Construction of Concept Graph for Education Domain

Roshankumar Pradhan[1], Kaushal Tajane[2], Akash Yadav[3], Prof. Sonali Mane[4]

[1,2,3] Final year student, Department of Information Technology, BVCOE, Navi Mumbai.
[4] Professor, Department of Information Technology, BVCOE, Navi Mumbai.

Abstract — *Inspired by vast applications of data graph and the ever increasing demand in education domain, we would propose a system, to automatically construct Concept graph for education domain. Concept graph is an integrated information repository that interlinks heterogeneous data from different domains. By leveraging on heterogeneous data (e.g., pedagogical data and learning assessment data) from the education domain, this technique first extracts concepts of subjects or courses, and so identifies the tutorial relations between concepts. More specifically, it adopts the Wikipedia web page crawling on pedagogical data to extract instructional concepts, and employs comparison between two chapters of the subject input data to spot the relations with educational significance. We detail all the above efforts through an exemplary case of constructing a demonstrative concept graph for a selected subject, where the educational concepts and their prerequisite relations are derived from curriculum standards and concept based performance data of scholars. It would not only be helpful in the Education sector but anywhere where learning is a task.*

Keywords- *Concept Graph, Prerequisite relation, Knowledge graph, Educational Domain, Text Mining.*

I. INTRODUCTION

Concept graph serves as an integrated information repository that interlinks heterogeneous data from different domains. Google's Knowledge Graph is such a prominent example that represents real world entities and relations through a multirelational graph. Existing generic knowledge graphs have demonstrated their advantages in supporting a large number of applications, typically including semantic search (e.g., Google's Knowledge Panel), personal assistant (e.g. Apple's Siri) and deep question answering (e.g., IBM's Watson and Wolfram Alpha). However, those generic knowledge graphs usually cannot well support many domain-specific applications, because they require deep domain information and knowledge. Education is one of such domains, and in this work, we mainly focus on how the concept graph for education can be automatically constructed. In education domain, Concept graphs are often used for subject teaching and learning in school, where they are also called concept maps.

Moreover, popular massive open online course (MOOC) platform such as Khan Academy, also adopt them for concept visualization and learning resource recommendation. Such Concept graphs are usually constructed by experienced teachers or domain experts in a manual way. However, such a manual construction process is actually time consuming and not scalable to large number of concepts and relations. What's more, the number of courses and subjects grows fast on MOOC platforms, so it is much more difficult, or even impossible, to manually construct concept graphs for each new course. Students need prior knowledge for thorough understanding of educational content. This need imparts an implicit order in learning educational concepts. Determining this order requires significant human time and effort. Furthermore, relying on expert knowledge to determine this order is subject to inconsistencies due to 'expert blind spot', which means expert's cognition and learner's cognition on the same concept often do not well align. In other words, even the domain experts or experienced teachers may easily misunderstand learners' cognitive process. As a result, those manually created Concept graphs can be suboptimal or misleading for learners.

II. RELATED WORK

Google's knowledge graph, a variety of generic knowledge graphs, such as Reverb, Google Vault, Freebase, and Microsoft's Probase, have been constructed by industry and academia, mainly utilizing data collected from the Internet. In educational domain, few studies focus on systematic construction of concept graphs[10], but there are some recent works investigating different relation extractions between certain known educational entities: Devendra et al.[1] automated induce of prerequisite structures of multiple units in a course, they propose a generic algorithm to use educational material and student activity data from heterogeneous sources to create a Prerequisite Structure Graph; Wang et al.[2] extract concepts hierarchies from the textbooks, extracts important concepts in each book chapter using Wikipedia as a resource and from this construct a concept hierarchy for that book; Liang et al.[3] recovers prerequisite relations from course dependencies, Wikipedia
data is exploited to find prerequisite relations among universally shared concepts using both the Wikipedia article contents and their linkage structures; Chen et al[4] constructed knowledge graph for academic and online courses, the main concepts are extracted using neural network and data mining technique is used to find the prerequisite relation and Yu Lu et al.[5] constructed a system that build graph for online platforms. The most relevant work to our research is carried out by Carnegie Mellon University: The researchers utilize observed relations among courses to create a directed concept graph [6], but the relations are assumed to be known in advance.

