Int J Software Informatics, Volume 10, Issue 3 (2016), pp. 000–000
International Journal of Software and Informatics, ISSN 1673-7288

E-mail: ijsi@iscas.ac.cn
http://www.ijsi.org
Tel: +86-10-62661048

DOI: 10.21655/ijsi.1673-7288.00227

Visualization for Knowledge Graph Based on Education Data

Kai Sun, Yuhua Liu, Zongchao Guo and Changbo Wang

(School of Computer Science and Software Engineering, East China Normal University, Shanghai, P.R.China)

Abstract Knowledge graph, also known as scientific knowledge graph, can reveal the dynamic development rules in complex knowledge fields. How to clearly present the internal structure of knowledge graph is particularly important, however, the current visualization research based on knowledge graph is rare. In this paper, varieties of data related to education are mined from massive web data, and are fused together. Then knowledge graph which is centered on educational events is constructed utilizing extracted named entities and entity relations. We construct a visual analysis platform for education knowledge graph, EduVis, which can support users to do associated analysis of education, and enable users to obtain the public opinions. In EduVis, we design and implement a) a word cloud treemap to provide an overview of education knowledge graph, b) a layout of events relation network graph based on topological structure and timeline to explore in details, c) a click tracking path to record the history of users' clicks and help users to backtrack. The case studies show that the aforementioned visual analysis methods for our knowledge graph can meet users' demands for data analysis tasks.
Key words: visualization; education knowledge graph; web data

Sun K, Liu YH, Guo ZC, Wang CB. Visualization for knowledge graph based on education data. *Int J Software Informatics*, **Vol.10, No.3 (2016): 000–000. http://www.ijsi.org/1673-7288/10/227.htm**

1 Introduction

Nowadays, the number and variety of data are increasing rapidly with the arrival of the era of big data. Education data, which includes media coverage of educational events and related policies enacted by government at all levels, has aroused wide attention of decision makers and researchers. In the field of education, a disruptive change in information technology is quietly happening. Meanwhile, the public opinions of internet users on educational policy and educational events are reflected in the network via a variety of forms, such as micro-blog, forums, news

This paper is partially supported by Natural Science Foundation of China under Grant Nos. 61532002, 61672237; National High-tech R&D Program of China (863 Program) under Grant 2015AA016404; the Specialized Research Fund for Doctoral Program of Higher Education under Grant 20130076110008.
Corresponding author: Changbo Wang, Email: cbwangcg@gmail.com
Received 2016-08-15; Revised 2016-10-01; Accepted 2016-10-28.

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. An effective method is to use the knowledge graph technique which is widely used at present. Knowledge graph[1], whose essence is a semantic network to reveal the relationship between knowledge, is a new way to represent knowledge. It can express information effectively and infer new knowledge based on what we have accordingly. Knowledge graph based on web information such as DBpedia[2], YAGO[3], ReVerb[4], is well studied, most of which are constructed by entities and their relationships that extracted from web such as Wikipedia, Baidubaike, etc. Besides, another type of knowledge graph focuses on specific domains, such as academic citation relationship and life sciences. They can clearly demonstrate knowledge and directly solve problems in this domain. It is an important research topic that how to clearly visualize entities and relationships between them. Commonly used way of displaying knowledge graph is to describe the entity as a node and the relationship between two entities as a line, which can display characteristics of knowledge graph structure. However, the layout of the relationship between entities will produce clutter using traditional graph layout directly to show knowledge graph.

In this paper, we design EduVis, which assists educational administrators to make decision based on related data extracted from web. One of the major problems to solve is how to display these data in an organized, structured form and discover the potential patterns in data through visual analysis. We construct a knowledge graph utilizing education data and combine new visualization methods to present data characteristics and potential patterns. Our contribution includes (1) constructing knowledge graph based on education data that includes entities, entity relations extracted from web by adopting natural language processing technology, (2) combining word cloud with treemap to present the overview of knowledge graph for users, and visualizing education knowledge graph, and a visual analysis method based on topological structure of events network and timeline is designed and implemented, at the same time, a click tracking concerned by users is presented, (3) building education knowledge graph visualization platform, EduVis, which can assists users to make analysis visually.

