DISCOVERING RELATIONSHIPS AND PATHS IN KNOWLEDGE GRAPHS

PUNEET AGARWAL

AMAR NATH AND SHASHI KHOSLA SCHOOL OF INFORMATION TECHNOLOGY,

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JUNE 2019

DISCOVERING RELATIONSHIPS AND PATHS IN KNOWLEDGE GRAPHS

by

PUNEET AGARWAL

AMAR NATH AND SHASHI KHOSLA SCHOOL OF INFORMATION TECHNOLOGY

Submitted

in fulfilment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JUNE 2019

Certificate

This is to certify that the thesis titled **DISCOVERING RELATIONSHIPS AND PATHS IN KNOWLEDGE GRAPHS** being submitted by **Mr. PUNEET AGARWAL** for the award of **Doctor of Philosophy** in Computer Science and Engineering is a record of bona fide work carried out by him under my guidance and supervision at the Amar Nath and Shashi Khosla School of Information Technology, Indian Institute of Technology Delhi.

The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Maya Ramanath
Associate Professor
Department of Computer Science and Engineering
Indian Institute of Technology Delhi
New Delhi - 110016

Gautam Shroff
Chief Scientist and Vice President
Tata Consultancy Services Ltd.
Gurgaon - 110016

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisers Prof. Maya Ramanath and Dr. Gautam Shroff for their continuous support of my Ph.D. study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me at the time of research and writing of this thesis. I could not have imagined having better advisors and mentors for my Ph.D. study.

Besides my advisors, I would also like to thank the rest of my thesis committee: Prof. Amithabha Bagchi, Prof. B. Chandra, Prof. Aditeshwar Seth, and Prof. Sanjiva Prasad, for their insightful comments and encouragement, and also for the thought provoking questions which incented me to widen my research from various perspectives.

I thank Dr. Gautam Shroff, in his role as supervisor at my workplace in TCS Research, not only for motivating me to join the Ph.D. program, but also for his continuous guidance during the entire tenure. I also want to thank all the other colleagues at work in TCS Research, the insightful discussions with them have been a key source of learning for me. Most importantly, I express my gratitude to my employer Tata Consultancy Services Ltd. for allowing me to pursue Ph.D. and for the travel support related to my publications.

I also thank my fellow labmates in IIT Delhi, for the stimulating discussions, and support during entire tenure.

Last but not the least, I would like to thank my family: my parents for their love and encouragement, my wife Shilpi for managing everything at home when I was busy, and specially my daughter Navya for allowing me to invest my family time towards this endeavor. Without their immense support, it would not have been possible for me to attempt this enlightening journey.

Puneet Agarwal

Abstract

The use of knowledge graphs has proliferated into many different industrial applications in recent years, for example, in knowledge management systems, in infrastructure management systems, for the purpose of fraud detection, etc. Querying the knowledge graphs has become even more important because of such widespread industrial applications. We are interested in special types of knowledge graph queries where-in all the query keywords do not occur in the same entity, and the expected answer entity (one or more) may not contain any of the query keywords. The answer entities are connected to the entities which contain the keywords via various relationships of the knowledge graphs. As a result, standard techniques of document search are not applicable for this problem of querying the knowledge graphs, mainly because those techniques rank the documents which contain the query keywords.

In this thesis, we analyze two types of knowledge graph queries, '*relationship queries*', and '*factoid queries*'. The first one comprises a set of keywords often indicating different entities, and the second one is a natural language sentence that mentions one or more entities. Against both of these types of queries, the answer entities may not contain the query keywords, they (answer entities) may however be connected to the entities containing the query keywords in the knowledge graph.

