AUTOMATIC DEVELOPMENT OF ONTOLOGY IN AGRICULTURE DOMAIN

NEHA KAUSHIK

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 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:

Date:

Prof. Niladri Chatterjee
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 स्वत: 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 अर्थात इंट्रा-सबडोमेन और इंटर-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

LIST OF FIGURES

LIST OF TABLES