2 Related Work

For constructing knowledge graph based on web information, except works we have mentioned, there are NELL[5], TextRunner[6], and Probase[7] that improve the way of extracting information from unstructured text data and construct knowledge graph using structured information extracted from web. These works have been applied in semantic conceptualization, semantic inference and complex query which users commit to search engine. Knowledge graph based on web information has massive data and covers a variety of knowledge domains.

Another way to construct knowledge graph is to use data of specific domain aiming at solving some specific questions. Hulliyah[8] et al. construct knowledge graph using text data, and produce summary of original text based on knowledge graph. Y Zhu[9] et al. present CKGHV that constructs knowledge graph using character relationships of Romance of the Three Kingdoms, and take advantages of

traditional visualization methods including using colors to represent different types of relationship between characters such as enemies or brothers, and using the boldness of the line to indicate the strength of relationship. By combining knowledge graph with visualization analysis technique, users can clearly and directly analyze dynamic process of history events. Yao[10] et al. organize knowledge in the field of information security and construct knowledge graph that can be applied in knowledge representation, navigation, and analysis tasks.

In order to make users have insight into knowledge graph effectively and directly, we need to take advantages of visualization technology, and there are some works focusing on it. Lai[11] et al. display web graph data using graph layout that considers a webpage as a node, link between two webpages as an edge between two nodes, and can show and navigate web graph data according to users' interests. Gibson[12] et al. and Tang[13] et al. design graph layout by attributes of graph node, and they obtain better results than traditional force-directed layout when this method is applied in data of small world network. Tang et al. adopt the same idea to build graph layout based on measuring distance between nodes. Wu[14] et al. present SAL (Subgroup Analysis Layout) algorithm that improves traditional force-directed layout by combining role analysis and key attributes, which can be applied in 2D and 3D visualization tasks.

3 Framework

The goal of EduVis for education knowledge graph is to help users to explore the the hidden information in massive educational data, and to assist education manager to make decision. Just as shown in Fig. 1, various educational data are extracted from web including the network public opinion information. Then, the raw data are processed from many aspects and various visual layout methods are presented to display them. Finally, friendly man-machine interaction ensures that users can explore the hidden information conveniently.

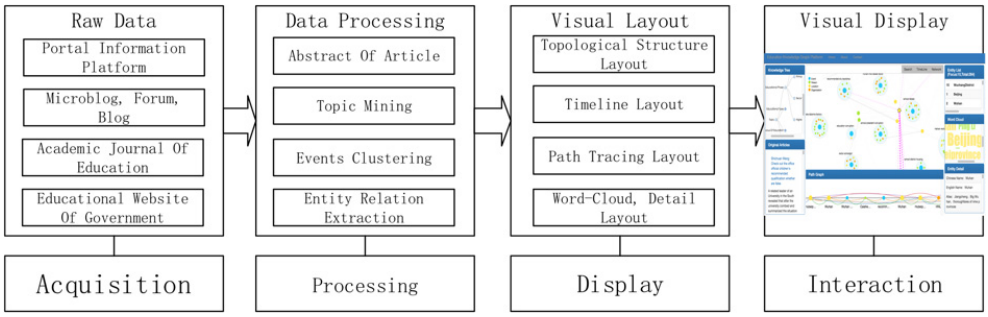


Figure 1. Constructing process of the visual analysis platform.

3.1 Analysis tasks

Our users mainly focus on the occurrence of educational events, responses of the public, correlations between events and policies, which can assist them to make decisions or adjust related policies. EduVis provides a new perspective to deal with data analysis tasks and reduces the workload of analysis compared with traditional

statistic methods. After discussing with education managers, we summarize the following 3 main analysis tasks.

(1) Users need to integrate and analyze information that includes people, place, organization and media reports related with educational data.

(2) Finding related educational events and the shared entities. For instance, events A and B are directly related, whether A and B are connected by entities like people or organizations.

(3) Displaying detail information about entities including persons, places and organizations in order to make users directly know the detailed content of events.

3.2 Data set

All the data used in this paper are extracted from web containing the network public opinion information from January to November 2015, and the total quantity of the data of the network public opinion information is more than 910,000 articles including url, title, public time, public source, author, content, comment information and so on. Among them, microblog and forum data account for 38.62%, education blogs 18.01%, academic journals of Education 17.65%, information portal platform 15.96%, government education website and traditional media 8.80% and 0.95%. The data set also includes information box data extracted from BaiduBaike and some other web data.

