Automatic Construction of Subject Knowledge Graph based on Educational Big Data

Ying Su
College of Information Science and Engineering,
Wuchang Shouyi University
Wuhan, Hubei Province, China
+86 18995637068
suying929@163.com

Yong Zhang*
Computer School,
Central China Normal University
Wuhan, Hubei Province, China
+86 15802766977
ychang@mail.ccnu.edu.cn

ABSTRACT
In this paper, we propose an automatic construction method of subject knowledge graph for educational applications. The subject knowledge graph is constructed based on educational big data by using a bootstrapping strategy to gradually expand knowledge points and connections between them. In this paper two different datasets are used. One is the subject teaching resources such as syllabuses, teaching plans, textbooks and etc., which is used to automatically construct the core of subject knowledge graph so as to reduce the dependence on the manual annotation. Meanwhile the high-quality of subject teaching resources is the guarantee of accuracy of the knowledge graph core. The other dataset is the massive Internet encyclopedia texts, which is used to expand and complete the subject knowledge graph. As to algorithm, this paper utilizes the BERT-BiLSTM-CRF model to automatically identify the subject knowledge points, and then evaluates the relationship between the knowledge points by calculating their semantic similarity, PMI and Normalized Google Distance between them. The experimental results show that BERT-BiLSTM-CRF outperforms the baselines significantly, and the three kinds of relationship evaluation models have achieved good results. Finally, computer science and physics science are taken as examples to construct the subject knowledge graphs successfully, which show the effectiveness of our method.

CCS Concepts
• Computing methodologies→ Information extraction

Keywords
knowledge graph; normalized google distance; point mutual information; BERT-BILSTM-CRF; intelligent education

1. INTRODUCTION
With the rapid development of the information industry and artificial intelligence, intelligent education has gradually become research hotspots in education. As we all know, the nature of education is to impart knowledge to students. Therefore, subject knowledge has naturally become the core resource of education, and the representation of knowledge has become one of the important tasks of intelligent education.

In traditional teaching mode in classroom, the subject knowledge system is usually divided into different courses and teaching activities are carried out in the order of chapters of the course textbook. In fact, the courses of the subject in education are interconnected, and the knowledge points of subject are essentially an interconnected whole. Therefore, researchers try to use knowledge graph (KG) to describe the knowledge structure of educational subject, and transform the traditional tree-like knowledge structure into a graph-like structure. Based on the subject knowledge graph, we will be able to break through the traditional linear teaching model and carry out a series of non-linear teaching research and applications, including micro-teaching, learning assessment, learning navigation, and personalized recommendation of learning resources based on the subject knowledge graph.

The construction of subject knowledge graph mainly uses natural language processing, data mining, machine learning and other technologies to automatically mine knowledge points and relationships from large-scale subject text data. However, due to technical limitations, the construction of subject knowledge graph heavily relies on training dataset labelled by the subject experts, which requires a lot of human labor. Thus the knowledge graph is hardly updated in time to accurately reflect the subject knowledge structure while subject knowledge is constantly evolving with the development of the subject.

To this end, this paper proposes an automatic construction method of Chinese subject knowledge graph based on educational big data. First, subject teaching resources are used to guide the construction of subject knowledge graph, which can greatly reduce the manual labor of subject experts. As we all know, in the long-term teaching activities, teachers have accumulated a lot of teaching resources, for example lesson plans, textbooks, syllabuses, examination papers and etc., which are all closely related to the corresponding subject. In particular, the syllabus is usually designed by subject experts in a standardized format, and clearly lists the knowledge points of corresponding course. So, we will make full use of the teaching resources of the subject to construct the core subject knowledge graph. Secondly, this method utilizes large-scale internet encyclopedia text to expand the subject knowledge graph and ensure the integrity of the subject knowledge graph as much as possible. Due to the limited scale of subject teaching resources, some subject knowledge points would be missed inevitably, and the links between knowledge points would be difficult to accurately assess. Therefore, this paper attempts to use the Baidu Encyclopedia to complete the subject knowledge graph like the DBPedia, which is constructed from Wikipedia.