III. SYSTEM OVERVIEW

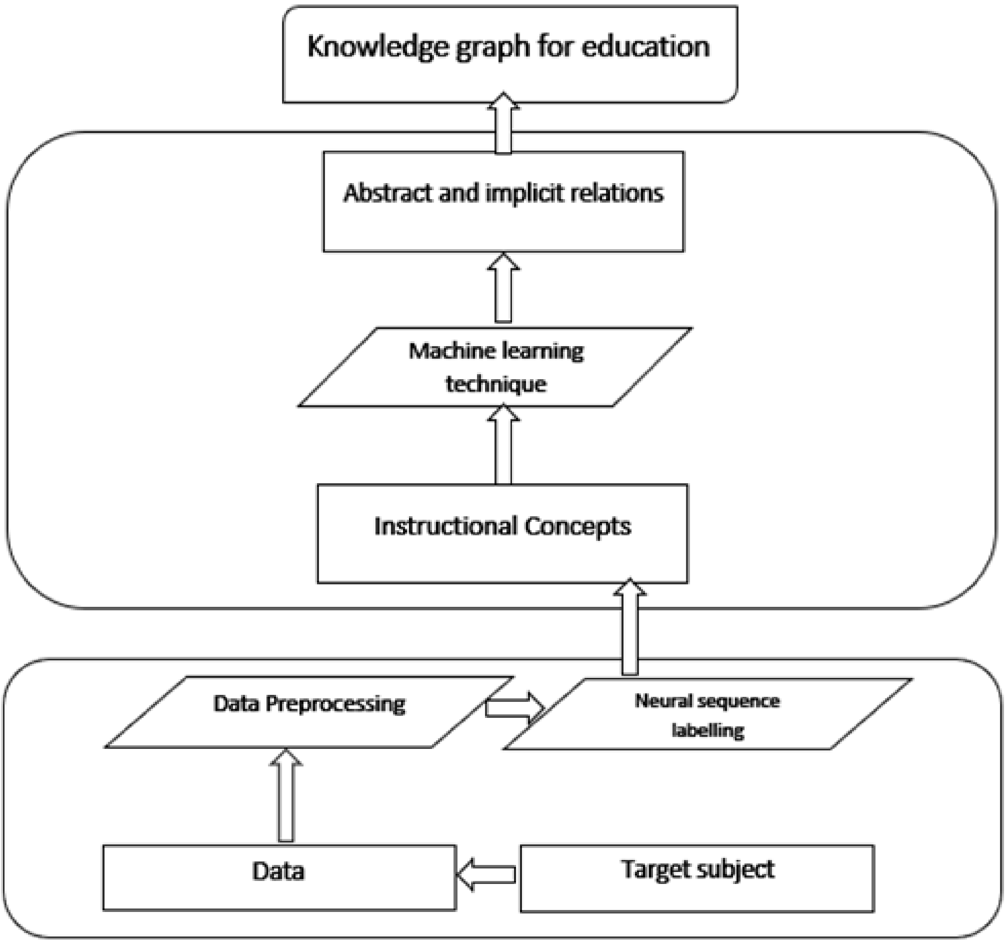

Knowledge graph for education
Abstract and implicit relations
Machine learning technique
Instructional Concepts
Data Preprocessing
Neural sequence labelling
Data
Target subject

Fig. 1: Block Diagram of the System

The block diagram of proposed system is illustrated in Figure 1. Its architecture mainly consists of three modules: Instructional Concept Extraction Module, Educational Relation Identification Module, Graph Generation. Their general descriptions of the modules are given as follows:

A. Instructional Concept Extraction Module:

The main objective of this module is to extract instructional concept for a given course and subject. This module takes the pedagogical data, that including the curriculum standards, textbooks and courses tutorials, which are collected from the education domain. The documents are in the printed form firstly they need to be converted into machine-readable text format. After the preprocessing data selection and format conversion natural language processing, name entity recognition especially neural sequence labelling can be deployed to extract the instructional concepts which is the main part of the system. The instructional concepts are the concept that is need to be mastered by the learner. This concept are the nodes to our Concept graph.