3.3 Framework views

EduVis includes 7 major views illustrated in Fig. 2(b): (1) Knowledge tree view has classification of educational events. Users can obtain related education knowledge graph by directly selecting different category or using search box. (2) Event network view is constructed to show connections of events by shared entities such as organization, people and place. (3) Click tracking view helps user to find connections between events by recording entities that user has focused in events network view and to backtrack for comparative analysis. (4) Original articles view shows titles and summaries of articles which are related with the events in events network view. Users can link to original webpage by clicking the title of article. (5) Entity detail view displays detail information of entity that user may interested in. (6) Entity list view shows a list including name, type, weight and degree of all entities. (7) Word cloud view produces word cloud related to events based on entities' information of event, making user directly find those highly weighted entities and their types.

4 Knowledge Graph Processing

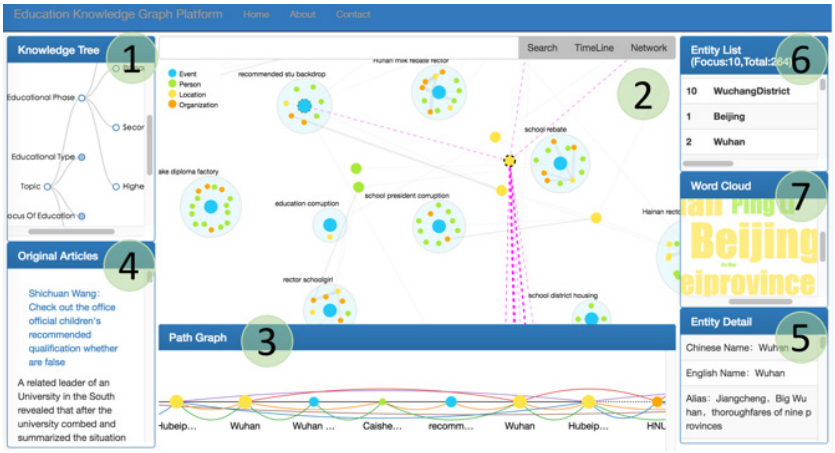
Topic Mining. Topic model is usually used to mine abstract topics in document sets in the field of natural language processing. LDA (Latent Dirichlet Allocation)[15] is adopted to process the data of educational public opinion and model for massive text corpus to find the themes hidden in it. Our data modeling and classification processing remove information unrelated to educational public opinion while retaining information related with that.

Entity Extraction and Normalization. We utilize NER techniques to recognize entities of various types in documents. Additionally, to improve the data quality of entities, we normalize entities by mapping surface forms to unambiguous

references. Given a collection of unnormalized entities M recognized by NER models, we filter out noisy or incorrect entities. The techniques for entity normalization include sub-string matching and entity disambiguation, introduced in Ref. [16].

Semantic Relation Extraction. Linguistic and statistical features are analyzed to identify candidate relation tuples[17], 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. We cluster entity pairs which have similar contexts together as a raw relation to label the extracted candidate relations, i.e., $R = \{(e_i, e_j, C_{i,j})\}$. The keywords for the raw relation R are labeled by extracting the frequent keywords in $C_{i,j}$ for all e_i and e_j pairs.

(a) Word cloud treemap



(b) Main view of the platform

Figure 2. Visual analysis platform for education knowledge graph.

5 Visualization Design

Education knowledge graph is mainly composed of entities, relations between entities and the relevant information. For making a clear display of education knowledge graph, we choose four kinds of entities and the relations between them to make a visualization analysis, including event, person, place and organization. We construct EduVis with event network view as the core view and word cloud treemap view, click tracking view, word cloud view and so on.

5.1 Module of word cloud treemap

Education knowledge graph is very large for users to view integrally, so we design an overview for education knowledge graph to give users the overall impression for it. After clicking the specific module of it, we can enter the main interface. We combine word cloud with treemap[18] that is used for structured data to visualize all education events and the education knowledge tree classified by education experts. In order to present clear classification and to acquire better visual effect, the treemap is colored according to Ref. [19]. The size of every rectangular changes according to the number of events of this category labeled in the top left corner. The layout schematic is shown in Fig. 3.

