Extracting information for teaching in knowledge graphs

Project for COMP8880 semester 1 2024
all code found at `https://github.com/BaconBreaker/COMP8880Project`

Thomas Brun Lau Christensen (u7865037)

Abstract

Considering the graph of all concepts, and how they are connected by how they can be used to explain each other, we consider the network of topics in tertiary level computer science and how it can be used to guide the learning path of students. We consider 2 different ways of accessing this graph through a surrogate, and show that using knowledge bases like Wikidata[3] shows the most promise, but that further work will still be required before we can derive useful insights into questions about how we learn and teach computer science at a tertiary level.

1 Introduction

When learning new concepts, one usually builds their understanding of said topic using already known concepts. E.g. using addition to understand multiplication or using mathematical relations to understand functions. This creates an interesting structure if we view each topic as a node and every time one concept can be used to explain the other, we connect an edge between them, we dub this as the concept graph.

This project will attempt to analyse the feasibility of working with this unobserved graph through an existing observed surrogate graph. We will limit the scope of the graph from the entire universe known knowledge, to computer science taught at a tertiary level.

We will finally explore the possibility of using the concept network, though its surrogate, to asses whether it could provide insights into how we structure learning a whole academic field. Trying to answer the questions of whether we can identify the most central topics that makes learning all topics in the field easier, and if we can identify the most fundamental topics of which everything else is derived.

2 Previous work

To the best of the authors knowledge, there is only piece of previous work that attempts to work on this problem. Wikidata for education (WD4E)[5] is an open-source project that aims to categorize high school curricula for all high school education in Ghana, with the aim to expand it to other countries and levels of education in the future. The WD4E project differs from this in that the goal of this project is to ascertain curricula or useful academic teaching insights in already existing networks, while the W4DE project aims to categorize a new network for the specific purpose of comparing curricula between multiple teaching institutions.

3 Approach

The method of which we will construct a surrogate graph to the concept graph of interest, is to find closely related real-world graphs and asses their suitability through exploratory data analysis. To this end we first consider two datasets and then we consider the data analysis methods used in the assessment.

3.1 Data

This projects attempts to use two different types of networks as surrogate for the concept network. We consider a hyperlink graph over a relevant online encyclopedia and a carefully chosen subset of a relevant knowledge graph.

For the hyperlink graph we consider the Free Online Dictionary Of Computing (FOLDOC)[1] website, which is an online dictionary containing short descriptions for a broad array of terms relevant in the computer science community. Each page will link to other pages pertaining to concepts that are relevant, thus creating a graph that explains which concepts are relevant to each other. The goal of using this network is to asses whether publicly sourced inter-connected encyclopedias can be used as a reliable surrogate to the concept graph we wish to study in this project. We used a publicly available scraping of the FOLDOC's hyperlink network[2] for this project.

The other network is constructed using a subset of a relevant publicly available knowledge graph. For this purpose we use the WikiData dataset[3], which is a publicly sourced knowledge whose variety of subjects is as wide as Wikipedia's. The nodes are connected by edges labelled as properties like 'father of' or 'published in', meaning that searching over specific relations is made possible in this setup. The goal of using the WikiData network is to asses whether a publicly sourced knowledge base can be used to construct suitable surrogate graphs to the concept graph this project entails.

For the WikiData graph a substantial amount of time was spent learning the structure of the graph, i.e. finding entities and properties suitable for the task, as well as learning sparQL[4], the database query language that the WikiData project uses in their online query service. To finally construct the dataset we

first define a core set of topics, then we consider a specific property, e.g. 'used by', and map all paths to any node in the core set. We repeat this for a set of manually chosen properties that fit our task, and the final dataset is the combination of all nodes and edges on these paths. To construct the core set, we consider the property type 'subclass of', and map all nodes that have a path to the subject 'Computer science' using the 'subclass of' property. The set of relations we then use to map the network are as follows

- 'Subclass of' (P279) which relates objects that are sub-themes in a large topics, e.g. 'artificial intelligence' is a subclass of 'computer science'
- 'part of' (P361) which relates concepts that are only part of a larger concept, e.g. 'loss function' is part of 'artificial neural network'
- 'Facet of' (P1269) which relates concepts that offers broader perspectives on others, e.g. 'artificial intelligence' is a facet of 'machine learning'
- 'used by' (P1535) which relates concepts utilized on others, e.g. 'activation function' is used by 'artificial neural network'

Notably there are some properties that we *didn't use* which are the following.

- 'Instance of' (P31) which relates specific elements of a concept, which proved to be way too specific and expanded the graph to a lot of unwanted areas of the total knowledge graph.
- 'Topics main category' (P910) which relates topics when they have the same main category. The main categories proved to be too specific for this relation to give any useful edges, and was thus not considered.