B. Educational Relation identification module:

The main objective of this module is to identify the educational relation that interlinks instructional concept that is extracted in the instructional concept extraction module to help the learning process. Since the educational relation are implicit and abstract, the prerequisite relation between the two chapters are identified in this module. Finally those identified relations connect instructional concept and construct the concept graph for the education that can be used by the teachers and the learners.

C. Graph generation Module:

The graph is generated using the graphviz. Graphviz is open source graph visualization software. Graph visualization is a way of representing structural information as diagrams of abstract graphs and networks. It has important applications in networking, bioinformatics, software engineering, database and web design, machine learning, and in visual interfaces for other technical domains. In this phase the concept graph is constructed using the important concept extracted in instructional concept extraction. The root node is the target subject and the important concept are the child nodes. In the following two sections, we will elaborate our design for these three modules respectively.

IV. INSTRUCTIONAL CONCEPT EXTRACTION

A. Data source and Preprocessing:

The desired node in our educational concept graphs represent instructional concept that are the cornerstone and to be mastered by the learner. The input data is mainly collected from the education domain and pedagogical practices, such as textbook, curriculum standards and Table of content of courses. The input data is the printed text. Thus, conversion into machine readable format is required and various format conversion technique can be used. For the conversion, the optical character recognition (OCR) [7] technique is used to handle the printed document. The tesseract OCR can be used to extract the text and convert it into machine readable format. After this pre-processing step of transforming input data into machine-readable text, the proposed system can then perform the instructional concept extraction.

B. Instructional Concept Extraction:

The input to natural language processing will be a simple stream of Unicode characters (typically UTF-8). Basic processing will be required to convert this character stream into a sequence of lexical items (words, phrases, and syntactic markers) which can then be used to better understand the content. Tokenization to divide up character streams into tokens which can be used for further processing and understanding.[8] Tokens can be words, numbers, identifiers or punctuation. Entity extraction – identifying and extracting entities (people, places, companies, etc.) is a necessary step to simplify downstream processing. The instructional concept are concepts that need to be mastered by the learner. For the extraction of the instructional concept from the target subject the beautifulsoup is used for pulling data from the Wikipedia pages. It converts incoming documents to Unicode and outgoing documents to UTF-8. The extracted words from the chapter is compared with the Wikipedia data. The Word2vec algorithm is used for the task of finding the important concept from the target subject and the Wikipedia page crawl data.

V. EDUCATIONAL RELATION IDENTIFICATION

A. Prerequisite relation identification:

Prerequisite relation is in accordance with the knowledge space theory, which argues that prerequisite exists as a natural dependency between concepts in human cognitive process. Specifically, a prerequisite relation from concept A to concept B means that a learner should master concept A first before proceeding to concept B.[9] To find the prerequisite between two chapter simple technique of comparison is used the word of the first chapter is compared with the words of the second chapter extracted if the word are present this are the prerequisite. The stop word dictionary is maintained all the irrelevant words are ignored for the comparison.

VI. GRAPH GENERATION

A. Graph Result

Last step is to generate graph of the important concept extracted in the instructional concept extraction phase the graph nodes contains the important concept that the learner needs to be mastered. The Graphviz is open *source* graph visualization software. Graph visualization is a way of representing structural information as diagrams of abstract graphs and networks. The root node is the target subject and the child nodes are the important concept extracted in the instructional concept extraction. The words extracted from the instructional concept are the nodes input.