Relationship queries have been studied for many years, using various terminologies, e.g., keyword search, Steiner tree in a graph etc., the solutions proposed in the literature so far either don't guarantee the optimality of answers retrieved or don't utilize distributed parallel processing paradigm. Such an approach can be used for scaling relationship queries to large graphs having millions of nodes and edges, such graph are now publicly available in the form of 'linked open data'(LOD). In this thesis, we present an algorithm for distributed keyword search (DKS) on large graphs, based on the graph parallel computing paradigm Pregel. We also present a proof of correctness of our algorithm. Even if terminated early, our algorithm produces approximate answers along with bounds. We describe an optimized implementation of our algorithm along with time-complexity analysis. Finally, we report experimental results on LOD data, and demonstrate efficiency of our approach on large graphs.

Further, answering natural language questions posed on a knowledge graph requires traversing an appropriate sequence of relationships starting from the mentioned entities. To answer complex queries, we often need to traverse more than two relationships. Traditional approaches traverse at most two relationships, as well as typically first retrieve

candidate sets of relationships using indexing etc., which are then compared via machine-learning. Such approaches rely on the textual labels of the relationships, rather than the structure of the knowledge graph. In this thesis, we present a novel approach KG-REP that directly predicts the embeddings of the target relationships against a natural language query, avoiding the candidate retrieval step, using a sequence to sequence neural network. Our model takes into account the knowledge graph structure via novel entity and relationship embeddings. We release a new dataset containing complex queries on a public knowledge graph that typically require traversal of as many as four relationships to answer. We also present a new benchmark result on a public dataset for this problem.

Finally, we conclude with a perspective on open problems in this domain, and highlight how the advances in this field can impact general capability of computational systems.

सार

नॉलेज ग्राफ का उपयोग हाल के वर्षों में कई अलग-अलग औद्योगिक अनुप्रयोगों में किया गया है, उदाहरण के लिए, नॉलेज प्रबंधन प्रणालियों में, बुनियादी ढांचा प्रबंधन प्रणालियों में, धोखाधड़ी का पता लगाने के उद्देश्य से, आदि। व्यापक औद्योगिक अनुप्रयोगों के कारण से नॉलेज ग्राफ को क्वेरी करना और भी महत्वपूर्ण हो गया है। हम विशेष प्रकार के नॉलेज ग्राफ प्रश्नों में रुचि रखते हैं जहां-सभी क्वेरी कीवर्ड एक ही इकाई में नहीं होते हैं, और अपेक्षित उत्तर इकाई (एक या अधिक) में कोई क्वेरी कीवर्ड नहीं हो सकता है। उत्तर इकाइयाँ उन संस्थाओं से जुड़ी होती हैं जिनमें नॉलेज ग्राफ के विभिन्न संबंधों के माध्यम से खोजशब्द होते हैं। नतीजतन, दस्तावेज़ खोज की मानक तकनीक नॉलेज ग्राफ़ को क्वेरी करने की इस समस्या के लिए लागू नहीं होती है, मुख्यतः क्योंकि उन तकनीकों में दस्तावेज़ होते हैं जो क्वेरी कीवर्ड होते हैं।

इस थीसिस में, हम दो प्रकार के नॉलेज ग्राफ प्रश्नों का विश्लेषण करते हैं, '*रिलेशनशिप* क्वेरी ' , और '*फक्टोइड क्वेरी*'। पहले वाले में अक्सर अलग-अलग संस्थाओं को दर्शाने वाले कीवर्ड का एक सेट होता है, और दूसरा एक प्राकृतिक भाषा का वाक्य है जिसमें एक या अधिक संस्थाओं का उल्लेख होता है। इन दोनों प्रकार के प्रश्नों के विरुद्ध, उत्तर निकाय में क्वेरी कीवर्ड नहीं हो सकते हैं, वे (उत्तर इकाइयाँ) हालांकि नॉलेज ग्राफ़ में क्वेरी कीवर्ड वाली संस्थाओं से जुड़े हो सकते हैं।