3.2 Exploratory data analysis method

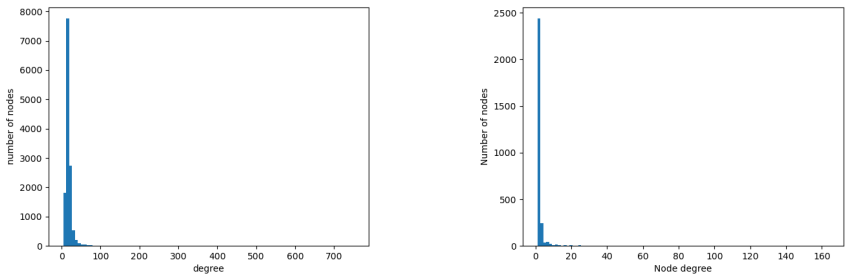
To asses the suitability of the two networks, we explore the networks using known metrics and visualization methods to explore the properties of the different graphs.

For each graph, we will note the size of the graph and the distribution of node degrees. Then we explore the most central nodes in the system using closeness centrality, betweenness centrality, and eigenvector centrality. We will compare the different resulting 'most central' nodes and discuss which method of measuring centrality makes the most sense in this setup. We will then use the information already described to visualize the whole network in a suitable manner.

4 Results

4.1 FOLDOC hyperlink network

The FOLDOC hyperlink network consists of $n = 13355$ nodes and $m = 120235$ edges, which means the average degree is ≈ 9. The node histogram seen in figure



(a) FOLDOC (b) Wikidata

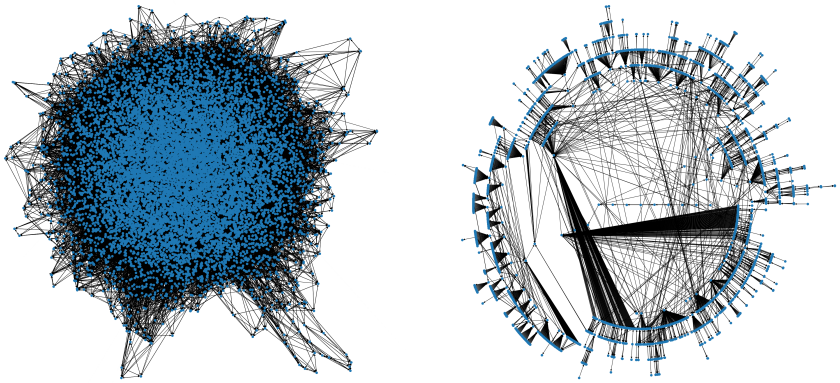
Figure 1: Histograms of node degrees of the FOLDOC and Wikidata datasets

FOLDOC		
Closeness Centrality	Betweenness Centrality	Eigenvector Centrality
'Unix', 0.3867	'operating system', 0.0598	'Unix', 0.3821
'Usenet', 0.3674	'Unix', 0.0559	'operating system', 0.2985
'operating system', 0.3528	'ASCII', 0.0554	'Usenet', 0.2182
'C', 0.3517	'C', 0.0347	'IBM', 0.1744
'Internet', 0.3441	'Internet', 0.0283	'C', 0.1682
'IBM', 0.3396	'Usenet', 0.0260	'Internet', 0.1498
'MS-DOS', 0.3291	'WEB', 0.0248	'MS-DOS', 0.1405
'protocol', 0.3286	'IBM PC', 0.0195	'IBM PC', 0.1355
'standard', 0.3247	'MS-DOS', 0.0184	'protocol', 0.1354
'IBM PC', 0.3222	'chat', 0.0164	'microprocessor', 0.1267

Table 1: Top 10 most central nodes using different centrality measures for the FOLDOC dataset.

Wikidata		
Closeness Centrality	Betweenness Centrality	Eigenvector Centrality
'computer science', 0.2478	'Q114737626', 0.0003	'computer science', 0.8238
'artificial intelligence', 0.0946	'artificial intelligence', 0.0002	'systems engineering', 0.3980
'Q114737626', 0.0803	'graph theory', 0.0002	'software development', 0.3979
'cognitive science', 0.0668	'machine learning', 0.0001	'software engineering', 0.0446
'mathematics', 0.0665	'algorithm', 0.0001	'computer programming', 0.0446
'emerging technology', 0.0660	'computability theory', 0.0001	'mathematics', 0.0123
'reasoning', 0.0660	'add-on', 8.994e-05	'information science', 0.0122
'graph theory', 0.0649	'artificial neural network', 8.6163e-05	'fundamental science', 0.0122
'theory of simplicial complexes', 0.0566	'Q114715608', 8.0800e-05	'component-based software engineering', 0.0044
'machine learning', 0.0547	'cryptographic primitive', 7.8993e-05	'theoretical computer science', 0.0010

Table 2: Top 10 most central nodes using different centrality measures for the Wikidata dataset.



(a) FOLDOC (b) Wikidata

Figure 2: Visualizations of the FOLDOC and Wikidata networks. FOLDOC is visualized using a spring layout while Wikidata is visualized as a tree-like structure.