ICBDE '20, April 1–3, 2020, London, United Kingdom

ACM ISBN 978-1-4503-7498-9/20/04…$15.00
DOI: https://doi.org/10.1145/3396452.3396458

Based on the above strategies, this paper proposes a bootstrapping method for construction of subject knowledge graph. Firstly we extracts core knowledge points from the subject syllabus, and evaluates the connection between knowledge points based on educational big data, thereby complete the construction of the core subject knowledge graph. Then, based on this core knowledge graph, more new knowledge points will be identified further and merged into the subject knowledge graph, and the connections between knowledge points are evaluated. Thus the subject knowledge graph is updated through this continuous iteration until no more knowledge points and connections could be found.

2. RELATED WORK

The concept of knowledge graph was first proposed by Google in 2012, and has since quickly attracted widespread attention. A large number of open fields and specific knowledge graphs have been continuously constructed and applied, such as Freebase, Google Knowledge Graph, Baidu Zhixin, Fudan's CN-DBpedia [1] and Academic graph OAG published by Tsinghua University and Microsoft [2].

For the construction of knowledge graphs, the current solutions are mainly divided into two types. The first is manually construction methods, such as "WordNet", "HowNet" and other semantic dictionaries, and collaboratively constructed large-scale knowledge such as "FreeBase" etc. However, this method requires a lot of manpower resources, and the knowledge graph constructed is difficult to update. Even for a large-scale knowledge base like "FreeBase", the amount of knowledge is still far from enough. Studies have shown that FreeBase only contains 6% of the parental relationship knowledge that has been found, and only 8% of the relationship between husband and wife.

The second way to build a knowledge graph is to use information extraction technology to automatically extract entities and entity relationships from large-scale texts to build a large-scale entity relationship database. For example, the large-scale knowledge base "DBPedia" is a cross-domain, multi-language, large-scale data set formed by using a knowledge extraction technology to extract entities and entity relationship from Wikipedia [3]. The current construction methods by information extraction mainly include two categories. The first is a rule-based method. The main principle is to construct a series of extraction rules, and then use the rule to extract entities and entity relationships from the text [4]. The core problem of this method is the construction of rules, which mainly includes manually constructing a rule base, or using machine learning methods to automatically learn and extract rules from labeled corpora [5][6]. The second is an automatic construction method based on machine learning, that is, an amount of corpus are manually labelled for training a learner which is used to automatically determine whether a certain word is an entity or an entity relationship [7]. But this type of methods also requires a large number of manually labeled data sets.

In order to build a large-scale knowledge base, and to get rid of the restrictions on manual annotation, in recent years, researchers have begun to use automatic extraction strategies such as bootstrapping and semi-supervision. Such as NELL Knowledge Base System (Never-Ending Language Learner), this system uses a small amount of artificial ontology, a small amount of predicate relationships between ontologies, and a small amount of manual intervention as a guide. The algorithm automatically extracts more entities and entity relationships from more than 1 billion web pages. Its optimization principle lies in iterative learning that never stops. It has been running continuously since its creation in January 2010 and has achieved good results [5]. In recent years, researchers have begun to apply deep learning to entity relationship extraction. Zeng et al. proposed an entity relationship classification method based on convolutional neural networks, where the experimental results show that the combination of deep convolutional network and expert knowledge base can significantly improve the performance of relation extraction [8]. Yoo and Jeong proposed an unsupervised method to automatically expand a knowledge graph by crawling and analyzing the news sites and social media in real-time. They utilized and fine-tuned the pre-trained multilingual BERT model for the construction of knowledge graphs，and then extract new relationships using the BERT-based relation extraction model and integrate them into the knowledge graph [9]. Mao et al proposed a bootstrapping knowledge graph builder, which employed a neural relation extractor resolving primary relations from input and a differentiable inductive logic programming (ILP) model that iteratively completes the knowledge graph [10]. Some researchers have tried various unsupervised method to construct domain knowledge graph for text and the experimental results have shown the effectiveness of these methods [11].

