ISSN: 2664-9640

Knowledge Extraction From Knowledge Graph: Research Based On COVID-19

Haoxing Wang *, Jiashu Liang, Yu Ma, Anquan Ren

School of Software, Yunnan University, Kun-ming, 650504, China

* Corresponding Author

Abstract

The outbreak of the Novel Coronavirus Pneumonia epidemic has brought a huge impact on the world and caused a global public health crisis. Therefore, it is essentialto accelerate the research on COVID-19. To this end, we build a scientific knowledgegraph of COVID-19 based on more than 20,000 papers which are from the latest early stage research and peer-reviewed research of Elsevier. In the process of constructing the COVID-19 scientific knowledge graph, we use an effective rule-based extraction methodto extract the title, author, research institution, publication time, keywords, abstract, and other key information of these papers. At the same time, we propose a novel joint extraction algorithm of entities and relations to extract the triple information in the abstract. Besides, we use the weight of topic words to divide the nouns in the abstract into each topic, followed by topic modeling using the LDA for mining in the abstract. Finally, we conduct data analysis based on the 2020 topic classification results. Our data analysis reflects the current progress of scientific research on COVID-19 to a certain extent, providing certain reference significance for future research.

Keywords

Global public health crisis, COVID-19, knowledge graph, LDA topic modelling, rule-based extraction method, pandemic, intelligent data analysis.

1. Introduction

In 2020, with the outbreak of Novel Coronavirus Pneumonia in many regions and countries around the world [1], a large number of relevant researchers have been making contributions to devote in COVID-19 research. The main clinical manifestations of the Novel Coronavirus Pneumonia are fever, dry cough, fatigue, etc. In severe cases, dyspnea may occur, with a certain lethal rate [2]. Therefore, accelerating research on COVID-19 is essential. While most of the relevant research results have been published in various academic journals in the form of professional literature, and the key knowledge covered in it has not been systematically combed, so it is not convenient for researchers to use existing knowledge to explore new discoveries quickly and make active contributions to the epidemic prevention work about Novel Coronavirus Pneumonia. With the development of big data and artificial intelligence (AI), data mining technology has been applied to various fields and has achieved certain success, such as discovery of new materials [3–6], virus tracing and propagation analysis [7–9], text mining [10–13], and image identification [14–20]. As a branch of AI, knowledge graph (KG) is a graph of data that is used to accumulate and convey knowledge of the real world. Its nodes represent entities of interest and edges represent the relations between these entities [21]. Entities can be real-world objects and abstract concepts, and relations represent connections between the entities [22]. The KG is used to identify, discover, and infer the complex relationship between things and concepts from the data. Its applications in the fields involve semantic search [23], intelligent question answering [24], and language understanding [25].

ISSN: 2664-9640

In this paper, we use data mining technology to construct a COVID-19 scientific knowledge graph (CSKG). This KG systematically summarizes the previous related research. Our work is based on more than 20,000 articles of the latest early research and peer-reviewed research on Coronavirus and COVID-19 provided by Elsevier1 , using rulebased extraction methods [26] to extract the paper title, author, research institution, publication time, keywords, published journal, reference, abstract, and other key information. Next, the extracted information is converted into a representation of a triple (head entity, relation, tail entity) ((h,r, t) for short) [27], in order to construct the CSKG. In addition, we perform data analysis on the entities in the CSKG. Through this data analysis, we found the following potential knowledge in the CSKG: (1) The focus of research on COVID-19 in 2020; (2) Research hotspots and evolution path of coronavirus. Our work can provide a great help for relevant researchers. They can use our CSKG for the following applications: (1) Retrieve related papers from the CSKG; (2) Explore research hotspots from it; (3) Combine the methods and conclusions of coronavirus research; (4) Discover the cooperative relationship between authors and institutions. The contributions in this paper are summarized as follows:

- We exploit a rule-based extraction method to construct a CSKG.
- We propose a joint extraction algorithm for entities and relations based on bootstrap ideas. Our algorithm does not need data annotation and greatly reduceslabor costs.
- We propose a method based on the LDA topic model and normalize the weight of topic words to divide each word into the generated topics.
- We use data analysis to mine Novel Coronavirus Pneumonia related knowledge and regularity from the CSKG.

The rest of this article is organized as follows. In Section 2, we introduce the work related to the application of AI in COVID-19. Section 3 focuses on the proposed algorithm designed in this work. The experiments and results analysis are introduced in Section 4. Finally, Section 5 summarizes this article and proposes the future research directions.

2. Related work

In this section, a brief review of the application of AI and data mining technology in COVID-19, is presented. At present, the application of AI technology in COVID-19 is mainly used in three aspects: (1) Modeling based on data-driven epidemic spread [28–34]. For example, Barenya Bikash Hazarika and Deepak Gupta in [33] used WCRVFL combined with RVFL and one-dimensional discrete wavelet transform to predict the spread of the COVID-19 in the five worst-hit countries. This model can achieve 60-day prediction. Sayantari Ghosh and Saumik Bhattacharya in [34] proposed a method based on probabilistic cellular automata and optimized it with an optimization algorithm, and then modeled and estimated daily live cases, the total number of infections, and the total number of deaths. (2) Detect positive COVID-19 cases using Chest CT images [35–40]. For example, Tao Zhou et al. proposed an integrated deep learning model for detecting COVID-19 from CT images in [40]. They trained three deep convolutional neural network models including AlexNet, GoogleNet, and ResNet for image features extraction. Finally, Softmax is used as the classification algorithm for the fully connected layer. The ensemble classifier is obtained through relatively majority voting. The results show that the overall classification performance of the ensemble model is better than that of the component classifiers. (3) Text data mining based on COVID-19 [41–45]. For example, Rosario Catelli et al. developed two multilingual deep learning systems to translate the Italian text of COVID-19 for low-resource language scenarios in [44]. The work in [45] combined multiple models such as BERT, Unigram, Bigram, SVM, Random Forest, etc. to determine the internal emotions of news, opinions, and other information on Twitter, and mining topics between April 2020 and August 2020 for analysis. Our proposed work is mainly

