TECHNIQUES FOR EFFECTIVE SEARCH AND EXPLORATION OVER KNOWLEDGE GRAPHS

Madhulika Mohanty

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
JULY 2020

TECHNIQUES FOR EFFECTIVE SEARCH AND EXPLORATION OVER KNOWLEDGE GRAPHS

by

MADHULIKA MOHANTY

Department of Computer Science and Engineering

Submitted

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

to the

Indian Institute of Technology Delhi

JULY 2020

Certificate

This is to certify that the thesis titled **TECHNIQUES FOR EFFECTIVE SEARCH AND EXPLORATION OVER KNOWLEDGE GRAPHS** being submitted by **Ms. MADHULIKA MOHANTY** for the award of **Doctor of Philosophy** in Computer Science and Engineering is a record of bona fide work carried out by her under my guidance and supervision at the Department of Computer Science and Engineering, 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

Acknowledgements

When I started out with PhD, I was freshly out of my masters and was very enthusiastic to study further. Little did I know back then that the journey I would embark upon would change my life forever. My PhD has been full of ups and downs but it has enriched me and taught me some wonderful life lessons. There are so many people who have been a part of this bittersweet journey and have contributed in their big or small ways that these few pages cannot suffice to thank them enough but I still make a humble attempt.

Firstly, I would like to thank my advisor, Maya Ramanath, for her guidance over the past years. I was a complete novice when I started out and she has painstakingly steered me throughout the PhD into the world of research. She rarely hand-held but her pointers during various stages of the work ensured that I could chart my own course to become independent and self-reliant. Discussions with her have helped me pull through the many lows during the PhD and her words of encouragement have helped me imbibe self-confidence. Above all, I thank her for being extremely patient with me at all times.

I thank my SRC members, Parag Singla, Gautam Shroff, and Amit Kumar, for their discussions and feedback which have helped me improve my work. I especially thank Parag Singla for asking many stimulating questions during the end sem presentations which eventually made me think deeper and helped refine my understanding and results.

I am also grateful to my thesis examiners, Senjuti Basu Roy and Arnab Bhattacharya, for their detailed comments and suggestions that helped me improve my work substantially and strengthened my contributions.

I would like to thank Gerhard Weikum and IMPECS for the internship opportunity at MPI-INF, Saarbrücken. Gerhard is a luminary researcher and I hope to be able to imbibe his level of discipline in life someday. The work I started there is one of the main contributions of my thesis and the experiences from there are going to last the lifetime. The friendships forged in Saarbrücken are few of the closest during my PhD. I hope to stay in touch with them even in the future.

I would also like to thank my colleagues at IIT Delhi. I do not want to list any names here as I fear I might unintentionally miss some of them. However, I do thank each of them for having touched my life in their own ways. All those times of distress and joys, shared over cups of caffeine, will be fond memories for years to come.

The most important pillar of my life has been my family – Mom, Dad, Riti, my maternal aunts – Swati and Muni, and Bapi. They have constantly encouraged me in all my pursuits. Whatever I achieve in life will always be due to their constant support and motivation at all stages. I hardly got to go home in the past few years and they have silently missed me but always ensured that I worked hard with dedication. I hope to spend more time with you all, henceforth. It is unfortunate that Maa is not here to see this day, but I am sure wherever she is, she is very proud of me.

I also thank my in-laws for being supportive of my career. I am immensely indebted to them for being considerate when I would skip family events due to deadlines. I especially thank my mother-in-law for shouldering many responsibilities on her own while I was busy with the thesis. This helped me focus exclusively on my work. They are as proud of me as my parents and I am fortunate to have them as my in-laws.

Lastly, I would like to thank my best friend and husband, Hari. No words can express what his unconditional love and unwavering faith meant to me during the trying times. If it were not for his moral support and understanding, I might have ended giving up on my PhD. He is highly dedicated to his own work and is able to multitask with household responsibilities without even as much as a sigh. Having him around has made life so much simpler. I especially acknowledge his support during the thesis writing period where he

managed everything to enable me to give undivided attention to writing. His constant reassurances instilled confidence in me to work harder with each failure and each joy turned manifold when shared with him. I believe this thesis belongs as much to him as it does to me, and lovingly dedicate the thesis to him.