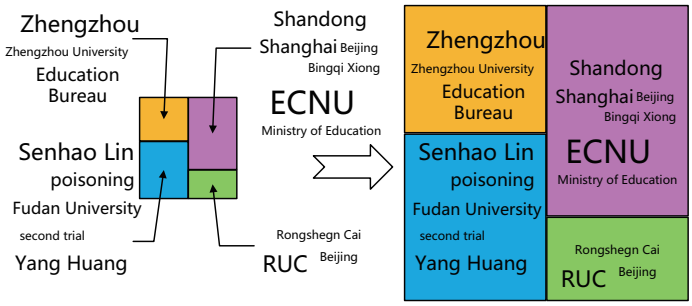


Figure 3. Combining treemap with word cloud.

5.2 Module of event network graph

Network graph is effective to make visual analysis of linked and relational data. We construct event network graph labeled 2 in Fig. 2(b) according to the features of education knowledge graph. The nodes of the graph represent the entities of knowledge graph, the colors of nodes represent different categories and lines between nodes represent the relations of entities, showing in Fig. 7.

5.2.1 Event network graph based on topology

Force-directed layout algorithm[20] is mainly used in building network graph. It can fully display the structure of network, so we choose it as basic layout algorithm and make improve on it. There are four steps for building the layout of event network graph based on topological structure.

Layout of Event Nodes. We adopt force-directed layout algorithm to position the nodes of event. To make the nodes of event have place to lay out their private entities, we set the value of repulsive force according to the weight of event nodes and

the number of the private entities, as shown in Fig. 4. Then we can get the positions of nodes which are used to compute the locations of other entities.

Layout of Entities Belonging to One Event. We adopt the method of ring layout to position the nodes which belong to one event at the ambient area of the corresponding event node. The calculating formulas of the method of ring layout are:

$$t = \frac{2\pi i}{len} + d1 \tag{1}$$

$$x = a + (r + d2)cost \tag{2}$$

$$y = b + (r + d2)sint \tag{3}$$

where, (a, b) is the center coordinate, r is the radius of circle, len is the number of private nodes, i is the ith of the private nodes, $d1$ and $d2$ are random numbers. To enhance the effect of cluster, we add $d1$ and $d2$. There is no need to optimize the layout of private nodes for the simplicity of the relations of entities. We can get the layout by using the method of ring layout, as shown in Fig. 5.

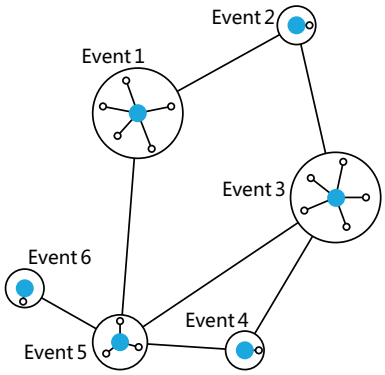


Figure 4. Force directed layout.

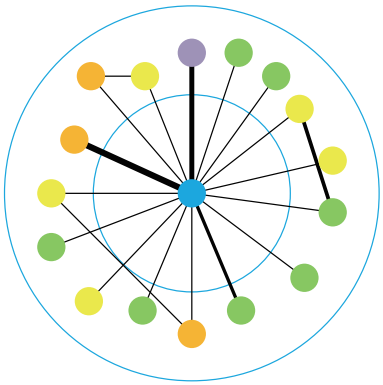
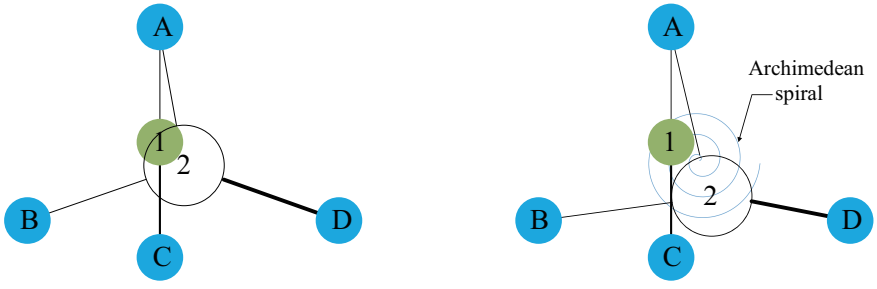


Figure 5. Ring layout.

Layout of Entities Belonging to Multiple Events. The entity which belongs to multiple events is located at the center of the multiple event nodes. In order to avoid covering the existing nodes, the method of spiral scanning is used for collision detection. The principle of spiral scanning method is Archimedes Spiral which is the trail generated by one point leaves the fixed position with mean speed and rotates around a fixed point with fixed angular speed as shown in Fig. 6.