ISSN: 2664-9640

used in the third aspect ”text data mining based on COVID-19”. Most existing text mining methods based on COVID-19 use a language model. Although language model is good in predicting ability, it lacks the form of data. In response to these problems, we have constructed a CSKG. As a form of information, KG has high-value semantic information and is rich in interpretability. Data mining on KG can acquire more information. At present, there is a lot of work on the construction of the COVID-19 knowledge graph. Recently, some knowledge graphs of COVID-19 are released by OpenKG. For example, Wang et al.[2] released the knowledge graph of COVID-19 Encyclopedias. They take viruses and bacteria as the main part and expand thecontents related to treatment and disease. Through the encyclopedia knowledge of these concepts, a knowledge graph of COVID-19 is formed, which could be applied to semantic retrieval, intelligent question answering, and document intelligent recommendation. The knowledge graph of COVID-19 science formed by Chen et al.[3] belongs to Zhejiang University, fused a large number of Novel Coronavirus Pneumonia research information by extracting the basic information of Novel Coronavirus Pneumonia, antiviral drugs, and genetic relationships into a KG. The knowledge graph of COVID-19 epidemiology is released by Li et al.4 . At present, the KG includes the basic concepts of epidemiology, epidemiological investigation, and so on. The knowledge graph of COVID-19 events is released by millet artificial intelligence laboratory5 They collect a series of new events related to the Novel Coronavirus Pneumonia and annotate semantics to some of the contents of the news. It supports forward and backward indexing of Novel Coronavirus Pneumonia events in time, and provides enumeration of event development context.

The above relevant KGs have made positive contributions to most fields of COVID-19, most of their data sources come from encyclopedia, virus databases, and news. However, the content of the COVID-19 knowledge graph regarding academic research fields does not yet exist. In order to fill this gap, we have made the CSKG, and conducted data mining and analysis on the CSKG, whose details are given in Section 3.

3. The Proposed Methodology

This section mainly focuses on the process and methods of constructing the CSKG and performing data analysis on it. The general KG construction process is to extract knowledge from structured, semi-structured, and unstructured data. Knowledge extraction also includes entity extraction and relation extraction, which can be used to extract the entities and relations that we are interested in. After knowledge extraction, the next step is entity alignment that includes entity disambiguation and co-reference disambiguation. The entity disambiguation is to resolve the problem of ambiguity that is caused by entities with the same name. The co-reference disambiguation resolves multiple reference items corresponding to the same entity object. Entity alignment is mainly used to eliminate contradictions and ambiguities in knowledge graph. The final step is knowledge storage and knowledge inference. Knowledge storage is used to store knowledge graph in the database, which is convenient for management, query, and visualization. Knowledge inference is a method to excavate the undiscovered knowledge from the existing knowledge graph so as to enrich and expand the existing knowledge graph.

In the process of constructing the CSKG, we first use a rule-based method to match a text string with a designed regular expression to extract interesting entities. Using rule-based extraction methods, we extract key information such as the title, author, research institution, keywords, abstract, and publication time, and convert the extracted information into a triple representation. Since the information in the abstract is of great value, we conduct further entity extraction and relation extraction in the abstract. At present, most models of entity extraction and relation extraction require labeled data,but annotating data often requires a lot of cost.

ISSN: 2664-9640

Therefore, we propose a bootstrap-based algorithm to implement a joint extraction for entity and relation, which does not require manual annotation of data.

After constructing the CSKG, we conduct data mining on the information in it, hoping to find information that is of interest to relevant researchers. Based on this information, an overview of the current research status of COVID-19 is given. We propose a data mining method based on the Latent Dirichlet Allocation (LDA) topic model [46], which can classify the words in the abstract by topic. Through this method, we can classify the valuable words in the abstract into a certain category of topics to obtain the current focus on coronavirus research and related important information. At present, the number of scientific research papers related to COVID-19 is increasing. This information is hidden in intricate papers and is difficult to detect using human efforts. We expect to mine this information through data mining related algorithms, so that we can obtain COVID-19 related information more quickly and accurately.

In Section 3.1, we introduce the related concepts and methods of entity and relation extraction. In Section 3.2, we introduce a bootstrap-based algorithm to implement a joint extraction for entity and relation. The algorithm is inspired by snowball [47]. In Section 3.3, we introduce the method used in the process of data mining, which is based on the LDA algorithm.

3.1. Emtity extraction and relation extraction

Entity extraction, also called Named Entity Recognition (NER) [48], refers to extracting entities with specific meaning from text. For example, person name, organizationname, location, time, etc. At present, entity extraction methods are mainly divided intothree categories: (1) Rule-based extraction methods; (2) Extraction methods based onstatistical models; (3) Extraction methods based on deep learning.

The rule-based extraction method is mainly used to manually formulate extracted rule templates for specific fields. The definition of rules usually consists of demonstrative words, central words, punctuation marks, keywords, etc. Then, the established extraction rules are pattern-matched with the text string to extract interesting entities.

The extraction methods based on statistical models mainly use fully labeled andpartially labeled corpus for model training. There are four common models that are mainly used in this field: Hidden Markov Model, Conditional Markov Model, Maximum Entropy Model, and Conditional Random Field Model. The extraction methods based on statistical models transform the problem of NER into a problem of sequence labeling.

The extraction methods based on deep learning mainly take the vector representationof the word in the text as input. After combining the context information of the word, the new vector representation of the word is considered as output. Finally, the labeling result is obtained as an output through the CRF model.

We adopt a rule-based extraction method to construct a CSKG. Because the other two methods need to annotate data, and the extraction templates for the title, author, research institution, keywords, abstract, publication time, and other information of the paper can be easily produced by hand.

After entity extraction, the text corpus obtains a series of discrete named entities without semantic relations among them. Therefore, the semantic relation between entities needs to be extracted from the text corpus. At present, there are three main types of relation extraction methods: (1) Rule-based relation extraction methods, which are mainly manual extraction templates that match entities with specific relation in the text; (2) Relation extraction methods based on supervised learning. These methods aim to train a supervised model on a large amount of labeled data to extract relation; (3) The relation extraction methods based on weakly supervised learning. Compared with supervised learning, these methods require only a small amount of labeled data to train the model, which mainly include distant supervision method and bootstrapping method.

