CONSTRUCTION AND APPLICATIONS OF A TECHNICAL KNOWLEDGE BASE

PRAJNA DEVI UPADHYAY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

AUGUST 2021

CONSTRUCTION AND APPLICATIONS OF A TECHNICAL KNOWLEDGE BASE

by

PRAJNA DEVI UPADHYAY

Department of computer science & Engineering

Submitted

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

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

AUGUST 2021

Certificate

This is to certify that the thesis titled **Construction and Applications of a Technical Knowledge Base** being submitted by **Ms. Prajna Devi Upadhyay** 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

I would like to thank my supervisor, Prof. Maya Ramanath, for her invaluable guidance and feedback that has led to the work in this thesis. She has always maintained very high standards for research and has been extremely critical and objective during our discussions. This has given me the opportunity to improve my research methods. I also thank her for the endless comments she provided on my writing and presentations which have helped me refine them. Her words of encouragement during trying times have boosted me to carry on. Working with her has also helped me being an independent researcher since she gave me the freedom to try new ideas, write and submit them on my own.

I would also like to thank my collaborators, Prof. Srikanta Bedathur and Prof. Tanmoy Chakraborty, for the enriching discussions that led to one of the major components of my Ph.D. I also thank my co-authors—Ashwini Purkar, Tanuma Patra, Ashutosh Bindal, and Manjeet Kumar for their hard work.

I thank my SRC chairpersons, Prof. Saroj Kaushik and Prof. Mausam, and committee members Prof. Parag Singla and Dr. L Venkata Subramaniam for their insightful queries and comments that helped me refine my ideas.

I would like to extend my thanks to the examiners of my thesis for providing me with valuable feedback on my manuscript.

I thank IBM AI Horizons Network and IITD Alumni Association for supporting my travel

to conferences.

I want to thank my friends in the lab—Madhulika, Neha, Chandrika, and Anirban, who were always available for discussions whenever I needed clarity. My friends in the hostel—Swati, Shubhrima, Akanksha, Poonam, and Lopa made this long and difficult journey easier to deal with.

I will forever remain indebted to Dipanjan, my husband, who has been extremely supportive in various stages of my Ph.D. I thank him for putting up with my mood swings, for pulling me out of them, for motivating me, and for celebrating my accomplishments. This thesis would not have been possible without him being by my side.

I have been very fortunate to have extremely loving and caring in-laws, and I thank them wholeheartedly.

I cannot thank my parents enough—they have been supportive throughout my long journey as a Ph.D. student, and have always celebrated my achievements. I thank my sisters—Ambika and Pratima for taking up responsibilities at home when I could not be around because of work commitments. I thank Pushkar—my little brother, who never fails to amaze me with his energy.

All of you inspire me to strive for betterment and aim for higher goals in research.

Prajna Devi Upadhyay

Abstract

Recent years have witnessed an exponential increase in the availability of technical material on the web, possibly due to the efforts to digitize education and open access to scientific publications. Anybody with an internet connection can pull up relevant technical material to study or research. However, it is not necessary that all of this readily available material is comprehensible to most users. A user who is interested to work on a research problem or a student willing to take up a new course may come across multiple topics about which she has little/no idea. She may find a number of learning materials on that topic on the web, however, due to the lack of prerequisite knowledge, she will have to perform multiple searches before obtaining a basic understanding of that topic. To keep up with the latest research, she has to identify and understand different aspects of the topic to conduct a survey. This can be overwhelming for her due to a large amount of technical material available on the web. It would be helpful if there was a system to recommend to her prerequisite concepts for basic understanding and research papers for advanced understanding of the topic.