VII. OUTCOME

To evaluate the proposed system, we construct an exemplary graph for Enterprise Network Design, which demonstrates the instructional concept extraction and the prerequisites and the concept graph for that chapter. Two chapters are given as input to the system and the instructional concept, prerequisite and the graph is constructed.

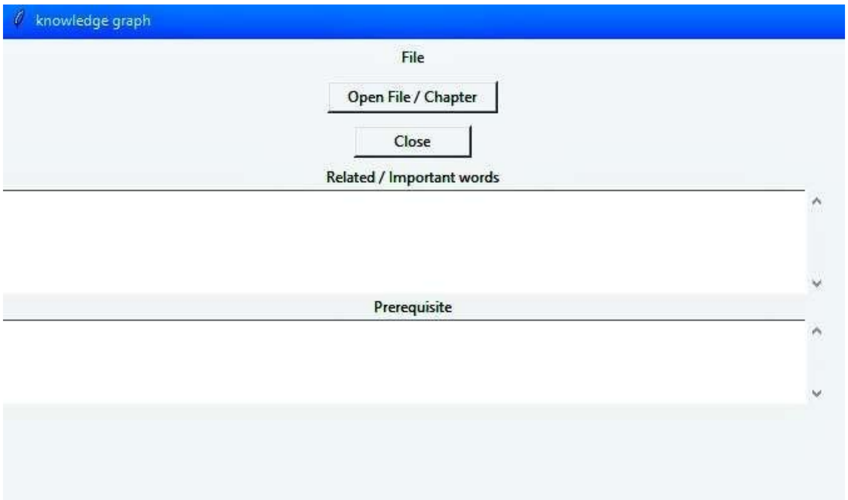
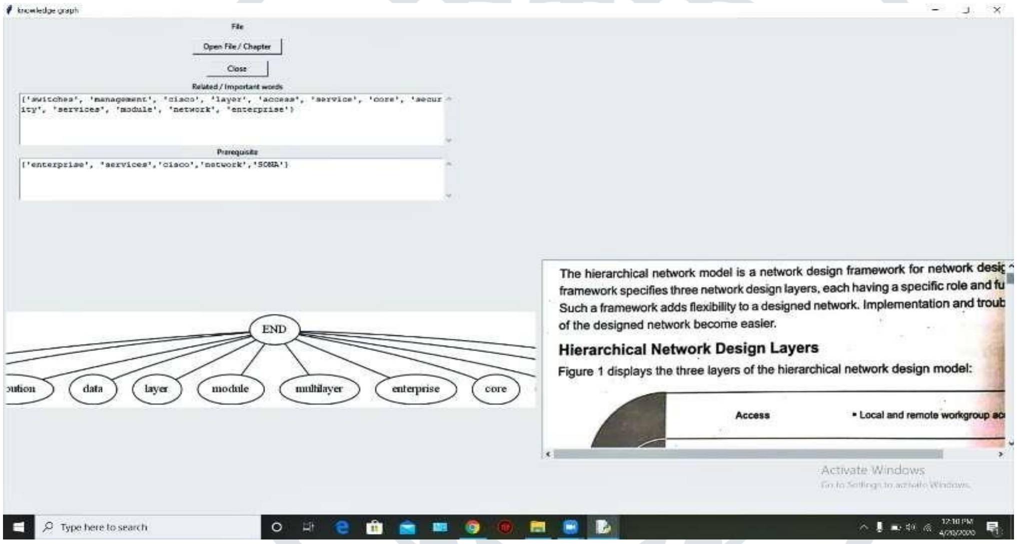


Fig .2: Startup GUI

This is the startup GUI the user needs to select the chapter for which he/she need the prerequisite and the concept graph. After the user selects the chapter the system will the pdf into machine readable format and the further process will be done and the output displayed in the particular output window. Here we have selected the chapters of END it consists of two chapters.



```
{'switches', 'management', 'cisco', 'layer', 'access', 'service', 'core', 'secur
ity', 'services', 'module', 'network', 'enterprise'}
```

```
{'enterprise', 'services','cisco','network','SONA'}
```