Madhulika Mohanty

Abstract

Knowledge Graphs (KGs) such as YAGO, DBPedia, and Freebase, form the backbone for applications including chatbots, personal assistants and question answering systems. They are constructed using Information Extraction (IE) techniques over various structured and unstructured sources. KGs represent the information in the form of a graph consisting of nodes to denote entities and edges to denote relationships between these entities. They are typically stored as Resource Description Framework (RDF) triples of subject, predicate and object, with subject and object being entities and the predicate denoting the relationship between them. Each of these triples is also associated with a score during the extraction process to denote the confidence of the IE technique in the accuracy of the information represented by the triple.

KGs are queried using either structured queries or relationship queries. Structured queries comprise of triple patterns having variables in the subjects, predicates or objects. Each triple pattern consists of at least one variable and queries for triples matching the constants, and variables are bound to the corresponding values in these triples. Each set of triples matching the triple patterns in a query forms a subgraph of the original KG

and is returned as an answer. Relationship queries comprise of unstructured keywords. Each keyword maps to multiple keyword nodes in the KG that have the keyword as a term in their labels. An answer to a relationship query is an interconnection between the keyword nodes (one for each keyword) and denotes how the queried nodes are related to each other. The answer graphs are ranked in decreasing order of their relevance to show only the top-k (for some value of k) most relevant answers to the users. The relevance is measured by assigning a score to each answer using a scoring mechanism.

Many of these KGs are huge in size having millions of nodes and edges. Hence, querying them effectively is non-trivial. Users querying KGs can range from beginners looking to casually explore the system without a particular information need in mind, to expert users querying for specific information needs. A common problem faced by the users is getting *empty* results. This is because of their lack of knowledge of the exact labels over nodes and edges in the KG. Also, since KGs allow schemaless addition of information in the form of triples, a given information may be represented by a variety of triples. Thus, structured queries seeking exact matches face the issue of *poor recall*. That is, *all* the desired answers are not fetched as the sought information is present in multiple forms and an exact match is not found with them. On getting unsatisfactory results, users try to relax their queries preserving original query intent. Nevertheless, coming up with useful relaxations is also challenging for these users. In this thesis, we propose *Insta-Search*, an interactive system which helps alleviate these issues by aiding the users in querying KGs. Insta-Search helps the users by providing autocompletion of partially entered query, giving instant feedback on the results that the current query would fetch and suggesting

possible relaxations. We evaluate the response times for each module to demonstrate that Insta-Search helps users at interactive speeds.

Insta-Search also supports automatic relaxations for structured queries to tackle poor recall caused by exact match semantics. Since the space of relaxations is quite large and users seek only top-k answers, existing systems employ top-k operators for early termination. However, they still process *all* the relaxations, many of whose matches do not contribute triples towards the top-k answers. We propose *Spec-QP* to eliminate this inefficiency by using a speculative approach to prune the relaxations which are unlikely to contribute triples towards the top-k answers. It makes use of precomputed statistics about the scores of the triples to speculate on the requirement of relaxations. We compare Spec-QP with an existing baseline to demonstrate its efficiency and accuracy.

On query submission, Insta-Search evaluates it to return top-k ranked results. Many of these results may be of the same kind and the user has to scroll through them to find a new answer. We propose *KlusTree*, which post-processes these results to present a summary of them. It uses a novel language-model (LM) based representation for the answer graphs for clustering the results based on the information conveyed by them. This enhances the result diversity in the top answers. We compare KlusTree with existing techniques for clustering graphs to show the advantages of using the LM based representation for the results. We compare with the renowned diversity-based reranking technique, Maximum Marginal Relevance (MMR), to show that KlusTree provides diversified results.

Hence, this thesis provides an integrated holistic system to facilitate smooth and effective exploration of KGs.