(a) Colliding layout (b) Adjusted layout

Figure 6. The layout of spiral scanning method.

The calculation formulas of plane Cartesian coordinates of Archimedes Spiral are:

$$r = x(1 + t) \tag{4}$$

$$x = r\cos(2\pi t) \tag{5}$$

$$y = r\sin(2\pi t) \tag{6}$$

After ascertaining the locations of all the nodes, we should draw the lines according to the positions of nodes and the relations between them. Every line has weight, and we set the weight value of the lines regarding to the times of the line shows up, a higher weighted value means more important relation between nodes. The thickness of the line and the depth of the color mean the weight of line, the higher weighted value means deeper of the color and thicker of the line.

5.2.2 Event network graph based on timeline

As shown in the legend of Fig. 7, blue node represents event, green node represents person, yellow node represents location, orange node represents organization. After clicking a node, it and the nodes which are directly connected with it will have dashed borders. The lines which are directly connected with it will become dashed lines. 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[21] illustrated in Fig. 8 generated by the same data as Fig. 7. Compared Fig. 8 with Fig. 7, 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.

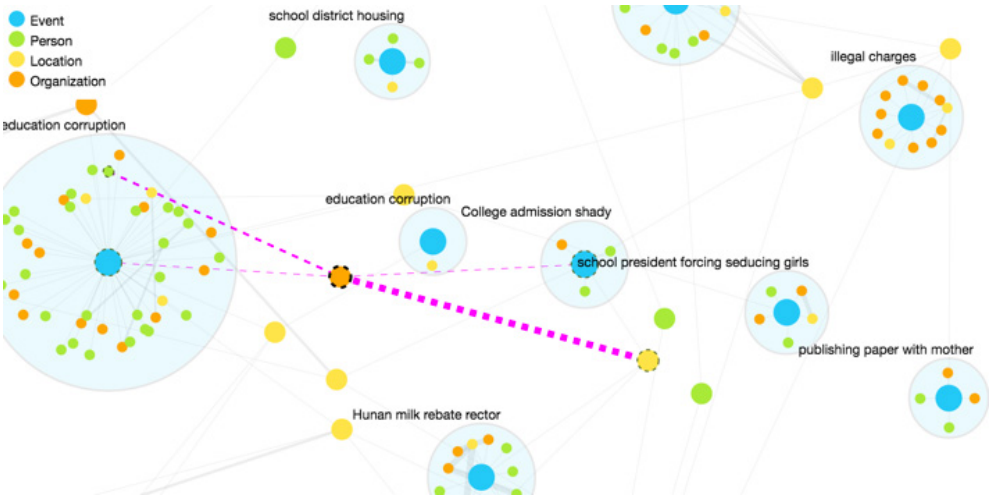


Figure 7. Event network graph based on topology.

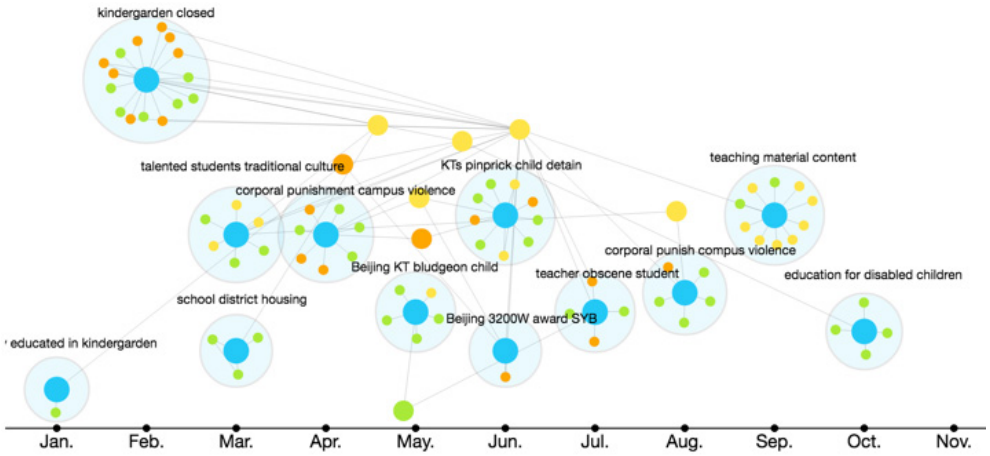


Figure 8. Event network graph based on timeline.

