Inspiratron.org - Natural language processing, machine learning and cybersecurity

Problems with knowledge graphs and perceptions about them

by Nikola Miloševi? - Wednesday, January 04, 2023

https://inspiratron.org/blog/2023/01/04/problems-with-knowledge-graphs-and-perceptions-about-them/

Knowledge graphs have for some time traction in both research and academia. Most of the traction started in the 1990s with semantic web, which was one iteration of Web 3.0 (before blockchain), but then it expanded with the new graph databases. Since Google in 2012 introduced knowledge graph in industry settings, it started penetrating industry. However, while in some of the tech companies people know benefits and limitations of knowledge graphs, in many other industries, that won't be the case. I have been tasked to work on a knowledge graph, and with that work, we have published a paper in Journal of Web Semantics [Miloševi?, Nikola, and Wolfgang Thielemann. "Comparison of biomedical relationship extraction methods and models for knowledge graph creation." *Journal of Web Semantics* 75 (2023): 100756.] (journal paper: https://www.sciencedirect.com/science/article/abs/pii/S1570826822000403?via%3Dihub, author preprint where you can access it for free https://arxiv.org/abs/2201.01647).

Management that do not understand fully the concept, usually think they would get all the knowledge from graph with single click, and that graph has all the information that they need. However, there are many issues to be solved first, before we can talk about graph replacing information retrieval engines. With the rise of chat agents, like GPTChat and Google Lamda, it is even questionable whether do we need knowledge graph (even thought, it can be part of for example Google Lamda's toolset), or we just leave knowledge representation to be created by neural network.

However, when it comes to issues with knowledge graph, I think, they are general issues with knowledge representation. Therefore, it may make sense to note them and they should be research directions (some of them may already be). Let me focus on few main issues here:

Ontology vs probability network

This boils down to usage of big data vs crafting knowledge almost manually. One usual way of generating knowledge graph is to use natural language processing and find entities of interest and relationships between them. However, in this case, often you will get thousands of relationships between the same pair of entities, especially if dealing with large collection of text (e.g. whole scientific literature). Also, you will get relationships of different type/class between same entity. NLP will naturally add some noise, because you can't make 100% accurate algorithm. So what you are ending up with is probability network, where there are probabilities that entities are related in some way based on number of edges between them of a particular relationship type. We will talk a bit later, how this can create issues, especially with emerging relationship types. Other option is to craft ontology-like network, where you have only one edge between nodes, and you are sure it is the right one. You can do it in multiple ways, you can start from NLP-generated network and review it, leaving only nodes that you deem right, or you can do whole process manually. However, you do it, you will lose a lot of information.

Conflicting statements

While probability network will retain all conflicting statements in the graph, the question once you have them is what to do with them. If you try reasoning around them, they can lead to confusion. You can ignore them, or let user handle them. I am not saying there is no way around them, but we still probably do not know what is the best way to deal with conflicting statements in graph. You can treat network as pure probability network, or assume that the right relationship is the one that has top number of edges, but this can be in some cases wrong. For example, there is new discovery conflicting previous belief, but past belief was there for some time, and therefore has many more edges, then it would for new discovery. With this, we go to the next one.

Validity of the statement

Human knowledge change over time, and if we process historical data, such as abstracts or other kind of literature, we will find statements that are no more true or valid. After some time, in scientific community may be created consensus and new kind of relationship may get more mentioned, but what are we doing with relationships when that increase in mentions of the right relationship still did not happen? Do we take into account impact factor or some information about source of document. This may be valid, but as well a lot of important discoveries were published in journals or other sources that were not so impactful. With new or rare statements we do not know whether it is some wild imagination of the authors or new discovery, at least not until we get more information (e.g. reproducibility studies, or something like that) So it is quite hard to know what is the latest, most accurate knowledge we have on a certain semantic relationship.

Reasoning and connectibility in graph

When we have graph, generated from large amount of literature, for example about biological entities, we may reach the problem of how many hops we want to look down the line and what will it tell us. If we go too few hops, we will get really obvious connections. If we go few hops, we can obtain new knowledge, however, we are running into some risks. One of such risk is that after few hops in graph, almost everything is connected to everything and usefulness of such information is questionable. Some graph learning may help us put some threshold on such reasoning, but that may be as well uncertain.

I would be curious if someone noticed some additional limitation of the knowledge graphs. If you did, please feel free to write down in the comments and I would be happy to edit this article and add them.