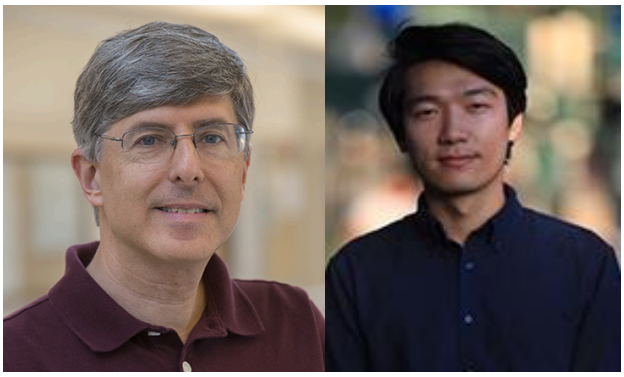
COVID Information Commons (CIC) Research Lightning Talk



Transcript of a Presentation by Steven Skiena and Xingzhi Guo (CUNY Stonybrook), February 2022

Title: Knowledge Graph Embedding Evolution for COVID-19

Funded by NSF Office of Advanced Cyberinfrastructure, Directorate for Computer & Information Science & Engineering (OAC/CISE) through the Northeast Big Data Innovation Hub Seed Fund Program.

Youtube Recording with Slides

February 2022 CIC Webinar Information

Transcript Editor: Saanya Subasinghe

Transcript

Xingzhi Guo:

स्लाइड *1*

तो हमें यहाँ होने के लिए धन्यवाद। मेरा नाम जिंगझी है। मैं चौथे वर्ष का पीएचडी छात्र हूं जो स्टोनी ब्रुक विश्वविद्यालय में प्रोफेसर स्टीवन स्कीना के साथ काम कर रहा है और प्रोफेसर स्कीना स्टोनी ब्रुक एआई संस्थान के निदेशक भी हैं। और यह प्रस्तुति इस बारे में है कि COVID के दौरान ज्ञान ग्राफ एम्बेडिंग कैसे बदलती है क्योंकि हमारी दुनिया बदल गई है। और मुख्य सामग्री SIGKDD 2021 में प्रकाशित एक हालिया पेपर पर आधारित है, जब वह यहां पोस्टडॉक थे, तब बोजियन [झोउ] के साथ सहयोग कर रहे थे।

स्लाइड *2*

ठीक है, तो चलिए आगे चलते हैं। तो सबसे पहले बात। ग्राफ एम्बेडिंग क्या है या अन्य इसे नेटवर्क एम्बेडिंग या नोड एम्बेडिंग कह सकते हैं? तो मूल रूप से, यह एक ऐसा फ़ंक्शन है जो ग्राफ में एक नोड को एक संख्यात्मक वेक्टर में मैप करता है जिसे हम एम्बेडिंग वेक्टर कहते हैं। तो जैसा कि निम्न उदाहरण से पता चलता है, हम इस ग्राफ में प्रत्येक नोड को इस 2D स्पेस में मैप कर सकते हैं। तो एम्बेडिंग मूल ग्राफ में नोड के अर्थ को पकड़ सकता है, लेकिन तथ्य यह है कि ग्राफ में करीब नोड्स में एम्बेडिंग स्पेस में समान वैक्टर हो सकते हैं। तो इस उदाहरण में, आप देख सकते हैं कि एक ही रंग वाले नोड्स भी इस 2D प्लान पर एक-दूसरे से मिलते हैं। और सबसे महत्वपूर्ण बात, इस कम आयाम संख्यात्मक एम्बेडिंग वेक्टर द्वारा, हम कई डाउनस्ट्रीम कार्यों के लिए मौजूदा मशीन लर्निंग एल्गोरिदम लागू कर सकते हैं। उदाहरण के लिए, नोड वर्गीकरण, समुदाय खोज या बाहरी पहचान के लिए नोड क्लस्टरिंग। और,