5.3 Module of click tracking graph

We offer users click tracking graph labeled 3 in Fig. 2(b) to backtrack when they want to repeat and contrast the analysis. In event network graph, users can click the nodes concerned in order to get the detailed information, meanwhile the entity clicked will be added to click tracking graph. If the node is not isolated from the last node clicked, we draw a solid line between them, otherwise the line is dashed. The newly added node d is connected to the nearest node that belongs to the same event through a curved line, as shown in Fig. 9.

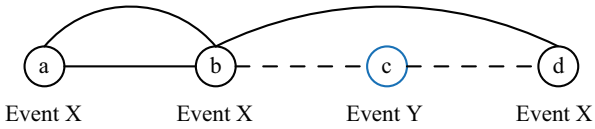


Figure 9. Adding node in click tracking graph.

When the mouse is moved to a line, it will be highlighted in bold, while the name of the event represented by the line shown at the bottom. Besides, the dot will be highlighted after being clicked, at the same time the event network graph based on topological structure will re-layout with the node as the core.

6 Case Study

We implement EduVis with the mentioned visualization methods. In this section, we will validate our methods through case studies.

6.1 Association analysis by place

Users need to find events happened in a place like Beijing, Shanghai, and related details information can help users clearly know what happened in this place.

We choose Beijing, the capital of China and the place where educational events frequently happen, as our case. After we input “Beijing” into the search box, we

can get the result illustrated in Fig. 10. We could find 7 events are connected by "Beijing", four of which have private entities connecting with entity "Beijing". All the events may have different scales, but they always have some connections with other events through some entities. Besides, one organization entity connected with some events has direct relation with entity "Beijing". We could find some entities such as "kindergarten teacher hit students by stick was administrative detention", "early educated students in kindergarten", "Beijing kindergarten teacher pricked students was administrative detention" connected with entity "Beijing", so users learn about the detailed description of educational events of specific type. Through above analysis, we can get the conclusion that many events happened in Beijing are about preschool education, campus security and children abuse. Users of EduVis, education decision makers, should strengthen supervision and make more effective policies in these areas.

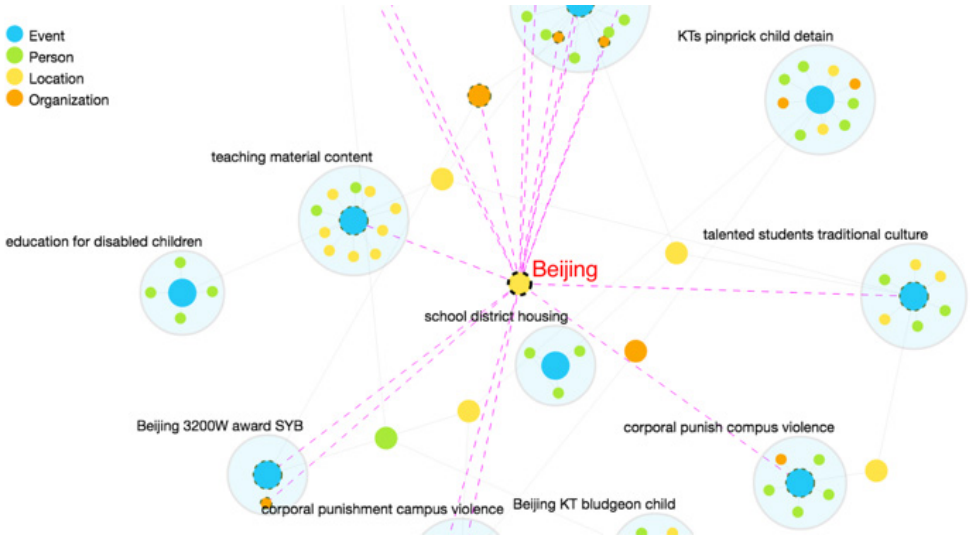


Figure 10. Association analysis by "Beijing".

