*Journal of Quantum Computing*
DOI: 10.32604/jqc.2022.026785
***Article***

Tech Science Press

# A Top-down Method of Extraction Entity Relationship Triples and Obtaining Annotated Data

**Zhiqiang Hu[1], Zheng Ma[1], Jun Shi[1], Zhipeng Li[1], Xun Shao[1,2], Yangzhao Yang[1,*], Yong Liao[1], Zhenyuan Gao[1] and Jie Zhang[1]**

[1]Shenzhen CyberAray Network Technology Co., Ltd., Shenzhen, 518042, China
[2]School of Regional Innovation and Social Design Engineering, Kitami Institute of Technology, Kitami, 090-8507, Japan
*Corresponding Author: Yangzhao Yang. Email: 13121590@bjtu.edu.cn
Received: 12 January 2022; Accepted: 28 April 2022

**Abstract:** The extraction of entity relationship triples is very important to build a knowledge graph (KG), meanwhile, various entity relationship extraction algorithms are mostly based on data-driven, especially for the current popular deep learning algorithms. Therefore, obtaining a large number of accurate triples is the key to build a good KG as well as train a good entity relationship extraction algorithm. Because of business requirements, this KG's application field is determined and the experts' opinions also must be satisfied. Considering these factors we adopt the top-down method which refers to determining the data schema firstly, then filling the specific data according to the schema. The design of data schema is the top-level design of KG, and determining the data schema according to the characteristics of KG is equivalent to determining the scope of data's collection and the mode of data's organization. This method is generally suitable for the construction of domain KG. This article proposes a fast and efficient method to extract the top-down type KG's triples in social media with the help of structured data in the information box on the right side of the related encyclopedia webpage. At the same time, based on the obtained triples, a data labeling method is proposed to obtain sufficiently high-quality training data, using in various Natural Language Processing (NLP) information extraction algorithms' training.

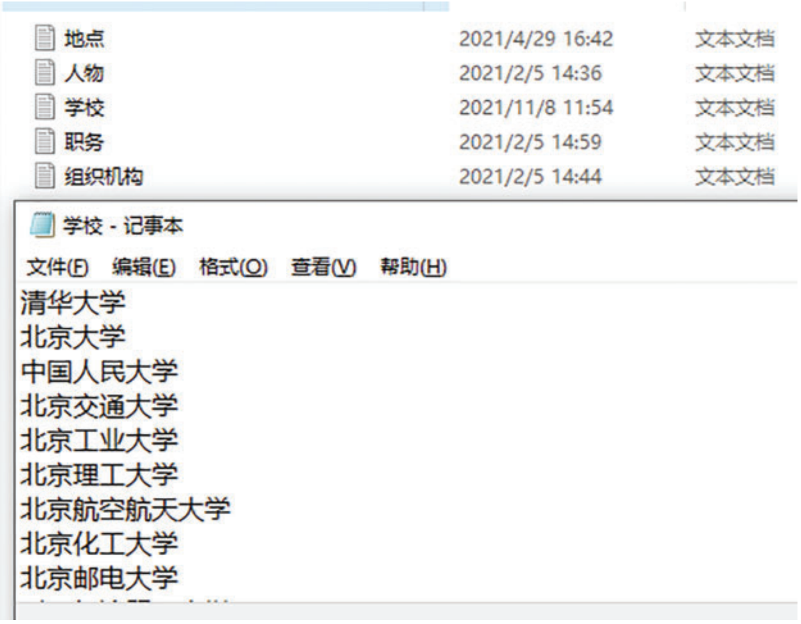
**Keywords:** Entity relationship triples; knowledge graph; top-down; social media; data labeling

## 1 Introduction

In 2012, Google introduced their KG, building on DBpedia and Freebase among other sources, for the optimization of its search engines [1]. Since then, KG has been applied to various fields of NLP. Using the concept of entity relationship triples, KG describes the objective world with structured data. That is, extracting entities, relationships, attributes and other knowledge elements from some publicly structured, semi-structured or unstructured data, such as (entity 1, relationship, entity 2), (entity, attribute, attribute value).

KG construction is generally divided into top-down and bottom-up two ways. Top-down refers to defining entity types and relationship types for KG first, and then adding qualified entities and relationships to the knowledge base. This construction method needs to use some existing structured knowledge bases as its basic knowledge base. Bottom up refers to extracting entities and relationships from some open linked data, selecting those with high confidence to join the knowledge base, and then building the relationship between entities.