ISSN: 2664-9640

In the process of constructing the CSKG, the relations are defined according to the inherent relations between the entities. For example, if the title, institutions, and authors have been extracted from a certain paper, the relations involved are author_is and association_is, represented as a triple (title, author_is, author) and (author, association_is, institution).

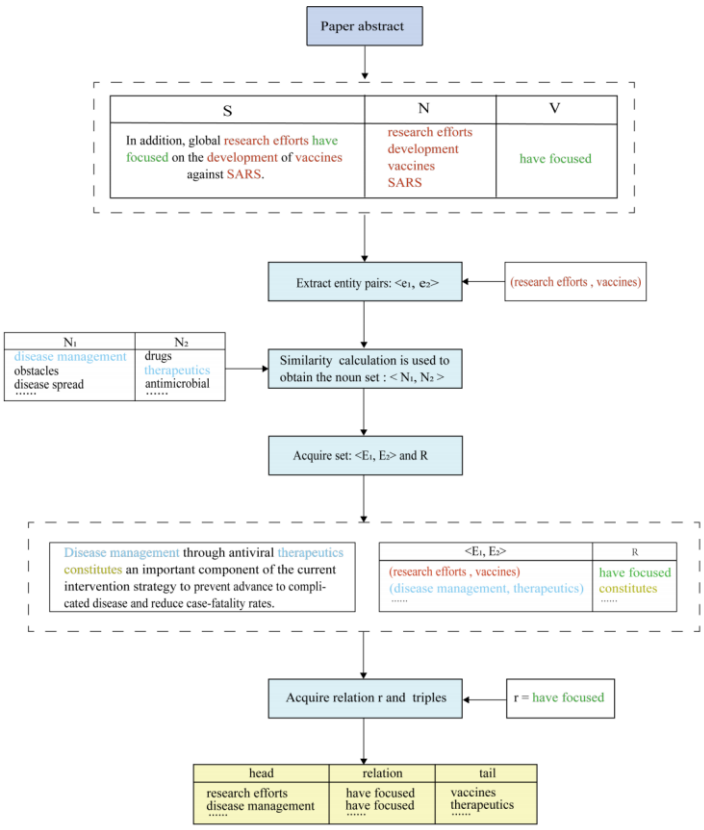

Figure 1: Framework diagram of entity and relationship joint extraction algorithm

Algorithm 1 Joint extraction algorithm for entity and relation

Input: The preprocessed dataset (S, N, V)
Output: A series of triples (head, relation, tail)

Calculate the word vectors of words in the corpus
for data in dataset do
  $(e_1, e_2) \leftarrow$sample(data)
  for each data in dataset do
    $(N_1, N_2) \leftarrow similarity(e_i, N) >$ threshold and Top K
  end for
  for n_1 in N_1 and n_2 in N_2 do
    $(E_1, E_2) \leftarrow n_1$ in S_i and n_2 in S_i
    $R = \{v_i\} \leftarrow v_i$ in S_i
  end for
  for r_i in R do
    Similarity (r_i, r_j in $R_{j!=i}$)
  end for
  Acquire $r \leftarrow$ number (similarity (r, R) > threshold) is most
  (head, relation, tail) $\leftarrow$ Acquire triples (E_1, r, E_2)
end for
return (head, relation, tail)

ISSN: 2664-9640

3.2. Joint extraction algorithm for entity and relation

This section introduces the joint extraction algorithm of entity and relation based on the idea of bootstrap. The algorithm framework is shown in Figure 1, and the pseudo code is shown in Algorithm 1. The algorithm has a basic assumption such that if there is an entity set H that is semantically similar to h in the triple (h,r, t), and an entity set T that is semantically similar to t, then the relation between the entities in H and T is also r. Compared with the traditional extraction algorithm, the advantages of our designed algorithm are as follows: (1) No need to annotate data, which greatly reduces labor costs; (2) Compared with the traditional bootstrap algorithm, there is no need to set an iteration seed, so the extracted information is more complete; (3) Ability to extract potential relations between multiple words in a sentence.

The algorithm steps are as follows:

Step 1: Input the preprocessed dataset. Each piece of data includes a sentence in the abstract, the nouns and verbs in the sentence. A detail introduction can be obtained from Figure 1.

Step 2: Use word2vec to calculate the word vectors of words in the corpus.

$$\mathrm{sim}(A,B) = \frac{A \odot B}{|A| * |B|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

where, A and B represent the word vectors.

Step 3: Iterate the dataset and use permutation and combination to obtain candidate entity pairs < e1, e2 > from the noun set in each data. Cosine similarity (refer to Eq. (1)) calculation is used to obtain the noun set < N1, N2 > that ranks top K with semantic similarity to each candidate entity and has a threshold greater than α. Select the pair of nouns that appear in a sentence at the same time through the pairwise combination between N1 and N2. These noun pairs form the candidate entity set < E1, E2 >, and the verb set of the sentence is added to the candidate relation set R. Use Eq. (1) to calculate the semantic similarity between the verbs in the candidate relation set R. Count the number of similarity thresholds greater than β between the current verb and other verbs. The verb with the largest number and greater than the set threshold is regarded as the relation of the candidate entity set < E1, E2 >. In this way, a series of triples (E1,r, E2) with relation r is obtained.

Then, we analyze the time complexity of Algorithm 1. Assuming that the size of dataset is x, the average size of noun set in each data is y, and y is much smaller than x. The time complexity of Algorithm 1 mainly depends on the operations including iterating dataset, taking out the noun pairs in each data, and calculating the similarity between a single noun in the noun pair and all nouns. The required time complexity of the above operations is $O(x)$, $O(y^2)$, $O(xy^{2})$, so the time complexity of Algorithm 1 is x^2y^4.