In summary, the construction and application of knowledge graphs has become a research hotspot at present. In particular, how to reduce the manual labor in the construction of knowledge graphs has been paid much attention. Meanwhile, in intelligent education, the knowledge graph for specific education subjects will become an important basic resource for educational applications, such as subject knowledge derivation, learning automatic evaluation, learning navigation, learning resource recommendation and etc. Therefore, automatic construction of knowledge graph for educational subject will play an important role in intelligent education.

3. METHODOLOGY

This paper uses two different text resources. One is a subject teaching resources, which mainly includes textbooks, syllabuses, and lesson plans. These text data are closely related to the subject and are a guarantee for constructing a high-quality subject knowledge graph. The other is an online encyclopedia resource, which belongs to an open large-scale online learning resource. Its purpose is mainly to improve the completeness of the knowledge graph of the subject.

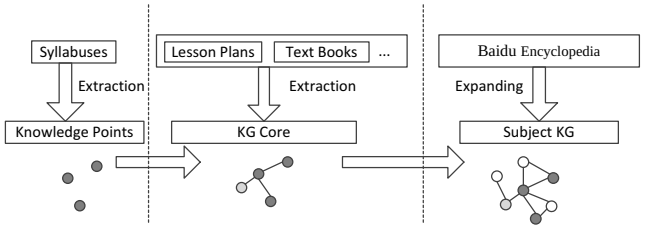

Figure 1. The architecture of subject KG construction

The method in this paper mainly includes three basic steps. First, we automatically extract the core knowledge points of the subject from the course syllabus; then, we use the subject teaching resources as the data source and use the subject knowledge point recognition algorithm to obtain more new knowledge points, and evaluate the strength of the association between the knowledge points to construct a small-scale core subject knowledge graph; finally, we use large-scale network encyclopedia resources as data sources to expand the core knowledge graph to obtain a complete subject knowledge graph, its basic principle is shown in Figure 1. In summary, this method includes two core algorithms, one is the

automatic identification of knowledge points in the subject, and the second is the evaluation of the strength of the association between the knowledge points.

3.1. Identification of Core Knowledge Points

Course syllabuses are usually prepared manually by experts in the subject area. Generally speaking, the syllabus will clearly indicate important or difficult knowledge points for each course. The knowledge points clearly identified by experts in these fields belong to the core of the entire subject.

The extraction of core knowledge points is relatively simple, because the syllabus usually has a relatively fixed and standardized format, and the important knowledge points are clearly listed in it. Thus, we only need to construct appropriate regular expressions according to the corresponding format, and design program to automatically extract the core knowledge points from the course syllabuses.

Taking computer science as an example, the syllabus of 12 core courses was processed in this paper, and a total of 322 core knowledge points were extracted.

3.2. Recognition of Knowledge Points

On the basis of core knowledge points, this paper needs to further automatically identify more knowledge points from large-scale subject texts, including subject teaching resources and online encyclopedia resources. To this end, this paper uses a joint recognition model based on BERT-BiLSTM-CRF, as shown in Figure 2.

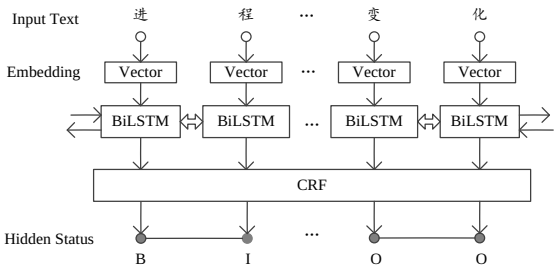

Figure 2. BERT-BiLSTM-CRF model

BERT-BiLSTM-CRF is the mainstream solution in the field of named entity recognition. In this model, the BERT model is a pre-trained model, which is mainly used to obtain the semantic vector (Embedding) of each Chinese character. This paper uses Google's open-source Chinese BERT model and its pre-trained results. In the core part of the model, the traditional CRF model is integrated with the BiLSTM deep neural network model, which makes the model have automatic feature selection capability and more powerful ability to capture contextual semantic information.