According to our business needs, the construction of KG in this article adopts a top-down approach, that is, entities keywords (Fig. 1) and relationships categories (Fig. 2) are defined firstly. Using the entities keywords as the head entities, grab the structured data in the information box on the right side of the related encyclopedia webpage (take Wikipedia as an example, Fig. 3). Each triple is composed of the above header entity and the tail entity, corresponding to each field in the information box, and the field is used to judge the relationship class and tail entity type. Thus, the extraction of top-down KG's triples based on encyclopedia data is completed, and then completes the most critical part of KG's construction.



| | | |
|---|---|---|
| 地点 | 2021/4/29 16:42 | 文本文档 |
| 人物 | 2021/2/5 14:36 | 文本文档 |
| 学校 | 2021/11/8 11:54 | 文本文档 |
| 职务 | 2021/2/5 14:59 | 文本文档 |
| 组织机构 | 2021/2/5 14:44 | 文本文档 |

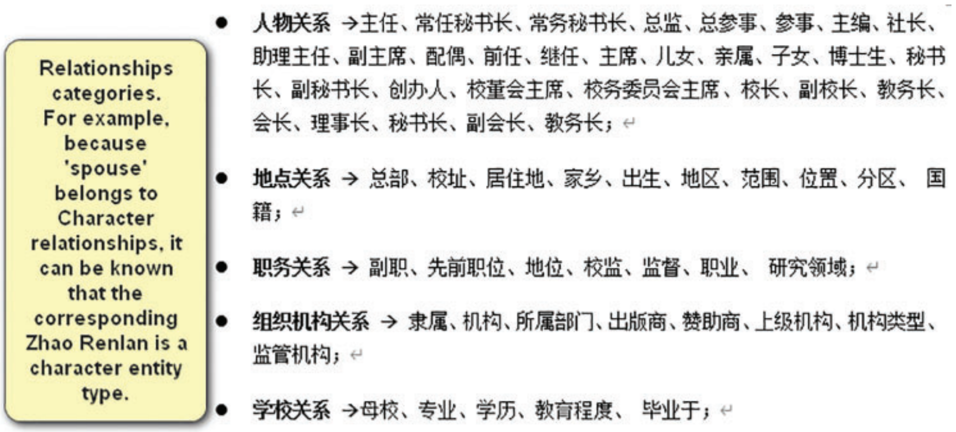
清华大学
北京大学
中国人民大学
北京交通大学
北京工业大学
北京理工大学
北京航空航天大学
北京化工大学
北京邮电大学

**Figure 1:** Some entities keywords

Another work of this paper is to use the triples obtained above, combine with the corresponding sentences as annotation data. Then after a small amount of manual intervention, a large amount of high-quality training data can be obtained, which can be used for the algorithm development of various NLP information extraction tasks.

## 2 Background and Related Works

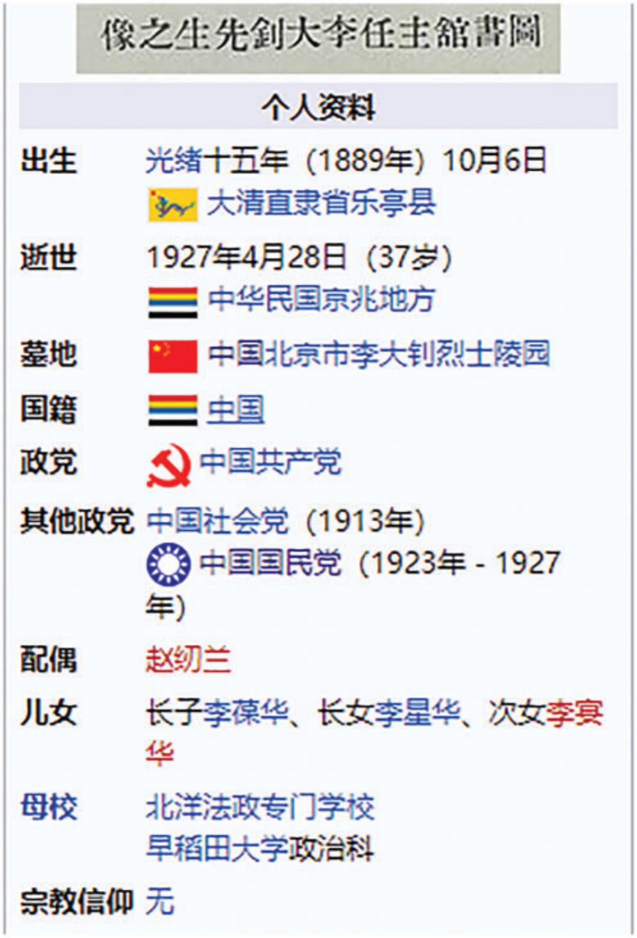
The KG is a data structure composed of entities, relationships and attributes. Through entities, relationships, and attributes, we can effectively organize knowledge that we understand. The construction and application of KG involve technologies such as databases, NLP, and semantic networks.



Relationships categories. For example, because 'spouse' belongs to Character relationships, it can be known that the corresponding Zhao Renlan is a character entity type.