3.3. Data mining method

The data mining method (shown in Figure 2) we used is based on the LDA topicmodel, which can extract topic words from the paper and can identify hidden topicsin large-scale papers. For example, consider a document d = [After the outbreak ofthe Novel Coronavirus Pneumonia epidemic in Wuhan, the Chinese government quicklyissued an order to lock down the city and mobilized medical resources from the wholecountry to Wuhan for support, effectively curbing the further spread of the epidemic]. Now Topic 1 will correspond to [Wuhan, Epidemic] and topic 2 will correspond to [Government, Closing the City, Mobilizing Medical Resources]. It can be seen that the LDA topic model can further explore the potential meaning of a document or find the between documents.

Given a set of documents d = {d1, d2, d3, · · · , dn}, since the LDA topic model does not consider the sequence order, the document can be regarded as a corpus of a large number of word list W = [w1,w2, · · · ,wm]. Assume that the topic set generated by the LDA topic model is t = {t1, t2, · · · , tk}.

ISSN: 2664-9640

First, the LDA topic model assumes that the prior distribution of document topics follows the Dirichlet distribution, for any document di, its topic distribution θi is shown in Eq. (2).

$$\theta_i \sim \mathrm{Dir}\ (\alpha) \tag{2}$$

Its probability density is shown in Eq. (3).

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \tag{3}$$

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \tag{4}$$

where, α is the hyperparameter of the Dirichlet distribution, which is a k-dimensional vector. k is the number of topics assumed in advance, and Γ(x) is the Gamma function.

Secondly, generate the topic zi j of the j-th word of the document di from the Multinomial distribution θi of topic:

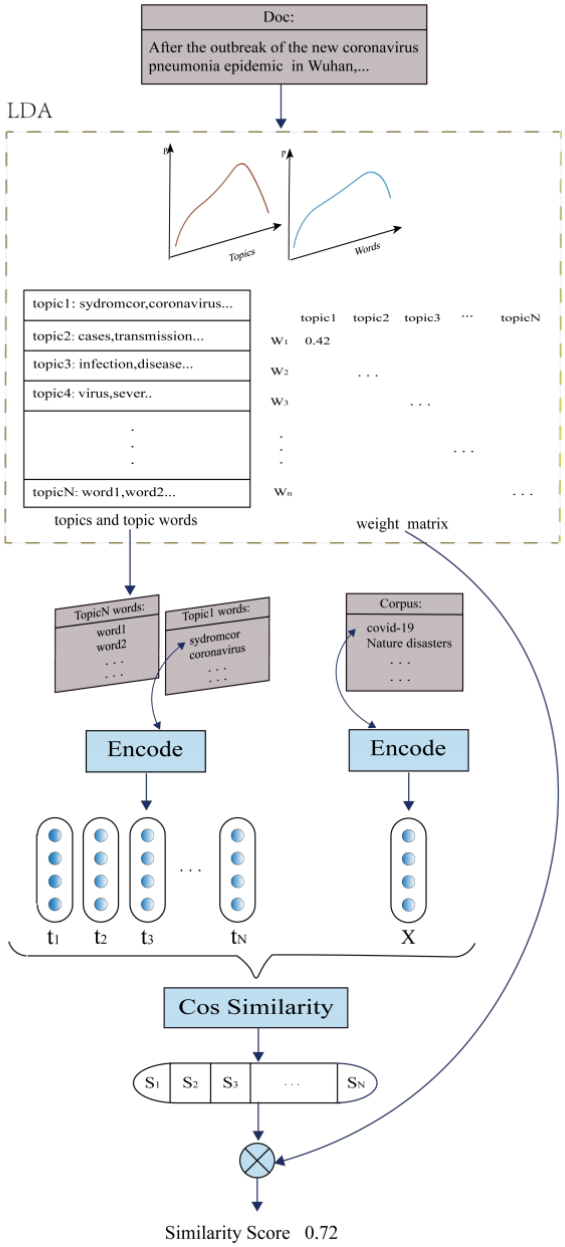

Figure 2: Framework of data mining. Doc stores the abstracts of all papers and Corpus stores the nounsin the abstract. First, the LDA topic model extracts N topics from Doc. Each topic is composed oftopic words and their weights. Then, select a noun X from Corpus and all the words

ISSN: 2664-9640

of a topic encodeto prepare for the next step. Finally, input the vector and topic words weight into the scoring function to calculate the similarity score between noun X and Topic.

$$z_{ij} = \text{Multi}\ (\theta_i) \tag{5}$$

$$\text{Multi}\ (x) = p\{x_1 = m_1, x_2 = m_2, \cdots, x_n = m_n\} = \frac{N!}{m_1!m_2!\cdots m_n!} p_1^{m_1} p_2^{m_2} \cdots p_n^{m_n} \tag{6}$$

where, Multi(θ_i) is the Multinomial distribution.

At the same time, the LDA topic model assumes that the prior distribution of topic words in a topic also follows the Dirichlet distribution, for any topic zi j, the topic word distribution is shown in Eq. (7).

$$\varphi_{z_{ij}} \sim \text{Dir}\ (\beta) \tag{7}$$

Its probability density is shown in Eq. (8).

$$p(\varphi \mid \beta) = \frac{\Gamma(\sum_{i=1}^{v} \beta_i)}{\prod_{i=1}^{v} \Gamma(\beta_i)} \varphi_1^{\beta_1 - 1} \cdots \varphi_v^{\beta_v - 1} \tag{8}$$

where, β is the hyperparameter of the Dirichlet distribution, which is a v-dimensional vector, and v is the number of all the words in the corpus.

Finally, the topic word $\delta_{z_{ij}}$ of topic z_{ij} is generated by sampling from the Multinomial distribution $\varphi_{z_{ij}}$ of topic words as shown in Eq. (9).

$$\delta_{z_{ij}} = \text{Multi}\ (\varphi_{z_{ij}}) \tag{9}$$

We use Eq. (10) to normalize $p(\varphi \mid \beta)$ as a weight to calculate the relation between the topic word and the noun in the paper.

$$\omega_{ij} = \frac{p_{ij}}{\Sigma_1^n p_{ij}} \tag{10}$$

where, ω_{ij} represents the weight of the j-th word of the i-th topic, and p_{ij} represents the probability of the j-th topic word of the i -th topic.

