EduVis: Visualization for Education Knowledge Graph Based on Web Data

Kai Sun
School of Computer Science and Software Engineering
East China Normal University
Shanghai, P.R.China
+86-15221559007
ksun@live.cn

Yuhua Liu
School of Computer Science and Software Engineering
East China Normal University
Shanghai, P.R.China
+86-13585526596
yhliu216@gmail.com

Zongchao Guo
School of Computer Science and Software Engineering
East China Normal University
Shanghai, P.R.China
+86-18516790965
gzongchao1@163.com

Changbo Wang *
School of Computer Science and Software Engineering
East China Normal University
Shanghai, P.R.China
+86-18502160116
cbwangcg@gmail.com

ABSTRACT
How to clearly present the internal structure of knowledge graph is particularly important, however, the current visualization researches about it are rare. We construct education knowledge graph utilizing extracted entities and entity relations, and construct a visual analysis platform, EduVis. In EduVis，we design and implement a) a layout of events network based on topological structure to explore in details, b) a layout of event network based on timeline to explore time information, c) a click tracking path to record the history of users' clicks and help users backtrack.

Categories and Subject Descriptors
I.3.8 [**Applications**]: Visualization

General Terms
Algorithms

Keywords
Visualization; Education knowledge graph; Web data.

1. INTRODUCTION
Education data, which includes media coverage of educational events and related policies enacted by governments, has aroused wide attention of decision makers and researchers. The public opinions on them are reflected in the network via a variety of forms, such as micro-blog, forums, news reports, etc. It is very meaningful to integrate, process and analyze the multi-source educational data utilizing big data and visualization techniques.

It is difficult for users to analyze so various and massive data effectively. How to clearly visualize entities and relationships between them is an important research topic. Commonly used method of displaying knowledge graph is to describe the entities as nodes and the relationships between entities as lines. However, the layout of relationships between entities will produce clutter using traditional graph layout directly to show knowledge graph.

VINCI '16, September 24-26, 2016, Dallas, TX, USA
ACM 978-1-4503-4149-3/16/09.
http://dx.doi.org/10.1145/2968220.2968227

In this paper, we design and implement EduVis illustrated in Fig.1 for education knowledge graph. Our contributions include (1) extracting entities and entity relations based on natural language processing techniques and constructing education knowledge graph (2) visualizing education knowledge graph based on topological structure of events network and timeline, and presenting a click tracking path concerned by users.

2. Knowledge Graph Pre-Processing
Topic Mining. In the field of natural language processing, topic model is used to mine abstract topics in document sets. We adopt LDA (Latent Dirichlet Allocation) to process the data of educational public opinion and model for massive text corpus to find the themes hidden in it. Our data modeling and classification processing retain information related with educational public opinion while removing information unrelated to that.

Entity Extraction and Normalization. NER techniques are utilized to recognize various types of entities in documents. After obtaining a collection of unnormalized entities M recognized by NER models, we filter out incorrect or noisy entities. We design a mapping function $f : m \rightarrow e$ such that for each entity m in the remaining entity set M' ($M' \subseteq M$), it maps m to its normalized, unambiguous form e. The techniques for entity normalization including sub-string matching and entity disambiguation, are introduced in [1].

Semantic Relation Extraction. It is hard to apply traditional pattern-based relation extraction methods, because web documents related to a certain topic and in a specific domain are relatively sparse. To solve this problem, we analyze statistical and linguistic features to identify candidate relation tuples [2], in the form of (e_i, e_j, $C_{i,j}$), where e_i and e_j are normalized entities, and $C_{i,j}$ are the contexts of e_i and e_j. For the purpose of labeling the extracted candidate relations, we cluster entity pairs which have similar contexts together as a raw relation, i.e., $R = \{(e_i, e_j, C_{i,j})\}$. We label the keywords for the raw relation R by extracting the frequent keywords in $C_{i,j}$ for all e_i and e_j pairs.

3. VISUALIZATION DESIGN
Education knowledge graph is mainly composed of entities, relations between entities and other relevant information. We construct EduVis with event network view as the core view combining timeline, click tracking path and so on.

* Corresponding author: Changbo Wang

Figure 1. EduVis for education knowledge graph

Module of Event Network. Network graph is effective to make visual analysis of linked and relational data. Link information always works together with group, collection and community [3]. We construct event network graph illustrated in Fig.2 according to the features of education knowledge graph to present the relations between events better. There are four steps for constructing the layout of event network graph based on topological structure.

1. We adopt force-directed layout algorithm [4] to position the nodes of events. Then we can get the positions of event nodes which are used to compute the locations of other entities.

2. We adopt ring layout method to position the nodes that belong to only one event randomly at the ambient area of the corresponding event node.