To build such a system, we first have to store the knowledge from the technical do-

main in the form of entities and relations as a knowledge graph. This will help design applications to consume this information in a systematic way. Knowledge graphs have been crucial for a number of semantic-aware applications. A number of such knowledge graphs, such as Yago, DBPedia, NELL, Wikidata, or Freebase have been constructed in the open domain and have supported tasks such as entity retrieval, question answering, or automatic organization of topics. Although there are multiple open-domain knowledge graphs, there are no known large-scale technical knowledge graphs, and especially in the domain of Computer Science. So, the first contribution of this thesis is TeKnowbase, which is a knowledge base in the domain of Computer Science. We use a combination of information extraction techniques to extract entities and relations from both structured and unstructured sources. TeKnowbase has been evaluated for its quality and is freely available.

The second contribution of this thesis is PreFace, which assists a beginner in the study of a topic in Computer Science by identifying its prerequisites using TeKnowbase. PreFace takes a topic in Computer Science as the query and returns a prerequisite graph, where the nodes represent the prerequisite concepts for the topic and the edges represent the prerequisite relationship. Additionally, it also identifies interesting aspects for the query and returns prerequisites grouped together as facets for each aspect. It achieves this by estimating a language model for the facet as well as the query using TeKnowbase. The facets are then ranked based on their relevance to the query and their coverage. The prerequisite graph, as well as the facets generated by PreFace, have been evaluated to be better than those generated using state-of-the-art prerequisite and facet retrieval

techniques.

Our final contribution is ASK (Aspect-based academic Search using domain-specific KBs), which recommends research papers for the different aspects identified by PreFace for advanced understanding of the queried topic. ASK first assists the user by providing relevant query suggestions for query and the aspect and then returns a ranked list of relevant research papers. The ranking of query suggestions, as well as research papers, is achieved using language models estimated for the query and aspect using TeKnowbase. The evaluation of papers as well as suggestions retrieved by ASK showed that they were superior to those returned by various state-of-the-art pseudo-relevant feedback or diversification techniques.

Overall, this thesis proposes techniques to make technical information more comprehensible for a user.

सारांश

हाल के वर्षों में वेब पर तकनीकी जानकारी की उपलब्धता में घातीय वृद्धि देखी गई है, संभवतः शिक्षा को डिजिटाइज़ करने और वैज्ञानिक प्रकाशनों की खुली पहुंच के प्रयासों के कारण। इंटरनेट कनेक्शन वाला कोई भी व्यक्ति अध्ययन या शोध के लिए प्रासंगिक तकनीकी सामग्री प्राप्त कर सकता है। हालांकि, यह आवश्यक नहीं है कि यह सभी आसानी से उपलब्ध सामग्री अधिकांश उपयोगकर्ताओं के लिए समझ में आ जाए। एक उपयोगकर्ता जो एक शोध समस्या पर काम करने के लिए इच्छुक है या एक नया पाठ्यक्रम लेने के इच्छुक छात्र को ऐसे कई विषयों का सामना करना पड़ सकता है जिनके बारे में उसे बहुत कम/कोई जानकारी नहीं है। उसे वेब पर उस विषय पर कई शिक्षण सामग्री मिल सकती है, हालांकि, पूर्वापेक्षित ज्ञान की कमी के कारण, उस विषय की बुनियादी समझ प्राप्त करने से पहले उसे कई खोज करनी होगी। नवीनतम शोध के साथ बने रहने के लिए, उसे सर्वेक्षण करने के लिए विषय के विभिन्न पहलुओं को पहचानना और समझना होगा। वेब पर बड़ी मात्रा में तकनीकी सामग्री उपलब्ध होने के कारण यह उसके लिए भारी पड़ सकता है। यह सहायक होगा यदि विषय की बुनियादी समझ के लिए पूर्वापेक्षा अवधारणाओं और उन्नत समझ के लिए शोध पत्रों की सिफारिश करने के लिए एक प्रणाली है।