We use Eq. (11) to calculate the similarity between the nouns in the abstract and each topic to classify each noun into possible topics according to the similarity score.

$$\text{score}\ (x, \text{topic}_i) = \Sigma_1^n \omega_{ij} \text{sim}\ (f(x), f(\text{topic}_{ij})) \tag{11}$$

where, $f\ (\cdot)$ represents the embedding function. Here we use word2vec as the embedding function to calculate the word vector of the word. Then we iteratively calculate the similarity between each noun and the j-th topic word of the i-th topic and multiply it with the weight of the topic word. The purpose of this operation is to weaken the error caused by the probability calculation and quantify the relation between words and topics in a better way. At the end, the final result of the calculation is added to obtain the similarity score between the word and the topic. When the score is greater than the set threshold, the word is classified under the topic.

Table 1: The data scale of the COVID-19 scientific knowledge graph

item	Number	Item	Number
title	16818	aggregationType	4
author	73919	reference	505631
Affiliation	50500	revisedDate	3453
abstract	10188	receivedDate	4793
keyword	45895	acceptedDate	4825
doi	17962	origLoadDate	4797
publication	1551	number of	745058
availOnlineDate	4722	entities	13
		type of relations	

ISSN: 2664-9640

4. Experiments Results and Discussion

Based on more than 20,000 articles on Coronavirus and COVID-19 provided by Elsevier, we construct the CSKG according to the method discussed in Section 3 and use data analysis on the entities in the CSKG. Our work aims to summarize related research in this field, assist researchers in decision-making analysis, and make positive contributions to the global epidemic prevention works. It is worth noting that we not do algorithm comparison work. For example, Algorithm 1 has a basic assumption such that if there is an entity set H that is semantically similar to h in the triple (h,r,t), and an entity set T that is semantically similar to t, then the relation between the entities in H and T is also r. In this way, our algorithm extracts a large number of triples, but some of them do not exist in the dataset, so the accuracy of the algorithm cannot be quantified. In addition, the algorithms in recent years are based on supervised learning or semi-supervised learning improved by deep learning models. Algorithm 1 is similar to unsupervised learning using unlabeled data, so the nature of algorithms is different.

4.1. Construction of knowledge graph

We use a rule-based method to extract key information such as the title, author, research institution, keywords, abstract, publication time, and other information of the paper by pattern matching with text strings using hand-made rule templates. The authors are further refined into first, second, and corresponding authors, and the relationship between authors and institutions is expanded. We convert the extracted information into a triple representation. The data scale of the CSKG is shown in Table 1.

Since the information in the abstract of the paper is of great value, we use the proposed entity and relation joint extraction model to further extract the entity and relation of the abstract. The number of entities extracted from the abstract is 5300, and the relation is of 2177 types.

After that, we build a query system based on the CSKG. The main functions implemented by the CSKG interface are: (1) Search for cooperating authors and institutions based on author name; (2) Search for cooperating authors and institutions based on institution name; (3) Query article information based on paper title; (4) Search for related papers based on keywords. Our CSKG and query system have been published on OpenKG[6]. We believe that the CSKG can accelerate COVID-19 related research andmake certain contributions to the fight against the disease.

Table 2: Topics generated by the LDA topic model

Topic	Topic Words	Number of Nouns
Topic1	[syndrome, coronavirus, respiratory, disease, severe, acute, health]	282
Topic2	[cases, transmission, associated, severe, coronavirus, risk, respiratory, pandemic]	252
Topic3	[infection, human, disease, coronavirus, patients, respiratory, severe]	797
Topic4	[virus, severe, disease, patients, coronavirus]	253
Topic 5	[clinical, patients, respiratory, coronavirus, pandemic, disease, acute]	523
Topic6	[host, health, public, disease, virus, spread]	3414
Topic7	[pollen, factor, current, spread, correlation, review, study]	2150

ISSN: 2664-9640

Table 3: Representative words under each topic: columns 1 to 4 in the table represent the unique nounsunder each topic, and the last column shows the nouns under all topics.

Topic 1	Topic 2	Topic 3	Topic 4	Multiple words
public health global	diarrhea cases	s1 subunit	novel viruses mimic	fatality rate
areas	respiratory viruses	sars-cov s protein	drug discovery	covid-19 pneumonia
controlling emergence environment events epidemiology preventive trends challenges . . .	lymphopenia diarrhea asthmatic bronchopneumonia rhinovirus infection infectious peritonitis lethal passive transfer . . .	chain reaction rna genome hcov-229e ic50 hela cells balb/c mice balb dna vaccine . . .	Biotechnology laboratory tests molecular methods pathogen detection high-throughput screening new technologies horizontal viral vaccines . . .	mers coronavirus ill severe pneumonia globally emergency department travel h7n9 . . .

4.2. Data analysis for the period 2020 to 2021

The abstract contains important information about the paper. Therefore, using the LDA topic model, we excavated the hidden topic information in the paper abstract and classified each noun in the abstract under each topic information, so as to discover the focus of each period. In the process of classifying nouns, we make full use of the topic word probability as the similarity weight to make more scientific classification effect.

In the period of 2020 to 2021, we generate 7 effective topics through the LDA topic model (as shown in Table 2). At the same time, the information-rich nouns in the abstract are also classified into various topics by weighted cosine similarity. As shown in Table 3, the representative nouns classified under the topic are displayed.

The words under topic 1 such as “global public health”, “environmental events”, “regions”, and “prevention” indicate the topics are related to the regionality of the epidemic. It can be seen that some studies take the region as the entry point, focusing on globalprevention and the impact of regional differences on the spread and outbreak of the disease. Researchers can discover research hotspots related to the spread of the disease by these terms. We observe that there are some words about respiratory diseases and their symptoms in the classification results, in the topic 2 including “diarrhea”, “asthma”, “human rhinovirus infection” etc. Therefore, topic 2 is the clinical manifestations of Novel Coronavirus Pneumonia or other respiratory diseases. Researchers can look for clinical manifestations of Novel Coronavirus Pneumonia or other respiratory diseases that they have not paid attention to yet. Topic 3 contains a large number of biochemical related words, such as “S1 subunit”, “sars-cov s protein”,

ISSN: 2664-9640

"RNA genome", "balb/c mouse" etc. Therefore, topic 3 is highly related to the research on the biochemical theory of the Novel Coronavirus Pneumonia. Topic 3 shows the current focus of research on the mechanism of the Novel Coronavirus Pneumonia. Researchers can discover current research methods, research hotspots or find research inspiration by observing the nouns under topic 3. According to the words under topic 4, such as "drug discovery", "high-throughput screening", "new virus simulation" and so on, we know that topic 4 shows the research methodology of pathology and the relevant situation of Novel Coronavirus Pneumonia vaccine research. Researchers can use the information under this topic to explore current research methods for viruses, vaccines, and related drugs.