- **人物关系** →主任、常任秘书长、常务秘书长、总监、总参事、参事、主编、社长、助理主任、副主席、配偶、前任、继任、主席、儿女、亲属、子女、博士生、秘书长、副秘书长、创办人、校董会主席、校务委员会主席、校长、副校长、教务长、会长、理事长、秘书长、副会长、教务长；↵
- **地点关系** → 总部、校址、居住地、家乡、出生、地区、范围、位置、分区、 国籍；↵
- **职务关系** → 副职、先前职位、地位、校监、监督、职业、 研究领域；↵
- **组织机构关系** → 隶属、机构、所属部门、出版商、赞助商、上级机构、机构类型、监管机构；↵
- **学校关系** →母校、专业、学历、教育程度、 毕业于；↵

**Figure 2:** Some relationships categories



像之生先釗大李任主館書圖

**个人资料**

| | |
|---|---|
| **出生** | 光绪十五年（1889年）10月6日<br>大清直隶省乐亭县 |
| **逝世** | 1927年4月28日（37岁）<br>中华民国京兆地方 |
| **墓地** | 中国北京市李大钊烈士陵园 |
| **国籍** | 中国 |
| **政党** | 中国共产党 |
| **其他政党** | 中国社会党（1913年）<br>中国国民党（1923年 - 1927年） |
| **配偶** | 赵纫兰 |
| **儿女** | 长子李葆华、长女李星华、次女李实华 |
| **母校** | 北洋法政专门学校<br>早稻田大学政治科 |
| **宗教信仰** | 无 |

**Figure 3:** Information box from Wikipedia

When constructing a KG, depending on whether to determine the data schema (or ontology) before collecting specific data, or collect specific data before determining the data schema, two information extraction methods are produced.

In the current NLP information extraction algorithms, the quality of labeled data often directly determines the effect of the algorithm. At present, the way to obtain labeled data is mainly through manual, which is time-consuming, laborious and expensive.

### *2.1 Bottom-up*

The bottom-up method refers to extracting high-confidence information from the open linked data or extracting information from unstructured text, and then refines the data schema based on the extracted information to complete the KG's construction. Constructing in this way because before the KG constructed, it is not clear about the scope of data collection or how to use the data. Collecting such data form a huge data set, then summarizing the characteristics of the data set form a framework which is called data schema. The public domain KG generally adopts this method, for the public domain KG involves massive information and includes all aspects of knowledge, the effect needs large and comprehensive. At the initial stage of construction, it is difficult to figure out the overall structure of the data, only to summarize and refine the characteristics based on the data set to form a data framework schema. For example, the KG of Google or Baidu belongs to the typical public domain. When using their search tools, users may input a wide range of contents or ask questions in various fields, which requires their KG should cover all kinds of knowledge. In the process of constructing such KG, with the continuous accumulation of data, data knowledge will be classified, then the data schema can be gradually presented.

### *2.2 Top-down*

The top-down method refers to determining the data schema firstly, then filling the specific data according to the schema, and finally forming the KG. The design of data schema is the top-level design of KG, and determining the data schema according to the characteristics of KG is equivalent to determining the scope of data's collection and the mode of data's organization. This method is generally suitable for the construction of domain KG. For an identified industry, data contents and data organization methods are relatively easy to determine. For example, the KG in legal field may be organized in form of legal classifications, legal provisions, legal cases, etc. Another example is the establishment of social media's KG. According to the main representatives of the field, the characters can be classified, the job, organizations, subordinates, friends, and main opinions of the characters can be counted, then the data schema can be designed based on these relationships. Collecting such characters and related data, a social media KG in certain field is formed finally. Generally speaking, the top-down method is applicable to those fields with clear knowledge contents and clear relationships. A typical example is Freebase project which data is mostly obtained from Wikipedia [2].

## 3 Triples Extraction and Annotation Data Acquisition

Taking the construction of social media KG as an example, with the help of expert knowledge, using a top-down method, this article summarizes the extraction of triples based on encyclopedia web pages and the acquisition of labeled data into three steps, including schema construction, triples extraction, labeled data acquisition.

### *3.1 Schema Construction*

In the process of KG construction, upper data pattern is provided from data schema which is the formal description of entity existence and the basis for the extraction of triples. The process of domain schema construction usually includes the following six steps: ontology requirement analysis, investigating reusable ontology, establishing domain core concepts, establishing concept hierarchy, defining classes and creating attributes, ontology evaluation and improvement [3]. According to different fields and different actual needs, the process of schema construction is also different. At present, the more recognized methods of building schema are: Skeleton method [4], TOVE method [5], Seven Step method [6], etc. Using these methods for reference, facing social media, organize and classify the existed encyclopedia structured data, and then combine the relevant knowledge of experts to define the types of entities and relationships we concern. Finally, an example of the construction of the schema RDF graph is completed, as shown in the following Fig. 4.