सार

नॉलेज ग्राफ्स (के.जी.) जैसे की यागो, डीबीपीडिआ और फ्रीबेस, अनेक अनुप्रयोगों, जैसे की चैटबॉट, व्यक्तिगत सहायक और प्रश्न उत्तर प्रणाली, के लिए मूल आधार है। वे विभिन्न संरचित और असंरचित स्रोतों पर इनफार्मेशन एक्सट्रैक्शन (आई.ई.) तकनीकों का उपयोग करके निर्मित किए जाते हैं। के.जी. एक ग्राफ के रूप में सूचना का प्रतिनिधित्व करता है -- नोड्स वस्तुएं और एड्जेस संबंधों को दर्शाते हैं। वे आमतौर पर रिसोर्स डिस्क्रिप्शन फ्रेमवर्क (आर.डी.एफ.) विषय, विधेय और लक्ष्य के तिकड़ी के रूप में संग्रहीत होते हैं, विषय और लक्ष्य वस्तुएं और विधेय उनके बीच संबंधों को निरूपित करते हैं। इन तिकड़ीओं (ट्रिपल्स) में से प्रत्येक तिकड़ी द्वारा प्रस्तुत सूचना की सटीकता में आई.ई. तकनीक के विश्वास को दर्शाने के लिए निष्कर्षण प्रक्रिया के दौरान एक स्कोर भी जोड़ा जाता है।

के.जी. को संरचित प्रश्नों या रिलेशनशिप क्वेरी का उपयोग करके सवाल (क्वेरी) किया जाता है। संरचित प्रश्नों में विषयों, विधेय या लक्ष्य में चर वाले ट्रिपल पैटर्न होते हैं। प्रत्येक ट्रिपल पैटर्न में स्थिरांक से मेल खाने वाले तिकड़ीओं के लिए कम से कम एक चर और प्रश्न होते हैं, और चर इन तिकड़ीओं में संबंधित मानों से बँधते हैं। एक क्वेरी में ट्रिपल पैटर्न से मेल खाने वाले ट्रिपल्स का प्रत्येक सेट मूल के.जी. का एक सबग्राफ बनाता है और एक उत्तर के रूप में वापस आ जाता है। रिलेशनशिप क्वेरी में अनस्ट्रक्चर्ड कीवर्ड होते हैं। प्रत्येक कीवर्ड के.जी. में कई कीवर्ड नोड्स को मैप करता है जिनके कीवर्ड उनके लेबल में एक शब्द के रूप में होते हैं। रिलेशनशिप क्वेरी का उत्तर कीवर्ड नोड्स (प्रत्येक कीवर्ड के लिए एक) के बीच एक अंतर-संबंध है और यह दर्शाता है की कैसे नोड्स एक दूसरे से संबंधित हैं। उत्तर रेखांकन केवल उपयोगकर्ताओं को 'k' सबसे अधिक प्रासंगिक उत्तर ('k' के कुछ मूल्य के लिए) को दिखाने के लिए उनकी

प्रासंगिकता के घटते क्रम में क्रमबद्ध हैं। प्रासंगिकता को स्कोरिंग तंत्र का उपयोग करके प्रत्येक उत्तर के लिए एक अंक प्रदान करके मापा जाता है।

इनमें से कई के.जी. आकार में लाखों नोड्स और एड्जेस वाले हैं। इसलिए, उन्हें प्रभावी ढंग से क्वेरी करना गैर-तुच्छ है। के.जी. की क्वेरी करने वाले उपयोगकर्ता शुरुआती (किसी विशेष जानकारी की आवश्यकता के बिना सिस्टम की खोज करने वाले) से लेकर विशिष्ट सूचना आवश्यकताओं के लिए क्वेरी करने वाले विशेषज्ञ उपयोगकर्ता हो सकते हैं। उपयोगकर्ताओं द्वारा सामना की जा रही एक आम समस्या 'खाली परिणाम' प्राप्त करना है। इसकी वजह केजी में नोड्स और एड्जेस पर सटीक लेबल के ज्ञान की कमी है। इसके अलावा, चूंकि केजी तिकड़ी के रूप में सूचना देता है, इसलिए दी गई जानकारी को विभिन्न प्रकार के तिकड़ीओं द्वारा दर्शाया जा सकता है। इस प्रकार, सटीक मिलान चाहने वाले संरचित प्रश्न 'खराब रिकॉल' के मुद्दे का सामना करते हैं। अर्थात्, 'समस्त' वांछित उत्तर प्राप्त नहीं होते हैं क्योंकि मांगी गई जानकारी कई रूपों में मौजूद होती है और उनके साथ एक सटीक मिलान नहीं मिलता है। असंतोषजनक परिणाम मिलने पर, उपयोगकर्ता अपने प्रश्नों को मूल क्वेरी आशय को संरक्षित करते हुए रिलैक्स करने का प्रयास करते हैं। फिर भी, इन उपयोगकर्ताओं के लिए उपयोगी छूट के साथ आना भी चुनौतीपूर्ण है। इस शोध-निबंध में, हम 'इंस्टा-सर्च' का प्रस्ताव करते हैं, एक इंटरैक्टिव सिस्टम जो उपयोगकर्ताओं को के.जी. को क्वेरी करने में मदद करके इन मुद्दों को कम करने में सहायता करता है। इंस्टा-सर्च उपयोगकर्ताओं को आंशिक रूप से दर्ज किए गए प्रश्नों को स्वत: पूर्णता प्रदान करके मदद करता है, वर्तमान प्रश्न के नतीजे पर तत्काल प्रतिक्रिया देता है और संभव छूट का सुझाव देता है। हम प्रत्येक अवयव के लिए प्रतिक्रिया समय का मूल्यांकन करते हैं कि इंस्टा-सर्च इंटरएक्टिव गति पर उपयोगकर्ताओं की मदद करता है।