In terms of training data construction, in order to avoid large-scale manual labeling, this paper draws on the idea of remote supervision in relation extraction, and uses an automatic labeling method of training data based on the core knowledge point set, that is, we are in the corpus Find the terms in the core knowledge point set. Any matching part is considered as an instance of the corresponding term, and the corresponding sentence is automatically labeled. In this paper, the training data is automatically labeled on the subject teaching resources and the network encyclopedia resources. After manual inspection, it is found that the training data marked by the remote supervision method does not contain much noise data.

Similar to the labeling of named entity recognition, this paper uses a concise and efficient BIO labeling mechanism. BIO labeling is the word-level position labeling for each entity, where the tag 'B' indicates the beginning character of a named entity, the tag 'I' indicates the interior characters of a named entity, and the tag 'O' indicates the characters that do not belong to any named entity.

3.3. Knowledge Point Association Assessment

The strength of the association of two terms usually includes two kinds, one is the semantic similarity of the two terms, and the other is the semantic association of the two terms.

The semantic similarity of two terms is usually evaluated by the context of the term. The implicit meaning is that the context of the term represents the semantics of the term. That is, the closer the contexts of the two terms are, the more similar the semantics of the two terms is.

The semantic relevance of two terms is usually evaluated by the probability that two terms appear in a piece of text at the same time, that is, if the two terms often appear together, there may be some kind of semantic association between the two terms, for example, in the course of operating system, although the two knowledge points of "memory allocation" and "page" are not semantically similar, they have strong semantic relevance.

For this reason, this paper comprehensively evaluates the strength of semantic association between knowledge points from these two aspects: semantic similarity and semantic relevance between knowledge points.

3.1.1 Semantic Similarity of Knowledge Points

For the semantic similarity evaluation of the two terms, this paper adopts the similarity measurement algorithm based on context distribution. The basic principle is to count all the context words of the two terms, convert their context representation into vector representation, and then calculate the cos value of the angle between the two vectors. The specific algorithm description is shown in Table 1.

Table 1: The process of knowledge points similarity evaluation

Given two knowledge points *x* and *y*
  Begin
    For *t* in set *(x, y)*
      The text fragments containing *t* are retrieved from the corpus.
      Take *k* words before and after *t* from each fragment as its context
      Accumulate the semantic vectors of all context words as the context vectors of *t*
    End For
    Calculate cos value of *x* and *y* context vectors
    If the cos value is greater than the given threshold *s*, then *x* and *y* are considered to be semantically similar.
End

In this algorithm, the parameter *k* is used to control the context window size *n*, that is, $n = 2k$ and the parameter s represents the threshold of semantic similarity.

In the experiment, the semantic vector representation of words uses Google word2vec pre-trained data. Although the BERT model can also obtain the semantic vector representation of words, related experimental results show that word2vec is more suitable for evaluating the context of two terms than the BERT model Semantic similarity.

3.1.2 Semantic Relevance of Knowledge Points

For the evaluation of the semantic relevance of two terms, this paper uses two different evaluation strategies, one is a point mutual information (*PMI*) algorithm based on educational text data, and the other is the Normalized Google Distance [12] (*NGD*) evaluation algorithm based on Baidu search engine.

3.1.2.1 PMI

The educational text data of this paper includes subject teaching resources and network encyclopedia resources, which are the core data sources for the construction of subject knowledge graphs. On this data, this paper uses pointwise mutual information to evaluate the connection between the two terms *x* and *y*. The formula is shown in Equation 1.

$$PMI(x,y) = \log_2 \frac{p(x|y)}{p(x)} = \log_2 \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

Where, *p(x)* represents the probability of the term *x* appearing in the dataset, *p(y)* represents the probability of the term *y* appearing in the dataset, and *p(x, y)* represents the probability of the simultaneous occurrence of *x, y* in a piece of text. Generally speaking, a piece of text can refer to a discourse, a paragraph, a sentence, or text snippet. In this paper, we assume that if two terms appear in the text at no more than *m* words, they are considered to co-occur once, that is, the size of the co-occurrence window is *m*.

In information theory, *PMI* is often defined as the connection between random variables, and it is also widely used in tasks such as term extraction in NLP. *PMI* meets non-negativity and symmetry. Its value has the following two characteristics: when the two terms are completely unrelated, *PMI* is zero; when two terms are relevant, the range of *PMI* is (0, ∞). Generally speaking, we will choose an appropriate threshold *t* according to the size of the corpus to measure whether there is semantic relevance between the two terms. When the *PMI* value of the two terms is greater than the threshold *t*, a connection between the two terms will be created.