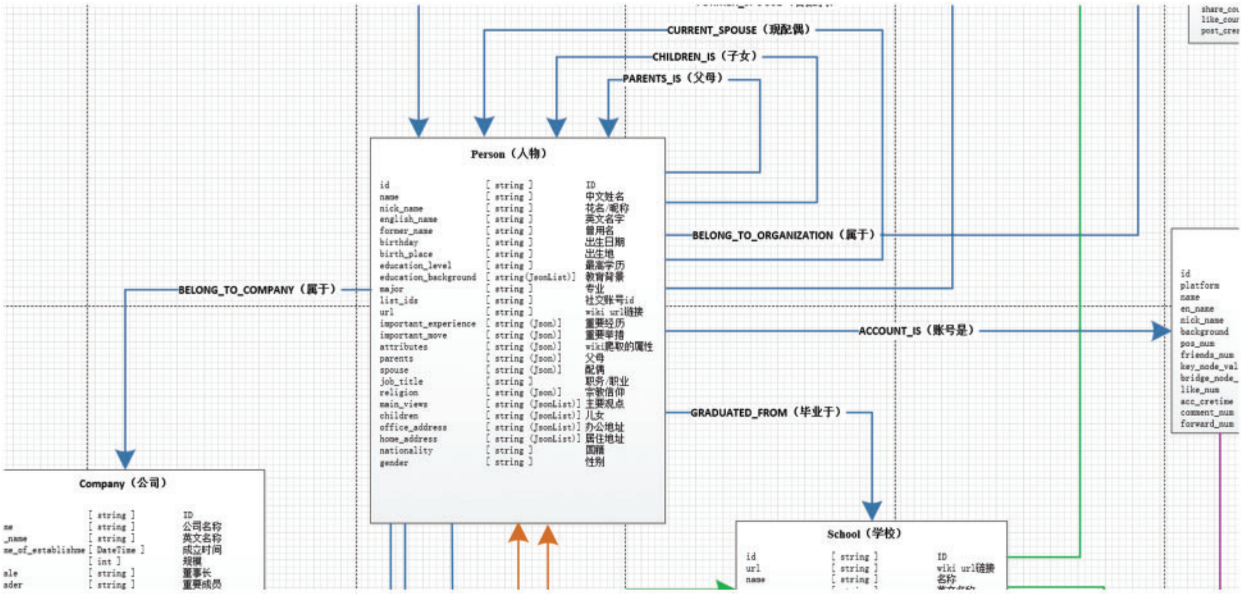

CURRENT_SPOUSE（现配偶）
CHILDREN_IS（子女）
PARENTS_IS（父母）
share_cou
like_cour
post_crea
Person（人物）
id [ string ] ID
name [ string ] 中文姓名
nick_name [ string ] 花名/昵称
english_name [ string ] 英文名字
former_name [ string ] 曾用名
birthday [ string ] 出生日期
birth_place [ string ] 出生地
education_level [ string ] 最高学历
education_background [ string(JsonList)] 教育背景
major [ string ] 专业
list_ids [ string ] 社交账号id
url [ string ] wiki url链接
important_experience [ string (Json)] 重要经历
important_move [ string (Json)] 重要举措
attributes [ string (Json)] wiki爬取的属性
parents [ string (Json)] 父母
spouse [ string (Json)] 配偶
job_title [ string ] 职务/职业
religion [ string (Json)] 宗教信仰
main_views [ string (JsonList)] 主要观点
children [ string (JsonList)] 儿女
office_address [ string (JsonList)] 办公地址
home_address [ string (JsonList)] 居住地址
nationality [ string ] 国籍
gender [ string ] 性别
BELONG_TO_ORGANIZATION（属于）
BELONG_TO_COMPANY（属于）
ACCOUNT_IS（账号是）
GRADUATED_FROM（毕业于）
id
platform
name
en_name
nick_name
background
pos_num
friends_num
key_node_val
bridge_node_
like_num
acc_cretime
comment_num
forward_num
Company（公司）
[ string ] ID
me [ string ] 公司名称
_name [ string ] 英文名称
me_of_establishme [ DateTime ] 成立时间
[ int ] 规模
ale [ string ] 董事长
ader [ string ] 重要成员
School（学校）
id [ string ] ID
url [ string ] wiki url链接
name [ string ] 名称

**Figure 4:** RDF graph of social media schema

### *3.2 Triples Extraction*

The extraction method in this article is divided into the following 3 steps:

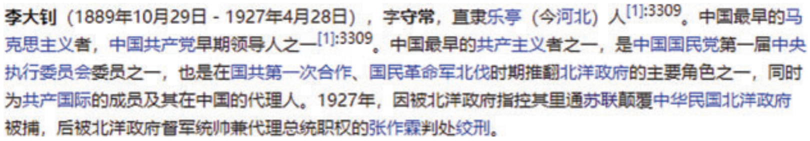
Step one, according to business scenario, combined with expert knowledge, sort out the entities keywords (Fig. 1) and relationships categories (Fig. 2) that need to be crawled. Among them, the entities keywords are the target objects should be crawled, the relationships categories are the fields in the information box on the right side of Wikipedia (Fig. 3), collecting these fields then classify them to determine the relationships and form Entity Relationship Triples.

Step two, crawling relative wiki pages based on the entities keywords in Fig. 2. For example, the person "Li Dazhao" (Fig. 3). Using the keyword "Li Dazhao" to crawl the data in the information box on the right side of the relevant Wikipedia as a header entity with type of person, then traversing the relationship keywords and finding the "spouse" field appears in the information box on the right side of Wikipedia belongs to "Character Relationships", therefore, the type of entity "Zhao Renlan" corresponding to this field is also a person, as the tail entity (Li Dazhao, Spouse, Zhao Renlan).