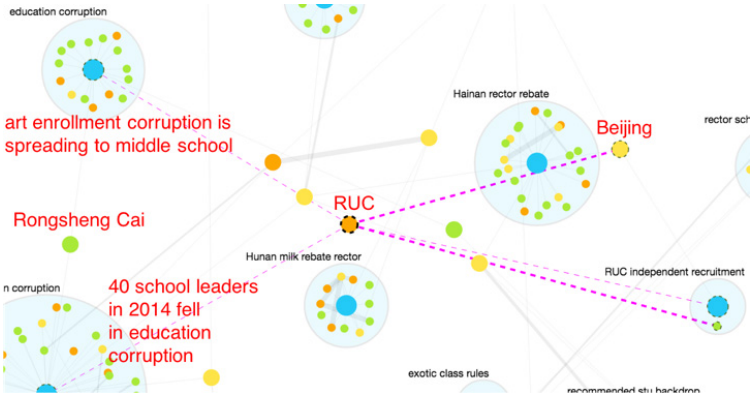
The event network graph based on timeline illustrated in Fig. 8 could be seen by clicking the Timeline button. We could find that the quantities of events happened in March and June are more than other months. After attentive analysis, we find that it is because the new semester begins in March and college entrance examination is held in June.

6.2 Analyzing events by shared entities

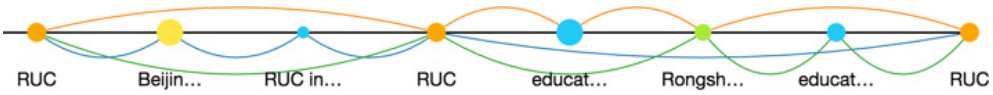
In the first interface of EduVis, we can find the prominent events in each category and education corruption is a hot issue in educational events, so we choose it as a case illustrated in Fig. 11. After clicking the module of education corruption, we can enter the main interface and click entity "Renmin University of China", then appearing 3 events "more than 60% students were recommended by experts and officials in RUC independent enrollment", "40 school leaders in 2014 fell in education corruption", and "art enrollment corruption is spreading to middle school". At the same time, RUC is connected with entity "more than 60% students were recommended by experts and officials in RUC independent enrollment"

through “Beijing”, when we clicked “Beijing”. At the moment, we can find “RUC”, “Beijing” and “more than 60% students were recommended by experts and officials in RUC independent enrollment” belonging to event “more than 60% students were recommended by experts and officials in RUC independent enrollment” and entity “RUC” belonging to the forementioned 3 events in click tracking view. Then, after clicking entity “40 school leaders in 2014 higher education corruption”, we can see entity “Rongsheng Cai” connecting with this event. Meanwhile, “40 school leaders in 2014 higher education corruption” and “art enrollment corruption covers high school” have the shared entity “Rongsheng Cai”.

(a) Classification of educational corruption



(b) Detail information of topology



(c) Click tracking graph

Figure 11. Analyzing events by shared entities.

Our new layout method shows relations of events and shared entities more clearly. Through the event network graph, each event and entities that belong to it have closed

position, which makes users directly have insight into this event. In this main view, there are shared entities by many events except events layout, and all of that can support users to analyze event itself and relationships between events through shared entities.

According to above analysis, users could find related events by an organization entity "RUC". Educational events connected with this entity involve in school corruption especially in university enrollment process. Additionally, users can focus on specific event, analyze entity shared by events like entity "Rongsheng Cai" that plays an important role in 2 events, and take advantage of entity description information for deeply and soundly understanding event and relation between events. After reading related reports, we find that Rongsheng Cai who provided help to candidates of special types of school enrollment process and accepted bribes more than 10 million RMB connected with events "more than 60% students were recommended by experts and officials in RUC independent enrollment", "40 school leaders in 2014 fell in education corruption". All of analytical process can assist users to understand what happened in the field of education and make decisions or policies.

7 Conclusion

Education is one of the most concerned fields at present and it is very meaningful to dig out the information hidden in educational data effectively. Knowledge graph technique can reveal the dynamic development rule in complex knowledge fields, which is widely used in various fields, and visualization technique can clearly show the inner structure of knowledge graph.

This paper discusses how to construct education knowledge graph combining with new visualization methods to present data characteristics and potential patterns. To begin with, education knowledge graph is constructed based on educational data that extracted from web data by adopting natural language processing technique. Next, we combine word cloud with treemap to present the abstract structure of education knowledge graph for users. Besides, a visualization method based on topological structure of event network and timeline is designed and implemented. At the same time, we present a method of click tracking concerned by users to analyze the connections of events with the exchange of users' attentions. Finally, EduVis which assists users to analyze educational data visually is constructed.