3.1.2.2 NGD

Due to the inevitable limitations of manually collected data, this paper also uses the Google Distance algorithm to evaluate the semantic relevance of two terms by searching for the frequency of co-occurrence of the two terms throughout the Internet. The formula of *NGD* is shown in Equation 2.

$$NGD(x,y) = \frac{\max\{logf(x), log\,f(y)\} - logf(x,y)}{logN - \min\{logf(x), logf(y)\}} \quad (2)$$

Where *N* is the total number of web pages indexed by Google, *f(x)* and *f(y)* are the numbers of hits for searching terms *x* and *y* respectively, and *f(x, y)* is the number of web pages in which both *x* and *y* occur. The closer *NGD(x, y)* approaches 0, the higher association between the two terms will be. If *NGD(x, y)* is greater than or equal to 1, the relevance between the two terms is small.

The *NGD* is a semantic similarity measure derived from the number of hits returned by the Google search engine for a given set of keywords. Keywords with the same or similar meanings in a natural language sense tend to be "close" in term of *NGD*, while words with dissimilar meanings tend to be farther apart.

3.2 Construction of Subject KG

This paper adopts a bootstrapping construction strategy for subject knowledge graph, as shown in Figure 3. The algorithm gradually adds more knowledge points and their relationships to the core knowledge graph through continuous iteration. The main steps are as follows.

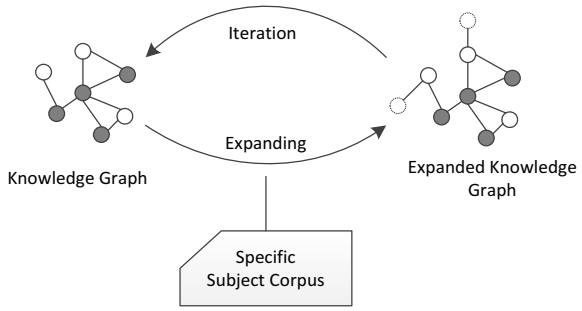

Figure 3. Schematic diagram of subject KG construction

(1) Based on the core knowledge points, we use three algorithms to evaluate the strength of the association between two knowledge points. If the strength of the association between two nodes is greater than a certain threshold, add one side. Through calculation, we can finally obtain an undirected graph of knowledge points, which we call the core knowledge graph;

(2) On the basis of the existing knowledge graph, we continue to search for text data related to existing knowledge points, identify more knowledge points based on the data, and evaluate the connections between their knowledge points;

(3) We calculate the minimum distance from the extended new knowledge point to the core knowledge point. If the distance is greater than k, the knowledge point will be deleted;

(4) Continue the loop from (2), and iteratively evaluate until no new knowledge points are found.

In the subject knowledge graph, the distance between knowledge points refers to the minimum number of edges passing from one knowledge point to another knowledge point. For example, the distance is one between two adjacent nodes since there is only 1 edge between them. The distance from a new knowledge point to the core knowledge points refers to the minimum value of the distances between the new knowledge point and each one of the core knowledge point. Therefore, the parameter *k* has become a key parameter for controlling the expansion of the knowledge graph.

4. EXPERIMENTS AND RESULTS

This paper takes computer science as an example, and collects subject teaching dataset (denoted by M) and online encyclopedia dataset (denoted by N). Then we adopt Distant Supervision [13] method to automatically label the datasets with 322 core knowledge points. The specific corpus size and labeled sentence size are shown in Table 2.

Table 2. Datasets used in this paper

Datasets	Size	# of labelled sentences
M	6M	8834
N	500M	48657

Based on these datasets, this paper conducts experiments in three aspects: automatic recognition of knowledge points, assessment of the strength of knowledge point associations, and construction of

knowledge graph, and the corresponding results are evaluated and discussed.