Similarly, we can get party relationship (Li Dazhao, Political Party, Communist Party of China), education relationship (Li Dazhao, Graduated From, Waseda University), nationality relationship (Li Dazhao, Nationality, Chinese) and so on.

Step three, handling MANY relationship class. The MANY relationship class is generated because the fields in the information box in Fig. 3 may exceed the definition of the relationships categories in Fig. 2. For example, the field "cemetery" in Fig. 3 is not included in the "Location Relationships", but there is still a relationship (Li Dazhao, Cemetery, Beijing Li Dazhao Martyrs Cemetery). For this case, this algorithm is uniformly named MANY class and numerous MANY cases need to be handled manually, which should be classified into the relationships categories in Fig. 2 or discard.

Note that the relationships categories in Fig. 2 are the key to our top-level design. It is not only used to determine whether there is a defined relationship between entities, but also determines the type of tail entity, which is the key to judge entity pairs can be combined to form a triple or not. Relationships categories are sorted manually in advance, and automatic expansion of the program is not supported. Entities keywords are also manually sorted at the beginning, but the program supports automatic expansion. Specifically, if there is a hyperlink (blue content in Fig. 5 below) in the encyclopedia page corresponding to the entities keywords, the content is considered to be an entity (Laoting, Hebei, Marxism . . . ) which is its 1-degree related entity. Collecting these 1-degree entities and adding them to the original entities keywords realizes the automatic expansion of entities classes. Crawling the relevant 1-degree entities and performing the same steps on them to achieve the 1-degree information expansion of entities keywords. After actual testing, a sufficient number of entities can get if we expand to 2-degree in generally.



**李大钊**（1889年10月29日 - 1927年4月28日），字守常，直隶乐亭（今河北）人[1]:3309。中国最早的马克思主义者，中国共产党早期领导人之一[1]:3309。中国最早的共产主义者之一，是中国国民党第一届中央执行委员会委员之一，也是在国共第一次合作、国民革命军北伐时期推翻北洋政府的主要角色之一，同时为共产国际的成员及其在中国的代理人。1927年，因被北洋政府指控其里通苏联颠覆中华民国北洋政府被捕，后被北洋政府督军统帅兼代理总统职权的张作霖判处绞刑。

**Figure 5:** Hyperlink text with some 1-degree related entities

### *3.3 Annotation Data Acquisition*

According to the triple obtained in 3.2, go to Wikipedia text to find a matching sentence. Saving these sentences as training set, which is likely to be obtained because there will be a specific description of the right information box (Fig. 3) in the wiki text. For example, triples (Li Dazhao, Political Party, Communist Party of China) and (Li Dazhao, Native place, Hebei Laoting) corresponding sentence is in Fig. 5. Save the index, type, relationship of the head and tail entities corresponding to each triple into the format we need through program, then complete the triples extraction and annotation data acquisition.

According to this cycle, the rapid, automatic and accurate triples extraction and annotation data acquisition algorithm is realized. 3.2 and 3.3 can be summarized as the following flow chart Fig. 6:

The rest is the triple fusion and triple storage part.

Triple fusion is to fuse the triples extracted from multiple data sources to construct the association relationships between the data, so as to ensure the consistency and accuracy of the data in the KG. Entity alignment is the main challenge in the process of knowledge fusion, which aims to judge whether two or more entities with different information sources are the same entity. For example, "BJTU" and "Beijing Jiaotong University" actually describe the same entity and can be merged.