हालांकि, यह उदाहरण एक स्थिर ग्राफ है जहां कोई नया किनारा नहीं है, कोई नया नोड नहीं है, सब कुछ तय है, लेकिन हमारे वास्तविक जीवन में हमारी दुनिया हमेशा बदल रही है। तो वास्तविक दुनिया रेखांकन और नोड्स करते हैं।

स्लाइड *3*

ठीक है, तो चलिए वास्तविक दुनिया बदलते ग्राफ पर एक नज़र डालते हैं - या हम इसे गतिशील ग्राफ कह सकते हैं। तो अब विकिपीडिया लिंक ज्ञान ग्राफ में, प्रत्येक नोड विकी लेख है जो आमतौर पर वास्तविक दुनिया की संस्थाओं का वर्णन करता है और प्रत्येक किनारा दो लेखों को जोड़ने वाला हाइपरलिंक है, जैसे कि हम एक पेपर लिखते समय उद्धरण करते हैं। तो यह ग्राफ है - यह बड़े पैमाने पर है। हमारे पास लाखों नोड्स और सैकड़ों लाखों किनारे हैं और यह बढ़ता रहता है जैसा कि आप इस छोटे से आंकड़े में देख सकते हैं। और निश्चित रूप से कुछ संस्थाएं कई बदल सकती हैं और इस गतिशील ज्ञान ग्राफ द्वारा कब्जा कर लिया गया है। और मैं वुहान शहर का एक विशिष्ट उदाहरण दिखाना चाहता हूं।

स्लाइड *4*

ठीक है, इसलिए COVID से पहले, वुहान शायद दुनिया भर के लोगों के लिए कम प्रसिद्ध है। और 2019 के अंत और 2020 की शुरुआत में, मुझे लगता है कि ज्यादातर लोग इसे COVID के प्रकोप के पहले स्थान के रूप में जानते थे। और मुझे लगता है - इसलिए मुझे लगता है कि यह या परिवर्तित इकाई के बदलने का एक अच्छा उदाहरण है। तो यहाँ आंकड़ा वुहान के एक विकिपीडिया लेख को दर्शाता है। और मैं इसके हाइपरलिंक को हाइलाइट करता हूं। और पहले दो पैराग्राफ भू [तार्किक] या वुहान से संबंधित कुछ ऐतिहासिक घटनाएं हैं। लेकिन 2019 में, अचानक, हम देखते हैं कि कई COVID संबंधित नए लिंक बनाए गए थे। तो ग्राफ एम्बेडिंग परिप्रेक्ष्य पर वापस, सवाल यह है कि इस गतिशील बड़े पैमाने पर ग्राफ में उन नोड एम्बेडिंग को कुशलतापूर्वक कैसे ट्रैक किया जाए ताकि हम नोड के एम्बेडिंग आंदोलन का पता लगा सकें और समय के साथ इसकी तुलना कर सकें ताकि हम देख सकें कि यह कैसे बदलता है?

स्लाइड *5*

ठीक है, अगला, इसलिए यह प्रश्न हमें एक नया एल्गोरिदम डिजाइन करने के लिए प्रेरित करता है जो एक समस्या को संभाल सकता है जिसे हम गतिशील बड़े ग्राफ़ में सबसेट नोड एम्बेडिंग कहते हैं। तो इस एल्गोरिथ्म का उपयोग करके हम कई पूर्वनिर्धारित नोड्स के एम्बेडिंग को ट्रैक कर सकते हैं। तो यह ग्राफ में चार नोड्स के बजाय नोड्स का एक सबसेट है जिसमें हम रुचि रखते हैं क्योंकि ग्राफ विकसित होता रहता है। तो मुख्य विचार व्यक्तिगत पृष्ठ रैंक का उपयोग करना है जो Google खोज द्वारा उपयोग किया जाने वाला बहुत सफल एल्गोरिदम है और अन्य तरीकों पर एक फायदा यह है कि हम केवल सबसेट नोड्स के सबसेट नोड्स के लिए केवल वही गणना कर सकते हैं जो हमें सबसेट नोड्स के लिए चाहिए, जबकि अधिकांश अन्य एल्गोरिदम को हर बार हर नोड के सभी एम्बेडिंग की गणना करनी होती है। लेकिन वे अंततः उनमें से केवल एक हिस्से का उपयोग करते हैं, इसलिए उनमें से बाकी सिर्फ बर्बाद हो जाते हैं। इसलिए हमारा एल्गोरिथ्म अधिक कुशल और तेज़ है, यह इस समस्या के लिए बहुत उपयुक्त है लेकिन अधिक जानकारी के लिए कृपया हमारे पेपर को देखें, जो अभी यहां हैं [https://arxiv.org/abs/2106.01570]। और जैसा कि यह उदाहरण अवधारणा को दिखाता है, हमारी विधि एक विशिष्ट नोड के एम्बेडिंग की गणना कर सकती है, इस मामले में वुहान, विभिन्न वर्षों में। और हम 2019 से 2020 तक एक विशाल एम्बेडिंग आंदोलन की उम्मीद कर सकते हैं।