4.1. Recognition of Knowledge Points

This paper uses the open source BERT-BiLSTM-CRF model [14] on GitHub, where the parameters are set to default values and the platform is TensorFlow 1.11.0. This paper also uses CRF and BiLSTM-CRF models as baselines for comparative experiments, where the open-source CRF ++ 0.54 is utilized and Google pre-trained word2vec word vector is used as input for BiLSTM-CRF model.

This paper evaluates the automatic recognition of knowledge points using 5-Fold cross-validation. The two types of training corpora are randomly divided into 5 parts, each of which uses 4 parts as the training set and the remaining 1 part as the test set. The evaluation measures include the precision (*P*), recall (*R*) and *F1* values. The experimental results for the identification of the core 322 knowledge points are shown in Figure 3.

Table 3. Results of recognition of knowledge points

	Datasets	*P* (%)	*R* (%)	*F1* (%)
CRF	M	85.5	81.8	83.6
	N	83.6	82.4	83.0
BiLSTM-CRF	M	92.7	88.4	90.5
	N	90.8	89.6	90.2
BERT-BiLSTM-CRF	M	95.2	92.8	**94.0**
	N	94.1	**93.4**	93.7

From the experimental results, it can be seen that the BERT-BiLSTM-CRF model performs best on text corpora. Its knowledge point recognition precision and recall are the highest. Compared with the traditional CRF model, The *F1* values on each data set were increased by 10.4 and 10.8 percentage points, respectively, and the experimental results were significantly improved compared to the BiLSTM-CRF model. The experimental results show that BERT-BiLSTM-CRF can effectively identify subject knowledge points.

In the experiment, in addition to the core knowledge points, the three models identified more knowledge point terms at the same time. In this paper, terms that appear more than 3 times at the same time are selected as new knowledge points, and the accuracy is manually evaluated. The evaluation results are shown in Table 4.

Table 4. Accuracy of recognition of new knowledge points

	Datasets	# of terms	# of correct terms	*Acc.* (%)
CRF	M	3674	2476	67.4
	N	4415	3122	70.7
BiLSTM-CRF	M	3145	2448	77.8
	N	4066	3217	79.1
BERT-BiLSTM-CRF	M	3028	2465	81.4
	N	3842	3187	**83.0**

It can be seen from the experimental results that the performance of the BERT-BiLSTM-CRF model is still the best, and the accuracy of its new knowledge points is the highest, reaching more than 80%. At the same time, the accuracy of the three models on the larger N dataset is better than that on the M dataset, and more new terms are identified on the N dataset. These results show that larger online encyclopedia resources are of great significance for the discovery of new knowledge points.

We also noticed that the accuracy of the three models has significantly decreased compared to the evaluation results of the core knowledge points. Through the analysis of the experimental data, we found that the structure of some knowledge point terms was relatively complicated, which can easily leads to errors in identification. For example, the term *"Extended Cartesian Product"* was identified as *"Cartesian Product"*, *"Database Security Control"* was identified as *"Security Control"* and etc.

4.2. Knowledge Point Association Assessment

In the experiment of knowledge point recognition, we found a total of 3234 knowledge points. This paper further conducted experiments on these knowledge points in three aspects: semantic similarity, semantic relevance and *NGD* distance, and then the experimental results were evaluated and analyzed.

In the semantic similarity experiment, we set several sets of typical values for the two parameters of window size n and similarity threshold s and evaluated the results. The *F1* values of the related evaluations are shown in Table 5.

Table 5. Performance (F1) of semantic similarity experiments

F1 (%)	*s=0.6*	*s=0.7*	*s=0.8*	*s=0.9*
n=10	45.8	51.4	62.8	60.7
n=20	53.7	68.4	**72.4**	68.4
n=30	50.6	62.1	65.4	64.5

It can be seen from the experimental results that when the window size is set to 20 and the threshold is set to 0.8, the experiment can achieve the best performance. Meanwhile we conducted further experiments for the algorithms of *PMI* and *NGD*.

In *PMI* experiments, we set the value of the relevance threshold *t* from 1 to 7, and carried out the experimental evaluation. The precision *P*, the recall *R* and *F1* of the relevance results are shown in Figure 5. It can be seen from the experimental results that when $t = 4$, the experiment can obtain the best *F1* value. When $t > 5$, the *R* drops sharply, but the *P* does not increase significantly.