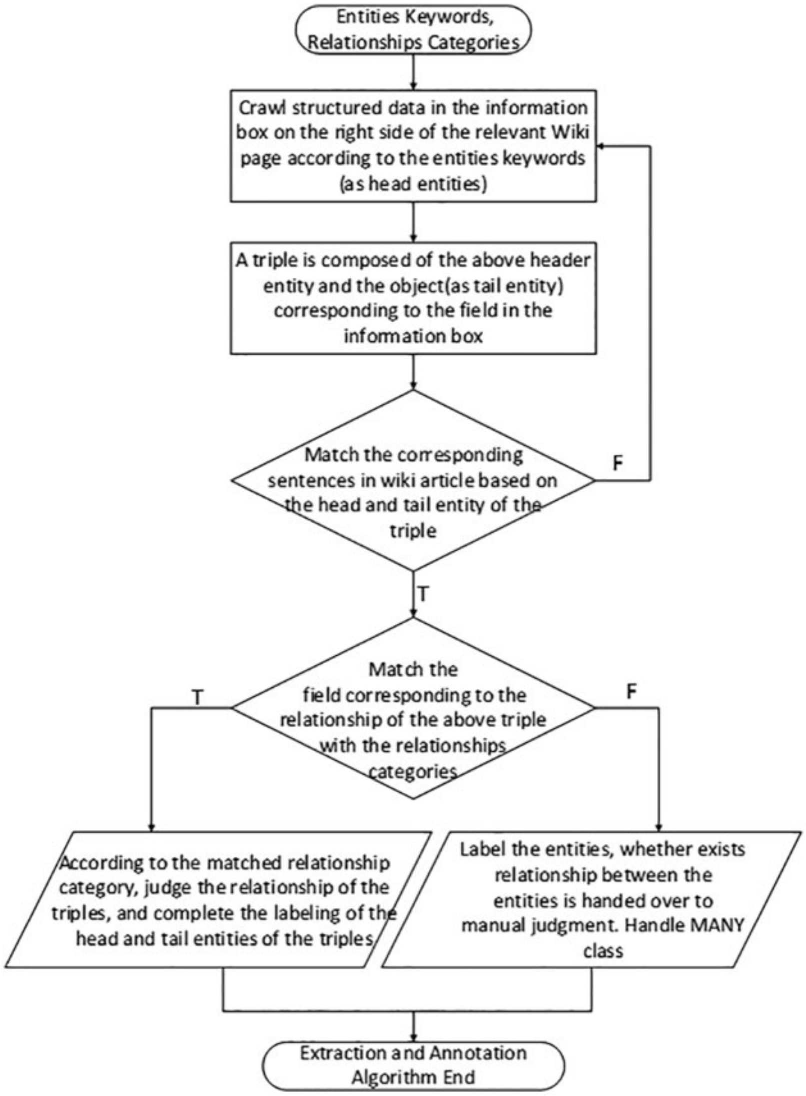

Entities Keywords, Relationships Categories
Crawl structured data in the information box on the right side of the relevant Wiki page according to the entities keywords (as head entities)
A triple is composed of the above header entity and the object(as tail entity) corresponding to the field in the information box
Match the corresponding sentences in wiki article based on the head and tail entity of the triple
F
T
Match the field corresponding to the relationship of the above triple with the relationships categories
T
F
According to the matched relationship category, judge the relationship of the triples, and complete the labeling of the head and tail entities of the triples
Label the entities, whether exists relationship between the entities is handed over to manual judgment. Handle MANY class
Extraction and Annotation Algorithm End

**Figure 6:** Flow chart of triples extraction and data labeling

Triple storage refers to storing the fused triples data in a database to support upper-level applications such as knowledge reasoning and knowledge calculation. Since most of the data stored in the KG is related-intensive data, and the graph database can store this type of data conveniently, graph data has become the mainstream storage method. The graph database uses the “graph data structure” to represent and store data, and realize fast query. It organizes, indexes and stores the relationship between nodes in the form of ‹Key, Value› pairs.

## 4 Result Shows

This paper finally constructed a social media KG containing 599173 entities in 13 categories with 2596416 relationships in 40 categories. The interface is shown in Fig. 7 below. In the figure, the circular nodes represent entities, and the marked edges represent relationships between the entities.