ऐसी प्रणाली बनाने के लिए, हमें पहले तकनीकी डोमेन से ज्ञान को एन्टीटीएस और रिलेशन्स के रूप में ज्ञान ग्राफ के रूप में संग्रहीत करना होगा। यह इस जानकारी को व्यवस्थित तरीके से उपभोग करने के लिए अनुप्रयोगों को डिजाइन करने में मदद करेगा। कई अर्थ-जागरूक अनुप्रयोगों के लिए ज्ञान ग्राफ महत्वपूर्ण रहे हैं। ऐसे कई नॉलेज ग्राफ , जैसे यागो, डीबीपीडिया, एनईएल, विकीडाटा, या फ्रीबेस का निर्माण खुले डोमेन में किया गया है और इसमें एंटिटी रिट्रीवल, प्रश्न उत्तर, या विषयों के स्वचालित संगठन जैसे कार्यों का समर्थन किया गया है। हालांकि कई खुले डोमेन नॉलेज ग्राफ हैं, लेकिन बड़े पैमाने पर कोई टेक्निकल नॉलेज ग्राफ नहीं

एन्टीटीएस और रिलेशन्स को निकालने के लिए सूचना निष्कर्षण तकनीकों के संयोजन का उपयोग करते हैं। टेक्नोबैस का मूल्यांकन इसकी गुणवत्ता के लिए किया गया है और यह मुफ़्त में उपलब्ध है।

इस थीसिस का दूसरा योगदान प्रीफेस है, जो कंप्यूटर विज्ञान में किसी विषय के अध्ययन में शुरुआती को टेक्नोबैस का उपयोग करके इसकी पूर्वापेक्षाओं की पहचान करने में सहायता करता है। प्रीफेस कंप्यूटर साइंस में एक विषय को क्वेरी के रूप में लेता है और एक पूर्वापेक्षा ग्राफ देता है, जहां नोड्स विषय के लिए पूर्वापेक्षा अवधारणाओं का प्रतिनिधित्व करते हैं और किनारे पूर्वापेक्षा संबंध का प्रतिनिधित्व करते हैं। इसके अतिरिक्त, यह क्वेरी के लिए दिलचस्प पहलुओं की भी पहचान करता है और प्रत्येक पहलू के लिए पहलुओं के रूप में समूहीकृत पूर्वापेक्षाएँ लौटाता है। यह पहलू के लिए लैंग्वेज मॉडल के साथ-साथ टेक्नोबैस का उपयोग करके क्वेरी का अनुमान लगाकर इसे प्राप्त करता है। फिर पहलुओं को उनकी प्रासंगिकता और उनके कवरेज के आधार पर रैंक किया जाता है। पूर्वापेक्षा ग्राफ, साथ ही प्रीफेस द्वारा उत्पन्न पहलुओं का मूल्यांकन अत्याधुनिक पूर्वापेक्षा और पहलू पुनर्प्राप्ति तकनीकों का उपयोग करके उत्पन्न लोगों की तुलना में बेहतर होने के लिए किया गया है।

हमारा अंतिम योगदान आस्क (डोमेन-विशिष्ट केबी का उपयोग कर पहलू-आधारित अकादमिक खोज) है, जो पूछे गए विषय की उन्नत समझ के लिए प्रीफेस द्वारा पहचाने गए विभिन्न पहलुओं के लिए शोध पत्रों की सिफारिश करता है। आस्क पहले क्वेरी और पहलू के लिए प्रासंगिक क्वेरी सुझाव प्रदान करके उपयोगकर्ता की सहायता करता है और फिर प्रासंगिक शोध पत्रों की एक रैंक की गई सूची देता है। क्वेरी सुझावों के साथ-साथ शोध पत्रों की रैंकिंग, टेक्नोबैस का उपयोग करके क्वेरी और पहलू के लिए अनुमानित लैंग्वेज मॉडल का उपयोग करके हासिल की जाती है। शोध पत्रों के मूल्यांकन के साथ-साथ आस्क द्वारा प्राप्त सुझावों से पता चला कि वे विभिन्न अत्याधुनिक सूडो रिलेवेंट फीडबैक या विविधीकरण तकनीकों द्वारा लौटाए गए सुझावों से बेहतर थे।

Contents

List of Figures

List of Tables