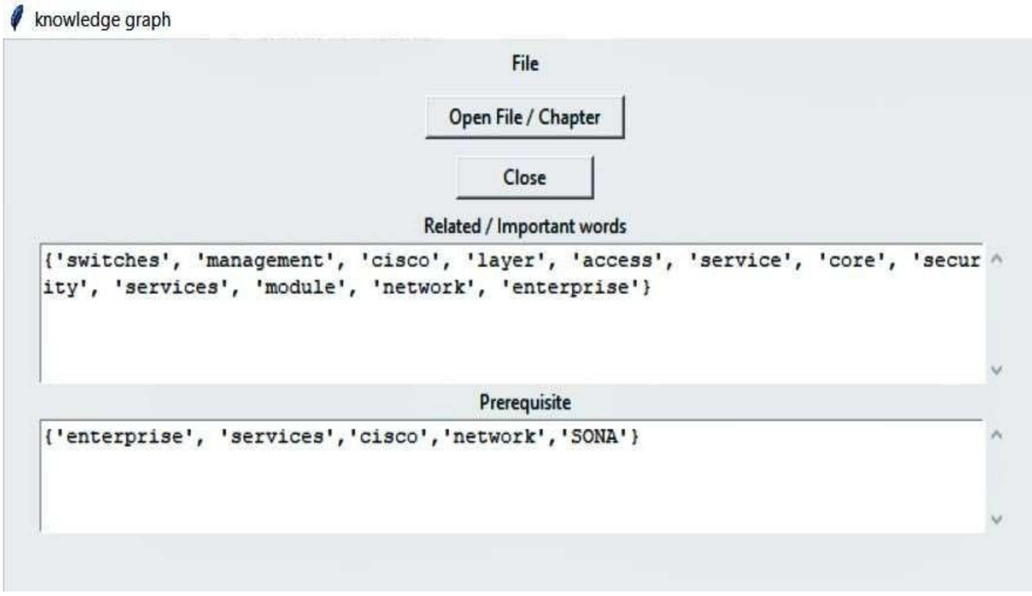
The hierarchical network model is a network design framework for network desig
framework specifies three network design layers, each having a specific role and fu
Such a framework adds flexibility to a designed network. Implementation and troub
of the designed network become easier.

Hierarchical Network Design Layers

Figure 1 displays the three layers of the hierarchical network design model:

Access • Local and remote workgroup ac

Fig.3: Output of the system



```
{'switches', 'management', 'cisco', 'layer', 'access', 'service', 'core', 'secur
ity', 'services', 'module', 'network', 'enterprise'}
```

```
{'enterprise', 'services','cisco','network','SONA'}
```

Fig.4: Instructional concepts and Prerequisite extracted

The 1st output window shows the words related to the concept. The 2nd output window shows the prerequisite words of the chapter and the graph and the pdf entered is shown in the bottom of the gui. Here it extracts the thirteen important concepts that is to be mastered by the learner and the prerequisite of the chapters.

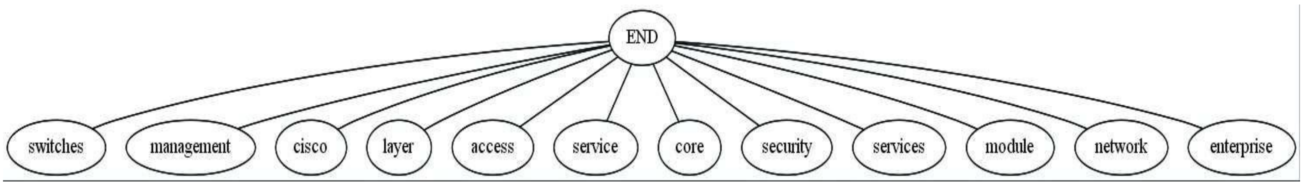


Fig.5: Concept Graph for One Chapter

This represent the final concept graph generated for the chapter END at the root node and the important concept the user needs to be mastered for the studying the chapter. The user gets the snapshot of the whole chapter from this graph. The is generated automatically at the end. Here we have constructed the graph for the subject enterprise network design.