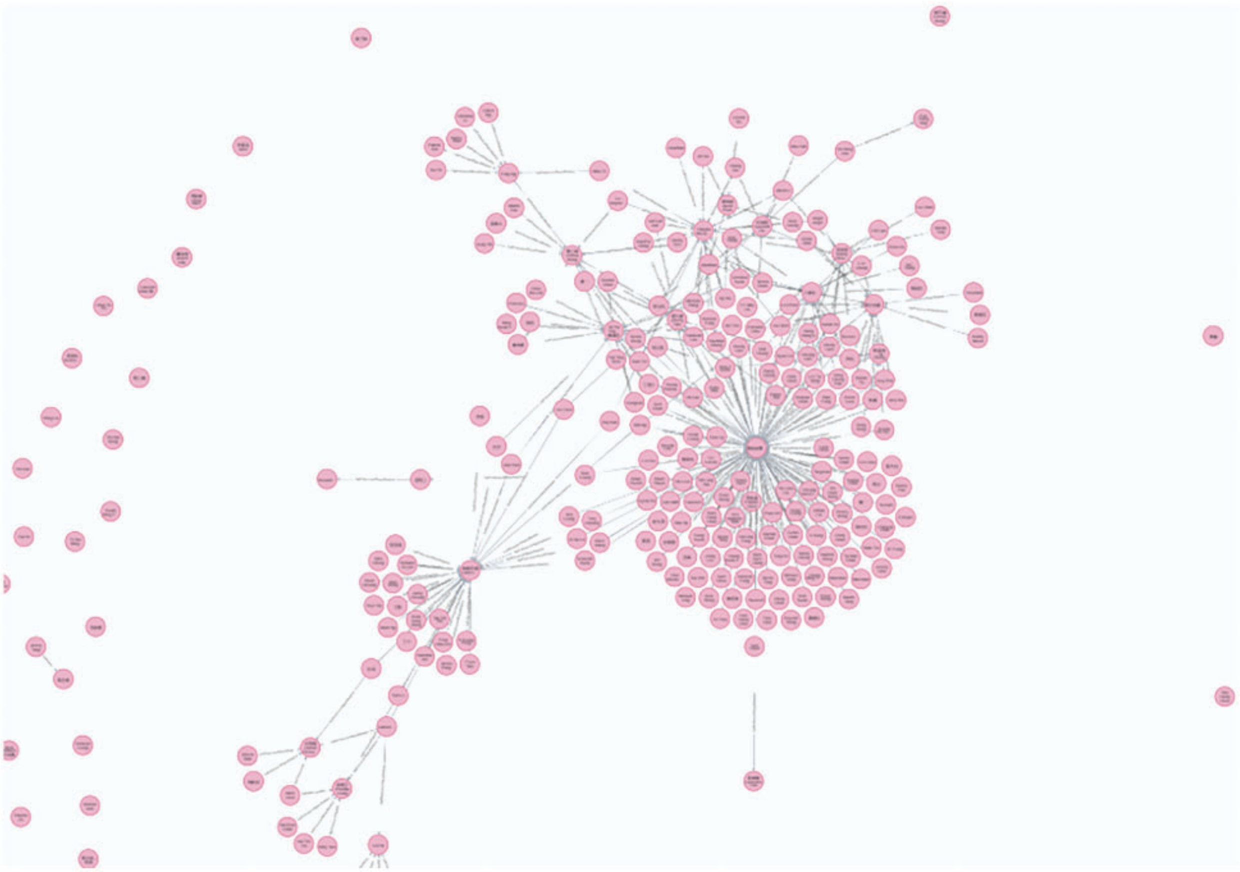
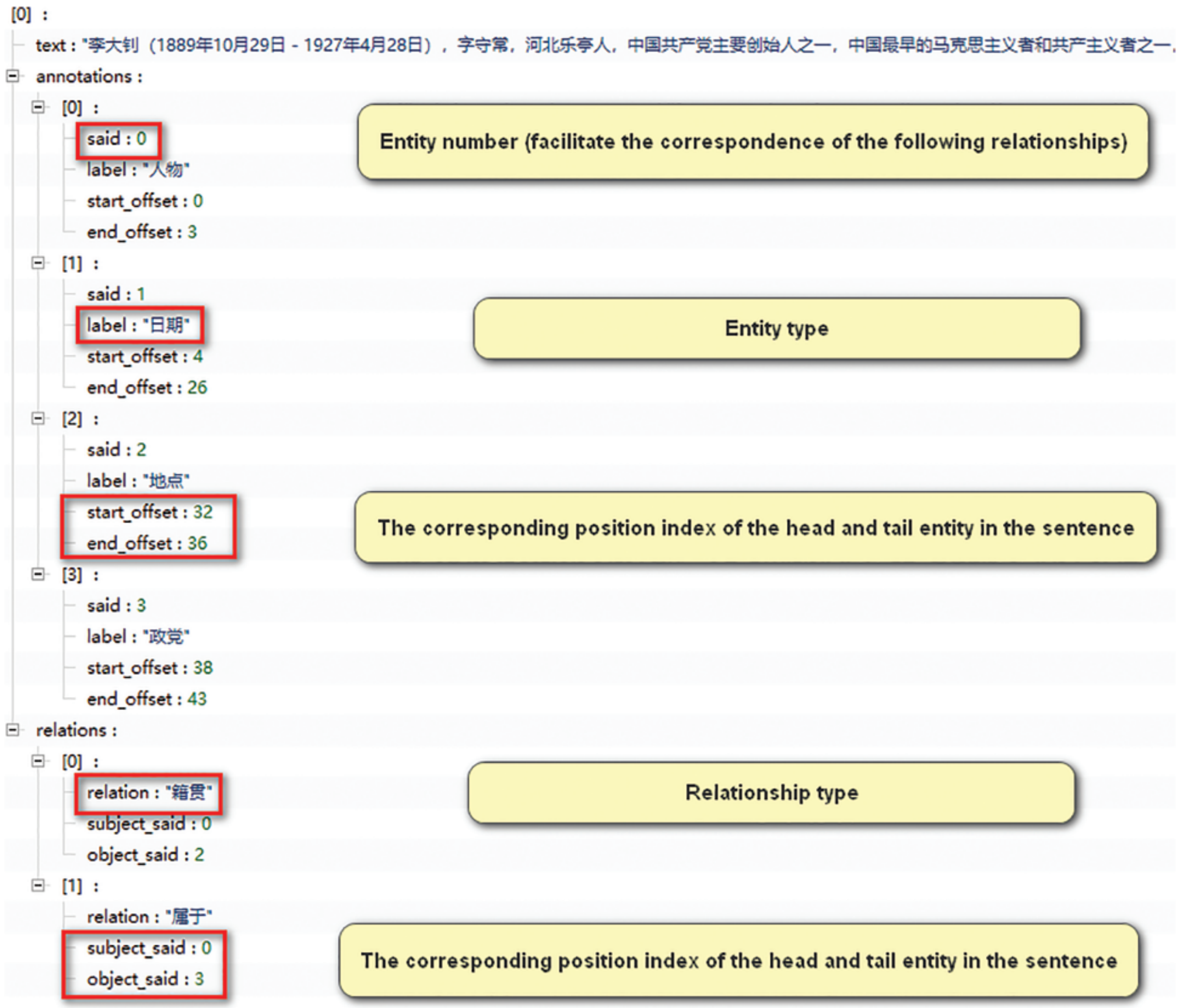