We extract the words belonging to most topics (N >= 5) as shown in Table 3. Through observation, we can find that the word "covid-19 pneumonia" appears in most of the topics, indicating that all the studies have been carried out around Novel Coronavirus Pneumonia after the outbreak of Novel Coronavirus Pneumonia in 2020. "fatality rate", "severe pneumonia", "globally", and other words indicate that the Novel Coronavirus Pneumonia has a larger impact and a certain mortality rate. "dengue" and "mers coronavirus ill" have similarities with the COVID-19, indicating that some researchers want to research the COVID-19 by studying similar viruses.

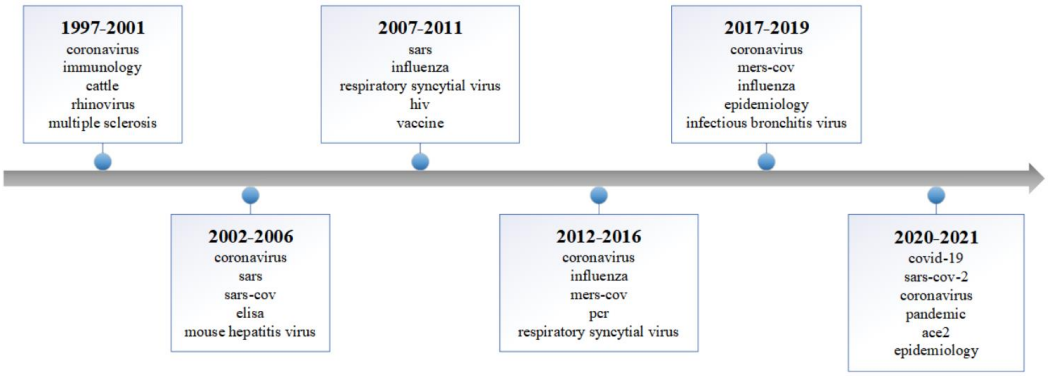

Figure 4: The evolution of research hotspots.

4.3. The evolution of research hotspots

Through data analysis of keywords (as shown in Figure 4), our work highlights the evolution of research hotspots over the time. The figure truly reflects the relationship between the disease outbreak time in the real world and research hotspots, which illustrates the effectiveness of our works.

From 1997 to 2001, it indicates the initial stage of research with diverse research content, the research focus on coronavirus, and immunology. From 2002 to 2011, the emergence of SARS and its high fatality rate attracted a lot of attention from researchers. The main research focus in this stage was on SARS. From 2012 to 2019, with the emergence of Middle East Respiratory Syndrome, another research boom in this field was raised. The main research hotspot at this stage was Middle East Respiratory Syndrome. From 2020 to 2021, with the global outbreak of Novel Coronavirus Pneumonia, this field has received unprecedented attention. All mankind is helping each other to overcome difficulties together.

4.4. Multi-path information mining

The triples extracted by the joint mining model contain a lot of path information, which is displayed in the form of a graph structure in the KG. We use the characteristics of the graph structure to mine the multi-path information between entities, hoping to obtain more information. The method used (as shown in Figure 5) is as follows: suppose there are triples A(E1, R1, E2), B(E2, R2, E3), and C(E1, R3, E3), then A, B, and C forms a closed triangle, we can propose an inference rule R1+R2≈R3 from the triangle. Through multi-path information mining,

ISSN: 2664-9640

we can obtain richer interpretability from E1$\xrightarrow{R3}$E3. For example, (Mr. Wang, BornIn, Wuhan), (Wuhan, IsProvinceOf, China), (Mr.Wang, NationalityIs, China), we can get a path reasoning rule, BornIn +IsProvinceOf ≈ NationalityIs. Moreover, further explanation is that Mr. Wang was born in Wuhan, and Wuhan belongs to China, so Mr. Wang's nationality is China. From the path of China $\xrightarrow{start}$outbreak $\xrightarrow{cause}$death and China $\xrightarrow{has\ shown}$death, we know that COVID-19 broke out in China, the outbreak of COVID-19 caused death, and COVID-19 caused death in China. Therefore, we can get a relation formula of start + cause ≈ has shown. Similarly, we learn from the path of China $\xrightarrow{lived}$Wuhan$\xrightarrow{arose}$ outbreak and China $\xrightarrow{have\ been\ discovered}$ outbreak that Wuhan is a city in China, Wuhan arose COVID-19 and COVID-19 has been found in China. So we get another relation formula: live + arose ≈ have been discovered. In this way, we can get an explanation of the triples. For example, the cause of deaths in China is the outbreak of COVID-19, the outbreak of COVID-19 in China is mainly concentrated in Wuhan.

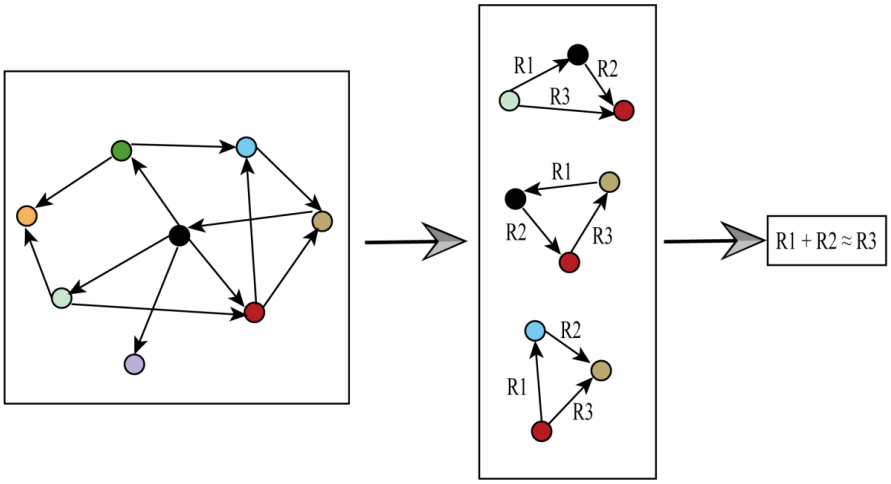

Figure 5: The multi-path information mining.