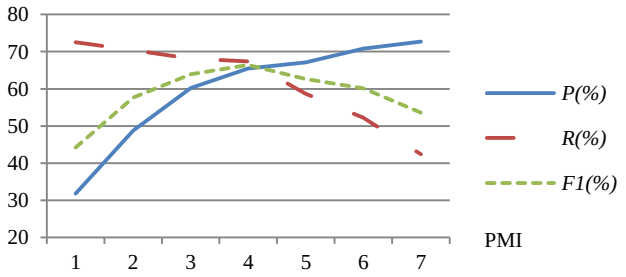

Figure 5. Performance of PMI experiments

In the Google Distance experiments, this paper evaluates the performance with different *NGD* threshold values. The precision *P*, and the recall *R* and *F1* values of the relevant results are shown in Figure 6. The smaller the *NGD* of the two terms, the higher the relevance. It can be seen from the results that when the *NGD* threshold is set to 1, the experiment can obtain the best *F1* value. Compared to *PMI*, the *NGD* experiment is based on massive Internet data, so its *F1* peak is also much higher than the experimental results of *PMI*.

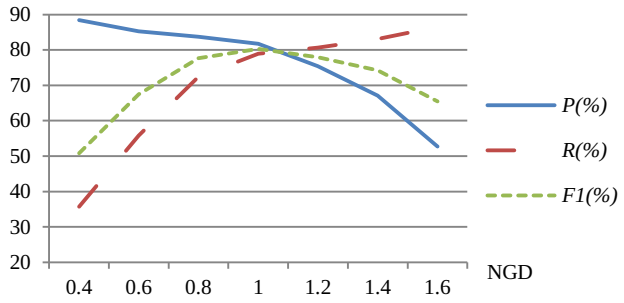

Figure 6. Performance of NGD experiments

4.3. Construction of Subject KG

On the basis of the above algorithm, this paper continuously finds new knowledge points and the connections between knowledge points through continuous iteration. At the same time, in order to control the expansion of the knowledge graph and limit the noise data, this paper assumes that the core knowledge point is the center of subject knowledge. If the distance between the new knowledge point and the core knowledge point is greater than 3, the knowledge point will be considered unrelated to the subject and deleted.

In the construction of knowledge graph for computer science, the window size *n* is set to 20, the similarity threshold s is set to 0.8, the *PMI* threshold is set to 4, and the *NGD* threshold is set to 1. After multiple experimental iterations, the final knowledge graph contains 2207 knowledge point nodes and 3122 edges. With the help of tools such as Gephi and sigma.js, this paper also carried out related data analysis, visualization, and exemplary applications of this subject knowledge graph. The visualization is shown in Figure 7.

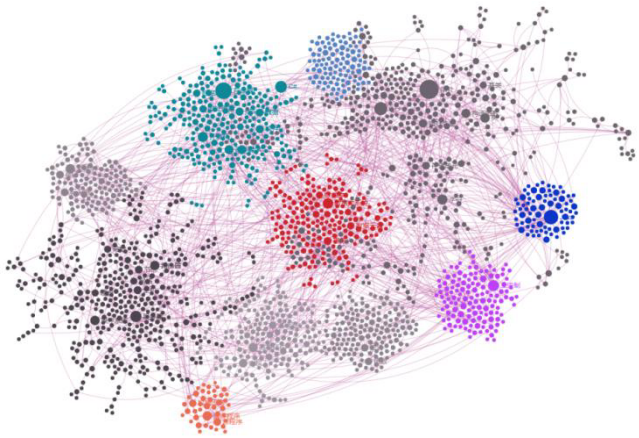

Figure 7. Visualization of KG for computer science

In order to verify the versatility and effectiveness of this method, this paper also conducted experiments on physics subject. A knowledge graph of physics was constructed for three parts: mechanics, heat and optics of the physics subject, which contains 1225 knowledge point nodes and 1722 edges. The experimental results show that our method is applicable for the construction of knowledge graphs of various subjects in education.