विभिन्न शब्दावली का उपयोग करते हुए कई वर्षों से संबंध प्रश्नों का अध्ययन किया गया है, उदाहरण के लिए, कीवर्ड खोज, ग्राफ़ में स्टेनर ट्री आदि, साहित्य में प्रस्तावित समाधान या तो पुनर्प्राप्त किए गए उत्तरों की इष्टतमता की गारंटी नहीं देते हैं या वितरित उपयोग नहीं करते हैं। समानांतर प्रसंस्करण प्रतिमान। इस तरह के दृष्टिकोण का उपयोग लाखों से अधिक नोड्स और किनारों वाले बड़े ग्राफ़ के संबंध प्रश्नों को स्केल करने के लिए किया जा सकता है, ऐसे ग्राफ़ अब सार्वजनिक रूप से `लिंक्ड ओपन डेटा '(एलओडी) के रूप में उपलब्ध हैं। इस थीसिस में, हम रेखांकन समानांतर कंप्यूटिंग प्रतिमान के आधार पर बड़े रेखांकन पर वितरित कीवर्ड खोज (DKS) के लिए एक एल्गोरिथ्म प्रस्तुत करते हैं। हम अपने एल्गोरिथ्म की शुद्धता का प्रमाण भी प्रस्तुत करते हैं। यहां तक कि अगर जल्दी समाप्त कर दिया, हमारे एल्गोरिथ्म सीमा के साथ अनुमानित जवाब पैदा करता है। हम समय-जटिलता विश्लेषण के साथ-साथ हमारे एल्गोरिथ्म के एक अनुकूलित कार्यान्वयन का वर्णन करते हैं। अंत में, हम एलओडी डेटा पर प्रयोगात्मक परिणामों की रिपोर्ट करते हैं, और बड़े रेखांकन पर हमारे दृष्टिकोण की दक्षता प्रदर्शित करते हैं।

इसके अलावा, नॉलेज ग्राफ पर रखे गए प्राकृतिक भाषा के सवालों का जवाब देने के लिए उल्लेखित संस्थाओं से शुरू होने वाले रिश्तों का एक उचित क्रम निर्धारित करना होगा। जटिल प्रश्नों का उत्तर देने के लिए, हमें अक्सर दो से अधिक रिश्तों का पता लगाना पड़ता है। पारंपरिक दृष्टिकोण अधिकांश दो रिश्तों पर चलते हैं, साथ ही आमतौर पर सबसे पहले अनुक्रमित आदि का उपयोग करते हुए रिश्तों के उम्मीदवार सेटों को पुनः प्राप्त करते हैं, जिनकी तुलना मशीन-लर्निंग के माध्यम से की जाती है। इस तरह के दृष्टिकोण नॉलेज ग्राफ की संरचना के बजाय रिश्तों के शाब्दिक लेबल पर निर्भर करते हैं। इस थीसिस में, हम केजी-आरईपी को एक उपन्यास दृष्टिकोण प्रस्तुत करते हैं जो सीधे एक प्राकृतिक भाषा क्वेरी के खिलाफ लक्ष्य संबंधों के एम्बेडिंग की भविष्यवाणी करता है, उम्मीदवार पुनर्प्राप्ति चरण से बचते हुए, एक अनुक्रम का उपयोग करके तंत्रिका नेटवर्क का अनुक्रम करता है। हमारे मॉडल नॉलेज ग्राफ संरचना को उपन्यास इकाई और संबंध एम्बेडिंग के माध्यम से ध्यान में रखते हैं। हम एक नए डेटासेट को एक सार्वजनिक नॉलेज ग्राफ पर जटिल प्रश्नों से मुक्त करते हैं, जिसका जवाब देने के लिए आम तौर पर चार रिश्तों के ट्रावेल की आवश्यकता होती है। हम इस समस्या के लिए एक सार्वजनिक डेटासेट पर एक नया मानदंड परिणाम भी प्रस्तुत करते हैं।

अंत में, हम इस डोमेन में खुली समस्याओं पर एक परिप्रेक्ष्य के साथ निष्कर्ष निकालते हैं, और इस बात पर प्रकाश डालते हैं कि कैसे इस क्षेत्र में उन्नति कम्प्यूटेशनल सिस्टम की सामान्य क्षमता को प्रभावित कर सकती है।

Contents

List of Figures

List of Tables