5. Conclusion

In order to assist scientists for researching the COVID-19 and accelerate the global epidemic prevention work, we proposed a rule-based extraction method to construct a CSKG. Researchers can use this KG to sort out research works about COVID-19. At the same time, we proposed a joint extraction model to extract the triple information in the paper's abstract. The model does not need to set seed, so that the extracted information is more complete. Finally, we used the LDA topic model to conduct topic mining, and divided the information-rich nouns in the abstract into various topics, hoping to extract some information that is difficult to obtain through the traditional way of reading papers. Our data analysis realistically reflects the real-world scenario, provides a good reference and guidance for scientific researchers in this field, and contributes to the epidemic prevention work about the Novel Coronavirus Pneumonia.

Despite the good contributions of our work to COVID-19, the proposed method has some limitations. For example, the time complexity of Algorithm 1 is relatively high and the calculation on a large dataset may incur high time cost. Focusing on these aspects can be possible future studies in this area of research. In future, we will consider extracting higher-level knowledge from the paper dataset. For example: (1) Dig out the host and transmission vector of the related virus in anticipation of discovering the connection between certain viral gene fragments and the host and transmission vector. (2) Extract the research methods and conclusions of related viruses, and make a systematic review of the research methodology for

ISSN: 2664-9640

virus. (3) Extract the treatment plan of known viruses, and give drug recommendations for some viruses that have no treatment drugs.

Acknowledgements

The authors are grateful to Elsevier, the text data used is provided by Elsevier free of charge. We would like to acknowledge the financial support provided by the Natural Science Foundation of China (NSFC) under Grant No. 61876166, No. 61663046.

References

[1] Nishiura, H., Jung, S., Linton, N.M., Kinoshita, R., Yang, Y., Hayashi, K., Kobayashi, T., Yuan, B., Akhmetzhanov, A.R.: The extent of transmissionof novel coronavirus in wuhan, china, 2020. Journal of Clinical Medicine (2020)

[2] Huang, C., Wang, Y., Li, X., Ren, L., Cao, B.: Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. The Lancet 395(10223) (2020)

[3] Raccuglia, P., Elbert, K.C., Adler, P.D.F., Falk, C., Norquist, A.J.: Machine-learning-assisted materials discovery using failed experiments. Nature 533(7601), 73–76 (2016)

[4] Baishya, S.S., B¨auml, B.: Robust material classification with a tactile skin using deep learning. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 8–15 (2016).https://doi.org/10.1109/IROS.2016.7758088.

[5] Xie, T., Grossman, J.C.: Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys. Rev. Lett. 120, 145301 (2018). https://doi.org /10.1103/PhysRevLett. 120.145301.

[6] Sch¨utt, K.T., Sauceda, H.E., Kindermans, P.J., Tkatchenko, A., M¨uller, K.-R.: Schnet – a deep learning architecture for molecules and materials. Journal of Chemical Physics 148(24), 241722 (2018).

[7] Babayan, S.A., Orton, R.J., Streicker, D.G.: Predicting reservoir hosts and arthropod vectors from evolutionary signatures in rna virus genomes. Science 362(6414), 577–580 (2018)

[8] Direkoglu, C., Sah, M.: Worldwide and regional forecasting of coronavirus (covid-19) spread using a deep learning model. medRxiv (2020)

[9] Lee, H., Jo, J., Lee, Y.O., Nuriye, K.Z., Abelmann, L.: Deep learning analysis of binding behavior of virus displayed peptides to aunps. In: FdezRiverola, F., Mohamad, M.S., Rocha, M., De Paz, J.F., Gonz´alez, P. (eds.) Practical Applications of Computational Biology and Bioinformatics, 12th International Conference, pp. 97–104. Springer, Cham (2019)

[10] Liu, B., Guo, W., Niu, D., Luo, J., Wang, C., Wen, Z., Xu, Y.: Giant: Scalable creation of a web-scale ontology. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. SIGMOD '20, pp. 393–409. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3318464.3386145. https://doi.org/10.1145/3318464.3386145

[11] Ananiadou, S., Pyysalo, S., Tsujii, J., Kell, D.B.: Event extraction for systems biology by text mining the literature. Trends in biotechnology 28(7), 381–390 (2010)

[12] Krallinger, M., Leitner, F., Valencia, A.: Analysis of biological processes and diseases using text mining approaches. In: Bioinformatics Methods in Clinical Research, pp. 341–382. Springer, ??? (2010)

[13] Wang, Z., Shang, J., Liu, L., Lu, L., Liu, J., Han, J.: Crossweigh: Training named entity tagger from imperfect annotations. EMNLP/IJCNLP (1), 5153–5162 (2019)

[14] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. international conference on learning representations (2015)

[15] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 558–567 (2019). https://doi.org/10.1109/CVPR.2019.00065

ISSN: 2664-9640

[16] Xu, B., Ding, S., Zhang, Y.: Image classification model based on GAT. Journal of Physics: Conference Series 1570, 012082 (2020). https://doi. org/10.1088/1742-6596/1570/1/012082

[17] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. arXiv preprintarXiv:1704.06904 (2017)

[18] Yang, Y., Hu, Y.Y., Zhang, X., Wang, S.: Two-stage selective ensemble of cnn via deep tree training for medical image classification. IEEE Transactions on Cybernetics PP(99) (2021)

[19] Yang, Y., Guo, J., Ye, Q., Xia, Y., Yang, P., Ullah, A., Muhammad, K.: Aweighted multi-feature transfer learning framework for intelligent medical decision making. Applied Soft Computing 105, 107242 (2021)

