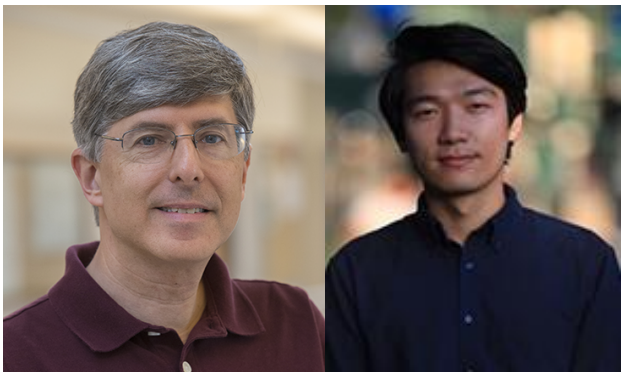
COVID Information Commons (CIC) Research Lightning Talk



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:

Slide 1

So thanks for having us here. My name is Xingzhi. I'm a fourth year Ph.D. student working with Professor Steven Skiena at Stony Brook University and Professor Skiena is also the Director of the Stony Brook AI Institute. And this presentation is about how knowledge graph embedding changes during COVID as our world changed. And the main content is based on a recent paper published at SIGKDD 2021, collaborating with Boajian [Zhou] when he was a postdoc here.

Slide 2

Okay, so let's go next. So first thing first. What is the graph embedding or others may call it a network embedding or node embeddings? So basically, it's a function that maps a node in the graph to a numerical vector which we call an embedding vector. So as the following example shows, we could map every node in this graph to this 2D space. So the embedding could capture the meaning of the node in the original graph but the fact that the closer nodes in the graph may have similar vectors in the embedding space. So in this example, you can see the nodes with the same color also get together to each other on this 2D plan. And most importantly, by this low dimension numerical embedding vector, we could apply existing machine learning algorithms for many downstream tasks. For example, the node

classification, node clustering for the community discovery or the outlier detection. And, however, this example is a static graph where there's no new edge, no new node, everything is fixed, but in our real life our world is always changing. So do the real world graphs and the nodes.

Slide 3

Okay so let's take a look at the real world changing graph - or we can call it a dynamic graph. So now in Wikipedia link knowledge graph, each node is Wiki articles usually describing the real world entities and each edge is the hyperlink connecting two articles, kind of like the citation we do when we're writing a paper. So this graph is - it's large scale. We have millions of nodes and hundreds of millions of edges and it keeps scaling up as you can see in this little figure. And certainly some entities may change many and have been captured by this dynamic knowledge graph. And I want to show one specific example of the city of Wuhan.

Slide 4

Okay, so before COVID, Wuhan is probably less famous to the people around the world. And at the end of 2019 and early 2020, I think most people knew it as the first place of COVID outbreak. And I think - so I think this is a good example of the changing of the or the changed entity. So here the figure shows a Wikipedia article of Wuhan. And I highlight the hyperlinks it has. And the first two paragraphs are the geo[logical] or some historical events related to Wuhan. But in 2019, suddenly, we see many COVID related new links were created. So back to the graph embedding perspective, the question is how to efficiently track those node embeddings in this dynamic massive graph so that we can detect the embedding movement of the node and compare it across time so we can see how this changes?

Slide 5

Okay, next, so this question motivates us to design a new algorithm that can handle a problem that we call a subset node embedding in dynamic large graphs. So using this algorithm we can track the embeddings of several predefined nodes. So it's a subset of nodes instead of the four nodes in the graph that we are interested in as the graph keeps evolving. So the key idea is to use personalized page rank which is very successful algorithm used by Google search and one advantage over other methods is that we could compute only what we need for the subset nodes of - for the subset nodes, while most other algorithms have to compute all embeddings of every node across every time. But they finally use only part of them, so the rest of them are just wasted. So our algorithm is more efficient and faster it's very suitable for this problem but for more details please refer to our paper, which are just here [https://arxiv.org/abs/2106.01570]. And as this example illustrates the concept, our method can calculate the embeddings of a specific node, in this case Wuhan, across different years. And we can expect a huge embedding movement from 2019 to 2020.

Slide 6

Okay, so next - so let's see the experiment result. First, we collect the English Wikipedia graph snapshot every day in 2020. And as I mentioned before is a very large dynamic graph. You can see in this table roughly we got 30,000 new edges inserted every day and we released the data in this GitHub repository

so it's very easy to access. And we keep track of Wuhan together with other hundreds of Chinese cities. And this figure shows how the changes we detected in the embedding space so as you can see, the curve of Wuhan is prominent. They have a huge peak and we highlight several peaks with annotation and found that they are all correlated with, correlated to the timeline of COVID. And there's another peak - you can see we spot one other peak in the city of Chengdu which reflects the U.S.-China diplomatic tension when the U.S. decided to close the consulate in Chengdu. So it has the serendipity we discovered in this study.

Slide 7

So, oops, so, uh - we have another experiment where we want to see the most changed cities in different time periods. So we rank the embedding movements of the tracked node at each time and found that, you can see, Wuhan usually is usually the top changed one among others as COVID evolves. And we highlight the news title in that time period so we can have an idea of what happened in that place.

Slide 8

Okay. So the major takeaways are: we have a very efficient node embedding algorithm that can capture the embeddings in very large dynamic graphs. Then we investigate the evolution of Wikipedia knowledge graph in 2020 and found the interesting node changes during COVID. For more detail, please refer to our paper and see the released resources. [https://arxiv.org/abs/2106.01570 and https://github.com/zjlxgxz/DynamicPPE] And yes, thank you, thanks!