5. CONCLUSIONS AND FUTURE WORK

In order to reveal the internal structure of education subject knowledge and serve teaching applications, this paper proposes an automatic construction method of subject knowledge graph based on educational big data. Generally the construction of knowledge graph usually requires a lot of artificial assistance, which limits its update and application. In this paper some strategies and algorithms are used to reduce the reliance on manual annotation resources and achieve better performance. Firstly, this paper comprehensively utilizes teaching resources of education subjects and open online encyclopedia resources. High-quality teaching resources are used to construct the core of the subject knowledge graph. Then the open internet encyclopedia resources are used to expand the core and complete the construction of the entire knowledge graph. Secondly, this paper presents a bootstrapping iterative strategy to construct the subject knowledge graph. At the beginning some core knowledge points are extracted from course syllabuses by regular expressions, and then based on the core knowledge points, BERT-BiLSTM-CRF model is used for knowledge point identification and three different algorithms are used to evaluate the semantic similarity and relevance between knowledge points. The experimental results show that the BERT-BiLSTM-CRF outperformed the baselines and the combination of three algorithms achieved better results. Finally we take computer science and physics as examples to successfully construct the knowledge graph, which show the effectiveness of our method.

In the future, we will continue our research on the evaluation of distance from a knowledge point to specific education subject to determine whether the knowledge point should be added to the knowledge graph. At the same time, we will also explore the application of the subject knowledge graph in higher education, including assessment of student learning effects, intelligent learning navigation, and recommendation of learning resources.

6. ACKNOWLEDGEMENTS

This work was partially supported by National Natural Science Foundation of China (61977032), Guiding Project of Science Research plan of Education Department of Hubei Province (B2018349) and Project of State Language Commission of China (ZDI135-99). We would thank the anonymous reviewers for their hard working. Yong Zhang is the corresponding author.

7. REFERENCES

[1] Bo Xu, Chenhao Xie, Yi Zhang, Yanghua Xiao, Haixun Wang and Wei Wang. Learning Defining Features for Categories. IJCAI 2016

[2] ArnetMiner. https://aminer.org/

[3] Meisam Booshehri, Peter Luksch.An Ontology Enrichment Approach by Using DBpedia. WIMS 2015: 5:1-5:11

[4] Jie Luo, Yifei Wang, Dongchen Jiang. Rule-based hidden relation recognition for large scale knowledge graphs. Pattern Recognition Letters 125: 13-20 (2019)

[5] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, J. Welling. Never-Ending Learning. In Proceedings of the Conference on Artificial Intelligence (AAAI), 2015.

[6] Pieter Heyvaert, Ben De Meester, Anastasia Dimou, Ruben Verborgh. Rule-driven inconsistency resolution for knowledge graph generation rules. Semantic Web 10(6): 1071-1086 (2019)

[7] Pengda Qin, Weiran Xu, William Yang Wang. Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning. ACL 2018

[8] Daojian Zeng, Kang Liu, Yubo Chen, Jun Zhao, Distant Supervision for Relation Extraction via Sparse Representation, EMNLP-2015.

[9] SoYeop Yoo, OkRan Jeong. Automating the expansion of a knowledge graph. Expert Syst. Appl. 141 (2020)

[10] Jiayuan Mao, Yuan Yao, Stefan Heinrich, Tobias Hinz, Cornelius Weber, Stefan Wermter, Zhiyuan Liu, Maosong Sun. Bootstrapping Knowledge Graphs From Images and Text. Front. Neurorobot. 2019

[11] Sarah Kohail. Unsupervised Induction of Domain Dependency Graphs - Extracting, Understanding and Visualizing Domain Knowledge. University of Hamburg, Dissertation, Germany, 2019

[12] R.L. Cilibrasi and P.M.B. Vitanyi. The Google similarity distance. IEEE Trans. Knowledge and Data Engineering, 19:3(2007), 370–383

[13] Mike Mintz, Steven Bills, Rion Snow, Daniel Jurafsky. Distant supervision for relation extraction without labeled data. ACL 2009

[14] https://github.com/macanv/BERT-BiLSTM-CRF-NER