इंस्टा-सर्च संरचित प्रश्नों के लिए स्वचालित छूट का भी समर्थन करता है ताकि सटीक मिलान शब्दार्थों के कारण खराब रिकॉल से निपटा जा सके। चूंकि छूट का परिमाण काफी बड़ा है और उपयोगकर्ता केवल शीर्ष-k उत्तर चाहते हैं, इसलिए मौजूदा प्रणालियाँ शीघ्र समाप्ति के लिए शीर्ष- k ऑपरेटरों को नियुक्त करती हैं। हालाँकि, वे अभी भी □समस्त' रिलैक्सेशन की प्रक्रिया करते हैं, जिनके कई मैच शीर्ष-k जवाबों की ओर योगदान नहीं करते हैं। हम एक सट्टा दृष्टिकोण का उपयोग करके इस अक्षमता को समाप्त करने के लिए 'स्पेक-क्यू.पी.' का प्रस्ताव करते हैं, जो शीर्ष-k उत्तरों की

ओर ट्रिपल योगदान करने की संभावना ना रखने वाले छूट को छांटता है। यह छूट की आवश्यकता पर अटकलें करने के लिए तिकड़ीओं के स्कोर के बारे में पूर्वनिर्मित आंकड़ों का उपयोग करता है। हम इसकी दक्षता और सटीकता को प्रदर्शित करने के लिए एक मौजूदा आधारभूत के साथ 'स्पेक-क्यू.पी.' की तुलना करते हैं।

उपयोगकर्ता द्वारा अपनी क्वेरी सबमिट किए जाने के बाद, इंस्टा-सर्च शीर्ष-k रैंक परिणामों को वापस करने के लिए इसका मूल्यांकन करता है। चूंकि इनमें से कई परिणाम एक ही तरह के हो सकते हैं, उपयोगकर्ता एक नया उत्तर खोजने से पहले उन पर स्क्रॉल करने के दौरान निराश होता है। हम 'क्लसट्री' को प्रस्तावित करते हैं, जो इनका सारांश प्रस्तुत करने के लिए इन परिणामों को पश्चात्-संसाधित करता है। यह उनके द्वारा व्यक्त की गई जानकारी के आधार पर परिणामों के क्लस्टरिंग के लिए उत्तर रेखांकन के लिए एक नव भाषा-मॉडल (एल.एम.) आधारित प्रतिनिधित्व का उपयोग करता है। यह शीर्ष उत्तरों में परिणाम विविधता को बढ़ाता है। हम ग्राफ क्लस्टरिंग के लिए मौजूदा तकनीकों के साथ क्लसट्री की तुलना करते हैं और परिणामों के लिए एल.एम. आधारित प्रतिनिधित्व का उपयोग करने के लाभों को प्रदर्शित करते हैं। हम प्रसिद्ध विविधीकरण तकनीक, मैक्सिमम रेलेवेंस रैंकिंग (एम. एम. आर.) आधारित रैंकिंग, के साथ तुलना करते हैं, यह दिखाने के लिए कि क्लसट्री विविध परिणाम प्रदान करता है।

अत:, यह शोध-निबंध के.जी. के सुचारू और प्रभावी अन्वेषण की सुविधा के लिए एक एकीकृत समग्र प्रणाली प्रदान करता है।

Contents

List of Figures

List of Tables