Figure 4.9 Example text for relation between crop and disease	106
Figure 4.10 Example text which needs co-reference resolution	108
Figure 4.11 Example text of Figure 4.10 after co-reference resolution.....	108
Figure 4.12 Modified Framework for Inter-subdomain Relation Extraction.....	109
Figure 5.1 Different classifications of the same concept ‘crop’ according to (a) product type, (b) crop species, (c) crop seasons.....	116
Figure 5.2 Partial Class View of Ontology1	117
Figure 5.3 Partial Class View of Ontology2	118
Figure 5.4 WordNet Hierarchy for crop.....	122

Figure 5.5 Screenshot of WordNet Hierarchy from web interface for crop	123
Figure 5.6 Example Formal Context, \mathbb{K}_1	132
Figure 5.7 Concept lattice diagram for \mathbb{K}_1	136
Figure 5.8 Partial View of Source Ontology O1	139
Figure 5.9 Partial View of Source Ontology O2	140
Figure 5.10 Partial view of \mathbb{K}_{ins}	141
Figure 5.11 Concept Lattice for part of \mathbb{K}_{ins} shown in Figure 5.10	141
Figure 5.12 Proposed Ontology Merging Algorithm	142
Figure 5.13 Partial view of Input_O1	145
Figure 5.14 Partial view of Input_O2	146
Figure 5.15 Partial view of Output_Proposed	147
Figure 5.16 Partial View of Out_Proposed	148

LIST OF TABLES

Table 2.2.1 Comparison of Ontology Merging Tools.....	45
Table 3.1 Regular Expressions used in the algorithm along with the corresponding patterns	63
Table 3.2 Difference between Actual Precision and Estimated Precision using Random Sample and Collection Text.....	68
Table 3.3 Estimation of Precision using Hypergeometric Distribution	69
Table 3.4 Precision and Recall of RENT, RAKE, TermRaider and TerMine	71
Table 4.1 Example output for a recent Self-supervised relation extraction scheme	76
Table 4.2 Position Vectors and Recency Vectors for 20 terms.....	81
Table 4.3. Table showing 20 words along with their Starting Points	82
Table 4.4 Table showing Mean and Standard Deviation of 20 terms	82
Table 4.5 Euclidean Distance between Example pairs	83
Table 4.6 Table showing 20 Terms along with their Frequencies of Occurrence	84
Table 4.7 Table showing similarity between different synsets of ‘agriculture’.....	87
Table 4.8 Table showing similarity between different synsets of ‘agriculture’ and ‘farming’	88
Table 4.9 Related Terms and Text occurring between them	90
Table 4.10 Sample Relationships identified.....	92

Table 4.11 Synonyms set obtained using Constraint 1 with their position vectors	95
Table 4.12 $dt1, t2$ values for four pairs of terms.....	96
Table 4.13 Examples for which $is_type_of([t1\ t2], t2)$ holds.....	97
Table 4.14 Patterns and example text for is_a relation.....	98
Table 4.15 Example text for $is_intercrop$ relation.....	99
Table 4.16 Results of RelExOnt on 10 Random Samples of Data.....	101
Table 4.17 List of relations and the corresponding subdomains.....	103
Table 4.18 Precision and Recall for the three relations using Baseline algorithm	107
Table 4.19 Precision and Recall for three relations using Modified algorithm	110
Table 5.1 Simple String-based Similarity Measures.....	120
Table 5.2 Li's Similarity values for Example Terms	124
Table 5.3 Cosine Similarity Values for Example Terms	126
Table 5.4 Comparison of word embedding models for identification of similar terms.....	130
Table 5.5 Derivation operators for formal context of Figure 5.6.....	134
Table 5.6 Comparison of Ontology metrics.....	144