स्लाइड *6*

ठीक है, तो अगला - तो चलिए प्रयोग परिणाम देखते हैं। सबसे पहले, हम २०२० में हर दिन अंग्रेजी विकिपीडिया ग्राफ स्नैपशॉट एकत्र करते हैं। और जैसा कि मैंने पहले उल्लेख किया है कि एक बहुत बड़ा गतिशील ग्राफ है। आप इस

तालिका में देख सकते हैं कि हमें हर दिन 30,000 नए किनारे डाले गए हैं और हमने इस GitHub रिपॉजिटरी में डेटा जारी किया है, इसलिए इसे एक्सेस करना बहुत आसान है। और हम अन्य सैकड़ों चीनी शहरों के साथ वुहान पर नज़र रखते हैं। और यह आंकड़ा दिखाता है कि हमने एम्बेडिंग स्पेस में जिन परिवर्तनों का पता लगाया, जैसा कि आप देख सकते हैं, वुहान का वक्र प्रमुख है। उनके पास एक विशाल शिखर है और हम एनोटेशन के साथ कई चोटियों को उजागर करते हैं और पाया कि वे सभी COVID की समयरेखा से संबंधित हैं। और एक और चोटी है – आप देख सकते हैं कि हम चेंगदू शहर में एक और चोटी देखते हैं जो अमेरिका-चीन राजनयिक तनाव को दर्शाता है जब अमेरिका ने चेंगदू में वाणिज्य दूतावास को बंद करने का फैसला किया था। तो यह गंभीरता है जिसे हमने इस अध्ययन में खोजा है।

स्लाइड *7*

तो, उफ़, इसलिए, उह - हमारे पास एक और प्रयोग है जहां हम अलग-अलग समय अवधि में सबसे अधिक बदले हुए शहरों को देखना चाहते हैं। इसलिए हम हर समय ट्रैक किए गए नोड के एम्बेडिंग मूवमेंट को रैंक करते हैं और पाते हैं कि, आप देख सकते हैं, वुहान आमतौर पर दूसरों के बीच शीर्ष पर होता है क्योंकि COVID विकसित होता है। और हम उस समय अवधि में समाचार शीर्षक को उजागर करते हैं ताकि हम अंदाजा लगा सकें कि उस स्थान पर क्या हुआ था।

स्लाइड *8*

ठीक। तो प्रमुख takeaways हैं: हमारे पास एक बहुत ही कुशल नोड एम्बेडिंग एल्गोरिदम है जो बहुत बड़े गतिशील ग्राफ़ में एम्बेडिंग को कैप्चर कर सकता है। फिर हमने 2020 में विकिपीडिया ज्ञान ग्राफ के विकास की जांच की और COVID के दौरान दिलचस्प नोड परिवर्तन पाया। अधिक विवरण के लिए, कृपया हमारे पेपर को देखें और जारी किए गए संसाधनों को देखें। [https://arxiv.org/abs/2106.01570 और https://github.com/zjlxgxz/DynamicPPE] और हाँ, धन्यवाद, धन्यवाद!