**Figure 7:** Screenshot of social media KG

The format of annotation data acquisition we save as shown in Fig. 8. It should be noted that the “relations” field cannot contain the relationships of all entities in the “annotations” field. For example, the entities in “annotations” also have the following relationship (“said:0”, political party, “said:3”) namely (Li Dazhao, Political Party, Communist Party of China). Even though there is no “Political Party” relationship in the “Relationships Categories” (not defined, Fig. 2), this algorithm will also label entities as much as possible. If the algorithm can determine the relationship between entities, it will be reflected in the “annotations” field, if not, it will be handed over to the next step of manual judgment. This can reduce the manual workload as much as possible and let us focus on judging the relationships between entities. If there is no sentence that completely corresponds to the triples in the entire wiki body, skip it and start the wiki crawling and matching of the next entity keyword.



```
[0] :
 text : "李大钊（1889年10月29日 - 1927年4月28日），字守常，河北乐亭人，中国共产党主要创始人之一，中国最早的马克思主义者和共产主义者之一，
 annotations :
   [0] :
     said : 0
     label : "人物"
     start_offset : 0
     end_offset : 3
   [1] :
     said : 1
     label : "日期"
     start_offset : 4
     end_offset : 26
   [2] :
     said : 2
     label : "地点"
     start_offset : 32
     end_offset : 36
   [3] :
     said : 3
     label : "政党"
     start_offset : 38
     end_offset : 43
 relations :
   [0] :
     relation : "籍贯"
     subject_said : 0
     object_said : 2
   [1] :
     relation : "属于"
     subject_said : 0
     object_said : 3
```

Entity number (facilitate the correspondence of the following relationships)

Entity type

The corresponding position index of the head and tail entity in the sentence

Relationship type

The corresponding position index of the head and tail entity in the sentence

**Figure 8:** The save format of annotation data acquisition

**Table 1:** Some common extraction tools

| Name | Description |
|---|---|
| LTP | Entity extraction |
| Pyhanlp | Entity extraction |
| OpenNLP [7] | Entity extraction |
| BosonNLP | Entity extraction |
| StanfordNER [8] | Entity extraction |
| DeepDive | Relation extraction |
| Wikimeta | Multilingual NER and sense tagging |
| LTP Cloud | Entity extraction, relation extraction, semantic role labeling |

(Continued)

**Table 1:** Continued

| Name | Description |
|---|---|
| BiLSTM + CRF [9] | Entity extraction, relation extraction |
| Bert [10] | Entity extraction, relation extraction, semantic role labeling, POS |

## 5 Conclusion

This paper discusses a method of extraction Entity Relationship Triples in top-down TG construction, and annotation data acquisition. Extracting triples depends on the structured data in the information box on the right side of the related encyclopedia webpage, obtaining annotated data depends on the result of triples extraction above with its corresponding sentences.

This paper is based on its own business and proposes a special method of extracting entity relationship triples. Of course, according to the different tasks and the features of data resources, there are different extraction tools have been released. However, considering the accuracy, we need to use these open-source tools carefully. Some commonly used extraction tools are organized as follows Tab. 1.

**Acknowledgement:** We are thankful to the peoples who have always encouraged and supported us.

**Funding Statement:** The authors received no specific funding for this study.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] A. Singhal, "Introducing the knowledge graph: Things, not strings," *Official Google Blog*, vol. 5, no. 16, pp. 1–10, 2012.

[2] K. Bollacker, R. Cook and P. Tufts, "Freebase: A shared database of structured general human knowledge," in *Proc. AAAI*, Ithaca, New York, USA, vol. 7, pp. 1962–1963, 2007.

[3] W. Zhang and Q. Zhu, "Research on construction methods of domain ontology," *Library and Information*, vol. 155, no. 1, pp. 16–19, 2011.

[4] M. Uschold and M. King, "Towards a methodology for building ontologies," *Edinburgh: Artificial Intelligence Applications Institute*, vol. 1, no. 1, pp. 1–13, 1995.

[5] M. S. Fox, "The tove project towards a common-sense model of the enterprise," in *Proc. Int. Conf. on Industrial, Engineering and other Applications of Applied Intelligent Systems*, Berlin, Springer, 1992.

[6] W. Li, J. Han and P. Jian, "CMAR: Accurate and efficient classification based on multiple class-association rules," in *Proc. of 2001 IEEE Int. Conf. on Data Mining*, Omaha, NE, USA, pp. 369–376, 2001.

[7] V. Khadilkar, M. Kantarcioglu, B. Thuraisingham and P. Castagna, "Jena-HBase: A distributed, scalable and efficient RDF triple store," in *Proc. ISWC-PD*, Boston, USA. pp. 85–88, 2012.

[8] D. Rinser, D. Lange and F. Naumann, "Cross-lingual entity matching and infobox alignment in wikipedia," *Information Systems*, vol. 38, no. 6, pp. 887–907, 2013.

[9] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, "Neural architectures for named entity recognition," *Neural architectures for named entity recognition*, arXiv Prepr. arXiv1603.01360, 2016.

[10] J. Li, A. Sun, J. Han and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020.