VIII. RESULT ANALYSIS

Sr No.	Chapters	Automatically Extracted from system	Manually Extracted	Accuracy (%)
1.	Chapter 1	12	14	85
2.	Chapter 2	6	9	66
3.	Chapter 3	9	13	70
4.	Chapter 4	6	10	60
5.	Chapter 5	7	10	70
Average Accuracy			70%	

Table No.1: Accuracy of the system

To find the accuracy of the system, we constructed the concept graph for five chapters of enterprise network design. For the concepts are extracted manually by the subject expert and matched with the concepts extracted automatically using the system. The table shows the accuracy of the system for different chapters. The average accuracy of the system is 70% as of now.

IX. CONCLUSION

We have implemented the Concept Graph for Education domain, which automatically constructs a graph consisting of word (concepts) related to the chapter for education domain. The system extracts instructional concepts and educational relations from heterogenous data sources, mainly including standard curriculum data and learning assessment data. For the instructional concept extraction, Wikipedia web pages crawling and word2vec models was employed, and for the prerequisite relation identification, the comparison between the occurrence of the concepts between two chapters was introduced. Furthermore, the precedence relationship between concepts was largely relevant and accurate. Further experiments, however, are needed. Firstly, the technique must be applied to a larger network of concepts and applied to a larger number of students thus the user models will be more varied. On a broader canvas, this system is feasible and have effectiveness to construct dedicated concept graph for the education domain. Such system would not go unnoticed among the growing youth of the world.

X. REFERENCES

[1]. Chaplot and K. R. Koedinger, "Data-driven automated induction of prerequisite structure graphs," in *Proceedings of the Educational Data Mining (EDM)*, 2016

[2]. S. Wang, C. Liang, Z. Wu, K. Williams, B. Pursel, B. Brautigam,S. Saul, H. Williams, K. Bowen, and C. L. Giles, "Concept hierarchy extraction from textbooks," in *Proceedings of the 2015 ACM Symposium on Document Engineering*. ACM, 2015, pp. 147–156

[3]. C. Liang, J. Ye, Z. Wu, B. Pursel, and C. L. Giles, “Recovering concept prerequisite relations from university course dependencies,” in *AAAI Conference on Artificial Intelligence*, 2017

[4]. Penghe Chen, Yu Lu*, Vincent W. Zheng, Xiyang Chen, Boda Yang “KnowEdu: A System to Construct Knowledge Graph for Education”

[5]. Penghe Chen, Yu Lu, Vincent W. Zheng, Xiyang Chen, Xiaoqing Li “An Automatic Knowledge Graph Construction System for K-12 Education”

[6]. H. Liu, W. Ma, Y. Yang, and J. Carbonell, “Learning concept graphs from online educational data,” *Journal of Artificial Intelligence Research*, vol. 55, pp. 1059–1090, 2016.

[7]. S. M. Beitzel, E. C. Jensen, and D. A. Grossman, “Retrieving ocr text: A survey of current approaches,” in *Symposium on Document Image Understanding Technologies, SDUIT*, 2003.

[8]. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,”*arXivpreprint arXiv:1412.3555*, 2014.

[9]. J.-P. Doignon and J.-C. Falmagne, “Spaces for the assessment of knowledge,” *International journal of man-machine studies*, vol. 23, no. 2, pp. 175–196, 1985.

[10].H. Liu, W. Ma, Y. Yang, and J. Carbonell. Learning Concept Graphs from Online Educational Data. In *Journal of Artificial Intelligence Research*, 55: 1059- 1090, 2016.