1a shows that the degree distribution is heavily centered around 15 with little variation and a very small tail. This means that although there exists some hubs, there are very few of them. In table 1 we see the most central nodes for the 3 different centrality measures. We see a high degree of agreement in the top 10, with 'Unix', 'Usenet', 'Operating system', 'C', 'IBM', and 'MS-DOS' being terms that are central in the graph, and thus central to understanding the most topics in the network. Finally, using that the graph has a very interconnected structure, we plot the network using the spring layout as seen in figure 2a, where it is again confirmed that the graph behaves like we would expect a hyperlink network to.

4.2 Wikidata knowledge graph

The manually chosen Wikidata sub-network consists of $n = 2866$ nodes and $m = 3281$ edges, which implicates that the graph is close to having a tree structure, since the average node degree is ≈ 1.144. This is again supported by the histogram of node degrees seen in figure 1b where we see a very sharp concentration of nodes having only 1 or 2 connections, indicating a strong 'branch-like' structure in the network. Figure 1b also shows that there exists hubs in the network, although there comparatively are few of them. Looking at table 2 we see that the agreement between the 3 centrality measures is less than in the FOLDOC network. There are nevertheless some agreement on the topics 'Computer Science', 'Artificial intelligence', and 'Mathematics'. Using that the network is close to a tree structure, we plot it as such in figure 2b where we

again confirm its branching structure.

5 Exploring the concept network

```
supervised learning -> support vector machine
classification algorithm -> support vector machine
machine learning -> supervised learning
    -> support vector machine
artificial intelligence -> machine learning
    -> supervised learning -> support vector machine
mathematics -> machine learning -> supervised learning
    -> support vector machine
computer science -> artificial intelligence -> machine learning
    -> supervised learning -> support vector machine
emerging technology -> artificial intelligence -> machine learning
    -> supervised learning -> support vector machine
reasoning -> artificial intelligence -> machine learning
    -> supervised learning -> support vector machine
cognitive science -> artificial intelligence -> machine learning
    -> supervised learning -> support vector machine
```

Figure 3: paths in Wikidata network that leads to 'Support vector machine'.

Judging on the analyses done en sections 4.1 and 4.2 we conclude that the FOLDOC network is not suited for the task ahead due to its over-connected nature and focus on overly specific terms, while the Wikidata is sparsely connected with a focus on general and broad topics.

Utilizing the Wikidata network we will now do a short investigation into how this graph can be used to guide learning paths. In figure 3 we see the different paths leading to the node 'support vector machine', which in turn should tell us what concepts SVM's build upon and we thus need to learn beforehand. We see that a few sensible paths exists with all of them traversing the nodes 'Machine learning', and 'supervised learning', which are the neighborhood of the SVM node.

6 Discussion

The FOLDOC network did not seem suitable as a surrogate to the concept network, as discussed in section 5. The Wikidata network did show more promise, but with its own pitfalls. The Wikidata dataset consists of very broad topics, that encapsulate a lot of knowledge at once, whilst being considerably smaller than the FOLDOC dataset, and also smaller than we would ideally like it to

be. This has the consequence that simpler and smaller topics are never explicitly modeled in the network, causing us to never see topics we would expect like 'addition', or 'functions', although this may be acceptable for tertiary level education purposes.

Mining the concept graph from already existing networks makes it improbable to find graph to out specifications. The more specific we are about the properties and topics modelled in the network, the less probable it would be to be able to find such an already mapped network. This creates an inherent disadvantage to using surrogate graphs for the analysis of the connections in concepts that is unavoidable. Thus fine-grained analysis is only feasible in graphs created for the specific purpose. This means that the optimal approach to the problem is found using graph created like done in the Wikidata for education project[5], but for different academic levels and areas.

7 Conclusion and further work

We have looked at two different datasets as candidates for a surrogate of the concept network, of which one is a hyperlink network of a relevant online dictionary and the other is a manually chosen subset of a large knowledge graph. We find that using a hyperlink network introduces a lot of unwanted edges, as the semantics behind a page linking to another page can be too broad. We also conclude that using knowledge bases shows promise, but requires more work into finding more suitable queries as well as a more extensive array of topics than what is currently offered.

Continuing the work on the question of supporting tertiary computer science education will require the manual construction of the concept graph, where we would ideally source it from multiple sources, just like the same approach done in Wikidata for education[5], although in a different academic field and at a different academic level.

References

[1] Foldoc - free online dictionary of computing. https://foldoc.org/. Accessed: 2024-05-08.

[2] Foldoc. http://vlado.fmf.uni-lj.si/pub/networks/data/dic/foldoc/foldoc.htm. Accessed: 2024-05-08.

[3] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September 2014.

[4] W3C. SPARQL 1.1 Query Language. Technical report, W3C, 2013.

[5] Wikimedia. Wikidata for education. https://www.wikidata.org/wiki/Wikidata:Wikidata_for_Education. Accessed: 2024-05-23.

All code can be found at https://github.com/BaconBreaker/COMP8880Project