In this paper, we only consider the basic relationship extraction instead of a more in-depth and accurate relationship extraction, so we will work further with this problem in future work to show more latent information of data. Moreover, the data type displayed is relatively single, so the information displayed by education knowledge graph is not perfect enough. This is another problem we should consider in the future.

Acknowledgements

Thanks for Xiaofeng He, Wenliang Cheng, Chengyu Wang, Yanhua Wang, Sicheng Dai, Yitao Liang, et. al help to provide the education data.

References

[1] Zhang L. Knowledge Graph Theory and Structural Parsing. Enschede, Canada: Twente University Press, 2002: 19–40.

[2] Lehmann J, Isele R, Jakob M, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 2015, 6(2): 167–195.

[3] Suchanek FM, Kasneci G, Weikum G. Yago: A core of semantic knowledge. Proc. of the 16th International Conference on World Wide Web. Alberta, Canada. ACM. 2007. 697–706.

[4] Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction. Proc. of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK. Association for Computational Linguistics. 2011. 1535–1545.

[5] Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka Jr ER, Mitchell TM. Toward an architecture for never-ending language learning. Proc. of Twenty-Fourth AAAI Conference on Artificial Intelligence. Atlanta, USA. 2010. 529–573.

[6] Banko M, Cafarella MJ, Soderland S, Broadhead M, Etzioni O. Open information extraction for the web. IJCAI, 2007, 7: 2670–2676.

[7] Wu W, Li H, Wang H, Zhu KQ. Probase: A probabilistic taxonomy for text understanding. Proc. of the 2012 ACM SIGMOD International Conference on Management of Data. Arizona, USA. ACM. 2012. 481–492.

[8] Hulliyah K, Kusuma HT. Application of knowledge graph for making text summarization (analizing a text of educational issues). International Conference on Information and Communication Technology for the Muslim World (ICT4M). IEEE. 2010. 79–83.

[9] Zhu Y, Cao X, Bian Y, Wu J. CKGHV: A comprehensive knowledge graph for history visualization. Proc. of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries. London, UK. IEEE Press. 2014. 437–438.

[10] Yao Y, Wang X, Meng X, Zhang X, Li B. ISEK: An information security knowledge graph for CISP knowledge system. 2015 5th International Conference on IT Convergence and Security (ICITCS). IEEE. 2015. 1–4.

[11] Lai W, Huang X. From graph data extraction to graph layout: Web information visualization. 2010 3rd International Conference on Information Sciences and Interaction Sciences (ICIS). IEEE. 2010. 224–229.

[12] Gibson H, Faith J. Node-attribute graph layout for small-world networks. 2011 15th International Conference on Information Visualisation (IV). IEEE. 2011. 482–487.

[13] Tang Y, Wang B, Fan Q. An improved graph layout algorithm of embedded node attributes. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(2): 228–237.

[14] Wu P, Li SK. Layout algorithm suitable for structural analysis and visualization of social network. Journal of Software, 2011, 22(10): 2468–2475.

[15] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993–1022.

[16] Jijkoun V, Khalid MA, Marx M, De Rijke M. Named entity normalization in user generated content. Proc. of the Second Workshop on Analytics for Noisy Unstructured Text Data. New York, USA. ACM. 2008. 23–30.

[17] Shen W, Wang J, Luo P, Wang M, Yao C. REACTOR: A framework for semantic relation extraction and tagging over enterprise data. Proc. of the 20th International Conference Companion on World Wide Web. New York, USA. ACM. 2011. 121–122.

[18] Shneiderman B, Wattenberg M. Ordered treemap layouts. Proc. of the IEEE Symposium on Information Visualization (INFOVIS'01). California, US. IEEE. 2001. 73.

[19] Tennekes M, De Jonge E. Tree colors: Color schemes for tree-structured data. IEEE Trans. on Visualization and Computer Graphics, 2014, 20(12): 2072–2081.

[20] Itoh T, Mueldser C, Ma KL, Sese J. A hybrid space-filling and force-directed layout method for visualizing multiple-category graphs. IEEE Pacific Visualization Symposium. 2009. 121–128.

[21] Nguyen PH, Xu K, Walker R, Wong BW. TimeSets: Timeline visualization with set relations. Information Visualization. 2015: 1473871615605347. DOI: 10.1177/1473871615605347.