[20] Yang, Y., Guo, J., Wang, P., Wang, Y., Yu, M., Wang, X., Yang, P., Sun, L.: Reservoir hosts prediction for covid-19 by hybrid transfer learning model. Journal of biomedical informatics 117, 103736 (2021)

[21] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J., Kirrane, S., Neumaier, S., Polleres, A., et al.: Knowledge graphs. CoRR abs/2003.02320 (2020) (2003)

[22] Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S.: A survey on knowledge graphs: Representation, acquisition and applications. arXiv preprintarXiv:2002.00388 (2020)

[23] Wang, T., Gu, H., Wu, Z., Gao, J.: Multi-source knowledge integration based on machine learning algorithms for domain ontology. Neural Computing and Applications 32(1), 235–245 (2020)

[24] Shah, S., Mishra, A., Yadati, N., Talukdar, P.P.: Kvqa: Knowledge-aware visual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8876–8884 (2019)

[25] El-Kahky, A., Liu, X., Sarikaya, R., Tur, G., Hakkani-Tur, D., Heck, L.: Extending domain coverage of language understanding systems via intent transfer between domains using knowledge graphs and search query click logs. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4067–4071 (2014). IEEE

[26] Str¨otgen, J., Gertz, M.: Heideltime: High quality rule-based extraction and normalization of temporal expressions. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 321–324 (2010)

[27] Yao, L., Mao, C., Luo, Y.: Kg-bert: Bert for knowledge graph completion. arXiv preprint arXiv: 1909.03193 (2019)

[28] Mojjada, R.K., Yadav, A., Prabhu, A., Natarajan, Y.: Machine learning models for covid-19 future forecasting. Materials Today: Proceedings (2020)

[29] Zeroual, A., Harrou, F., Dairi, A., Sun, Y.: Deep learning methods for forecasting covid-19 time-series data: A comparative study. Chaos, Solitons& Fractals 140, 110121 (2020)

[30] Satu, M., Howlader, K.C., Islam, S.M.S., et al.: Machine learningbased approaches for forecasting covid-19 cases in bangladesh. Machine Learning-Based Approaches for Forecasting COVID-19 Cases in Bangladesh (May 30, 2020) (2020)

[31] Rostami-Tabar, B., Rendon-Sanchez, J.F.: Forecasting covid-19 daily cases using phone call data. Applied Soft Computing 100, 106932 (2021). https://doi.org/10.1016/j.asoc.2020.106932

[32] Hernandez-Matamoros, A., Fujita, H., Hayashi, T., Perez-Meana, H.: Forecasting of covid19 per regions using arima models and polynomial functions. Applied Soft Computing 96, 106610 (2020). https://doi.org/10.1016/j.asoc.2020.106610

[33] Hazarika, B.B., Gupta, D.: Modelling and forecasting of covid-19 spread using wavelet-coupled random vector functional link networks. Applied Soft Computing 96, 106626 (2020). https://doi.org/10.1016/j.asoc.2020. 106626

[34] Ghosh, S., Bhattacharya, S.: A data-driven understanding of covid-19 dynamics using sequential genetic algorithm based probabilistic cellular automata. Applied Soft Computing 96, 106692 (2020). https://doi.org/10.1016/j.asoc.2020.106692

[35] Marques, G., Agarwal, D., de la Torre D´ıez, I.: Automated medical diagnosis of covid-19 through efficientnet convolutional neural network. Applied Soft Computing 96, 106691 (2020). https://doi.org/10.1016/j.asoc.2020. 106691

ISSN: 2664-9640

[36] Aslan, M.F., Unlersen, M.F., Sabanci, K., Durdu, A.: Cnn-based transfer learning–bilstm network: A novel approach for covid-19 infection detection. Applied Soft Computing 98, 106912 (2021). https://doi.org/10. 1016/j.asoc.2020.106912

[37] Yousri, D., Abd Elaziz, M., Abualigah, L., Oliva, D., Al-qaness, M.A.A., Ewees, A.A.: Covid-19 x-ray images classification based on enhanced fractional-order cuckoo search optimizer using heavy-tailed distributions. Applied Soft Computing 101, 107052 (2021). https://doi.org/10.1016/j. asoc.2020.107052

[38] Nour, M., C¨omert, Z., Polat, K.: A novel medical diagnosis model for covid-19 infection detection based on deep features and bayesian optimization. Applied Soft Computing 97, 106580 (2020). https://doi.org/10. 1016/j.asoc.2020.106580

[39] Gupta, A., Anjum, Gupta, S., Katarya, R.: Instacovnet-19: A deep learning classification model for the detection of covid-19 patients using chestx-ray. Applied Soft Computing, 106859 (2020). https://doi.org/10.1016/j.asoc.2020.106859

[40] Zhou, T., Lu, H., Yang, Z., Qiu, S., Huo, B., Dong, Y.: The ensemble deep learning model for novel covid-19 on ct images. Applied Soft Computing 98, 106885 (2021). https://doi.org/ 10.1016/j. asoc.2020.106885

[41] Prastyo, P.H., Sumi, A.S., Dian, A.W., Permanasari, A.E.: Tweets responding to the indonesian government's handling of covid-19: Sentiment analysis using svm with normalized poly kernel. Journal of Information Systems Engineering and Business Intelligence 6(2), 112–122 (2020)

[42] Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., Zhu, T.: Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. PloS one 15(9), 0239441 (2020)

[43] Drias, H.H., Drias, Y.: Mining twitter data on covid-19 for sentiment analysis and frequent patterns discovery. medRxiv (2020)

[44] Catelli, R., Gargiulo, F., Casola, V., De Pietro, G., Fujita, H., Esposito, M.: Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. Applied Soft Computing 97, 106779 (2020). https://doi.org/10.1016/j.asoc.2020.106779

[45] Garcia, K., Berton, L.: Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. Applied Soft Computing 101, 107057 (2021). https://doi.org/10.1016/j. asoc.2020.107057

[46] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)

[47] Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. Proceedings of the Fifth ACM Conference on Digital Libraries (DL) (2000). https://doi.org/ 10.1145/336597.336644

[48] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)