3. The entity which belongs to multiple events is located at the center of the multiple event nodes. But the positon may already be occupied by a node. In order to avoid covering the existing nodes, the method of spiral scanning is used for collision detection. The principle of the method is Archimedes Spiral, a trail generated by one point leaves the fixed position with mean speed and rotates around a fixed point with fixed angular speed. The nodes belonging to the events are considered a collection which is located according to step 2.

4. After ascertaining the positions of all the nodes, we should draw the lines according to the relations between them. The weight of each line is set as the times it appears, and the higher value means the more important relationship between nodes.

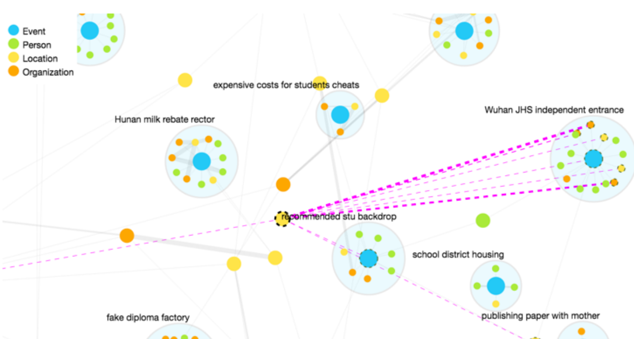

Figure 2. Event network graph based on topology

Module of Event Network Based on Timeline. Utilizing event network based on topological structure, we can get a better layout to mine involved entities and the relationships between different event communities. Nevertheless, we can't acquire time information in it. So we position event communities on the time axis according to the happened time of them. According to the value of the weights, we arranged them in turn. When overlapping occurs, we adopt the method of moving slowly up to avoid it. The strategy of positioning other entities and relations is same as module of event network. Event network graph based on timeline [5] illustrated in Fig.3 generated by the same data as Fig.2. Compared Fig.3 with Fig.2, we have no trouble in finding the correlation between distinct communities, the happening times of the events and the regularity of events in the time distribution by switching the two module arbitrarily.

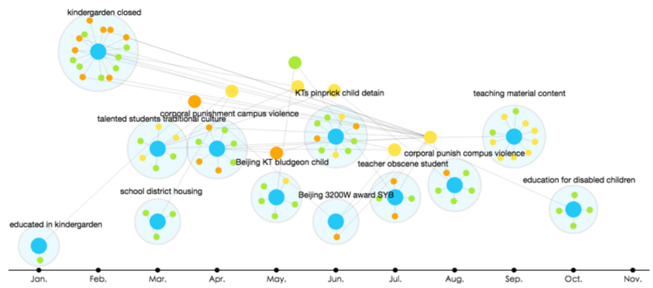

Figure 3. Event network based on timeline

Module of Click Tracking Graph. Inspired by [6], we offer users click tracking graph at the bottom of Fig.1 to backtrack when they want to repeat and contrast the analysis. In event network graph, users can click the concerned nodes to get the detailed information, meanwhile the clicked entity will be added to the record of click tracking graph. If the node is not isolated from the last clicked node, we draw a solid line between them, otherwise the line is dotted. The newly added nodes are connected to the nearest nodes that belong to the same event.

4. ACKNOWLEDGEMENT

This paper was supported in part by Natural Science Foundation of China under Grants 61272199 and the Specialized Research Fund for the Doctoral Program of Higher Education under Grants 20130076110008.

5. REFERENCES

[1] Jijkoun, V., Khalid, M. A., Marx, M., & De Rijke, M. 2008. Named entity normalization in user generated content. *The Workshop on Analytics for Noisy Unstructured Text Data*. (New York, USA, 2008) ACM, 23-30.

[2] Shen, W., Wang, J., Luo, P., Wang, M and Yao, C. 2011. REACTOR: a framework for semantic relation extraction and tagging over enterprise data. In *Proceedings of the 20th international conference companion on World Wide Web*. (New York, USA, 2011) ACM, 121-122.

[3] Jianu, R., Rusu, A., Hu, Y., & Taggart, D. 2014. How to display group information on node-link diagrams: an evaluation. *IEEE Transactions on Visualization & Computer Graphics*, 20(11), 1530-1541.

[4] Walshaw, C. 2000. A multilevel algorithm for force-directed graph drawing. In *Proceedings of Graph Drawing*. Springer Berlin Heidelberg, 171-182.

[5] Nguyen, P. H., Xu, K., Walker, R., and Wong, B.W. 2015. TimeSets: Timeline visualization with set relations. *Information Visualization*. (Oct 2015) DOI=10.1177/1473871615605347.

[6] Kucher, K., Schamp-Bjerede, T., Kerren, A., Paradis, C., & Sahlgren, M. (2016). Visual analysis of online social media to open up the investigation of stance phenomena. *Information Visualization, 15*(2), 93-116.