Soft Computing (2021) 25:8337–8355
https://doi.org/10.1007/s00500-021-05756-8

METHODOLOGIES AND APPLICATION

Recent trends in knowledge graphs: theory and practice

Sanju Tiwari[1] · Fatima N. Al-Aswadi[2] · Devottam Gaurav[3]

Accepted: 16 March 2021 / Published online: 16 April 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
With the extensive growth of data that has been joined with the thriving development of the Internet in this century, finding or getting valuable information and knowledge from these huge noisy data became harder. The Concept of Knowledge Graph (KG) is one of the concepts that has come into the public view as a result of this development. In addition, with that thriving development especially in the last two decades, the need to process and extract valuable information in a more efficient way is increased. KG presents a common framework for knowledge representation, based on the analysis and extraction of entities and relationships. Techniques for KG construction can extract information from either structured, unstructured or even semi-structured data sources, and finally organize the information into knowledge, represented in a graph. This paper presents a characterization of different types of KGs along with their construction approaches. It reviews the existing academia, industry and expert KG systems and discusses in detail about the features of it. A systematic review methodology has been followed to conduct the review. Several databases (Scopus, GS, WoS) and journals (SWJ, Applied Ontology, JWS) are analysed to collect the relevant study and filtered by using inclusion and exclusion criteria. This review includes the state-of-the-art, literature review, characterization of KGs, and the knowledge extraction techniques of KGs. In addition, this paper overviews the current KG applications, problems, and challenges as well as discuss the perspective of future research. The main aim of this paper is to analyse all existing KGs with their features, techniques, applications, problems, and challenges. To the best of our knowledge, such a characterization table among these most commonly used KGs has not been presented earlier.

Keywords Knowledge graphs · Knowledge extraction · Learning techniques · Reasoning

1 Introduction

KGs have become one of the most effective and efficient knowledge integration techniques that have been most accepted in both academic and industry circles recently these years. KGs are specially organized to present entities from any area and the relations between these entities. For sharing linked data globally, several KGs such as Freebase Bollacker et al. (2008), DBpedia Lehmann et al. (2015), YAGO Suchanek et al. (2007), NELL Carlson et al. (2010), Wikidata Vrandecic (2012), Google Knowledge Graph Singhal (2012) and Google Knowledge Vault Dong et al. (2014) are published. These KGs are most widely used for associating data sets on the Linked Open Data (LOD) cloud, entity linking or web search and question answering applications. These KGs are also categorized based on size such as large or vast networks and large scale knowledge base (Google Vault, Google Knowledge Graph) while YAGO is termed as ontology but it referred both Knowledge Graph as well as Knowledge Base. Based on different definition and information, knowledge graphs are expressed as cleaned knowledge base that is ontology population (creating instances). But still they are suffering for storage cost and richer computation and become quite precious for several applications such as spam detection, machine learning applications and can be availed by knowledge graphs Zhao et al. (2019).

There are several knowledge representation models that are used to store KG such as Resource Description Frame-

✉ Sanju Tiwari
tiwarisanju18@ieee.org

Fatima N. Al-Aswadi
fatima7aswadi@gmail.com

Devottam Gaurav
gauravpurusho@gmail.com

[1] Universidad Autonoma de Tamaulipas, Ciudad Victoria, Mexico

[2] Faculty of Computer Sciences and Engineering, Hodeidah University, Hodeidah, Yemen

[3] Dept of Computer Science and Engineering, Indian Institute of Technology, Delhi, India

work (RDF), Resource Description Framework Schema (RDFS), JavaScript Object Notation (JSON) and so on. RDF is the most used model to describe the metadata and semantics of information. This model is efficient to organize the huge amount of knowledge and presenting entities as a triple store. A triple is a form of "Subject-Predicate-Object" or "Subject-Attribute-Value". These triplets collectively weave a network of knowledge to form a Knowledge Graph (KG).

As KGs are quite expressive to represent the structured data in a symbolic form (subject, object, predicate). They are not only databases or applications but also a core element of several operations and a practice of understanding the information and decision making to improve the efficiency. KG is considered as a cornerstone of several information systems that need to access the structured knowledge in a specific domain and an independent domain. As Google made announcement in 2012[1] for KG to represent the general world knowledge and paid a serious attention. KG follows triple forms to present the versatility. It is more general form to represent data on semantic web such as question answering, semantic search and knowledge-based systems are using triple form. A KG can have three properties in a triple form: Subject, Object, and Predicate. Knowledge linking and knowledge extraction are the core modules to design the KGs Wu et al. (2019) and regular and live extractions are considered two best strategies for knowledge extraction.

Data quality, accuracy, and completeness are the key actor to describe the usability of KGs. In earlier times KGs had presented as a heterogeneous information network Sun and Han (2012). Generally, KG is best suited to represent semantically structured information that easily interpreted by machines and it helps to empower several "BigData" applications in scientific and commercial domains. The integration of Google's KG is the best example that stores 18 billion facts about 570 million entities as a result of Google's search engine Singhal (2012) . KGs are considered as a significant phase to organize the text-based search engines into a semantic-based question answering system. Different approaches have been presented for KG construction Nickel et al. (2015) such as curated approach (e.g. WordNet, Cyc/OpenCyc, UMLS), automated unstructured approach (e.g. NELL, Knowledge Vault, PROSPERA, DeepDive), and automated structured approach (e.g Facebook, Google) and collaborative approach (e.g. Freebase, Wikipedia). According to the literature it is found that the progress of Wikipedia has been getting down and automatic KG construction methods getting more attention Suh et al. (2009).

As existing surveys mainly focused on Statistical Relation Learning Nickel et al. (2015), KG Refinement Paulheim (2017), Knowledge Graph Embedding Wang et al. (2017), KG Construction Wu et al. (2018), and Knowledge Representation Learning Lin et al. (2018). Wu and his team Wu et al. (2018) has introduced the construction techniques to prepare a Chinese knowledge graph. Paulheim Paulheim (2017) has presented a survey of approaches and evaluation methods for knowledge graph refinement. He has also discussed about few KGs for their existing features but not in descriptive manner. Ji et. al. Ji et al. (2020) presented a comprehensive review to cover different research topics such as knowledge graph representation learning, temporal knowledge graph, knowledge acquisition and completion and knowledge-aware applications.

The main motivation of the paper to present all existing KGs at one place with their different parameters. In this paper, we have presented a systematic overview of existing KGs in their present state and discussed different features of these KGs. To the best of our knowledge, such a characterization table among these most commonly used KGs has not been presented earlier. The presentation of KGs features covered KGs construction techniques, data quality, applications, problems and challenges, refinement, and evaluation. These features are the most significant aspects when it comes to considering which KG to use in a specific setting. The main contributions of this paper are:

1. Present a systematic review to analyse the related study and filtered relevant papers with selection criteria.
2. To present the most prominent KGs based on existing literature, with a characterization table of different features of these KGs.
3. To present and discuss the KGs construction techniques such as entity and relation extraction, supervised and unsupervised learning for knowledge extraction.
4. To present and explore different KG features and its applications, problems, challenges, refinement, and evaluation, etc.

The rest of this paper is organized as follows: Secton 2 presents the state-of-the-art of KGs to describe the foundation of KGs. Section 3 discussed the literature review to characterize different KGs by conducting a systematic review methodology. In Sect. 4, we have presented the Knowledge Extraction techniques for KGs that include entity and relation extraction, supervised and unsupervised learning for knowledge extraction. Section 5 explains several issues and challenges of different KGs. Section 6 presents the KG refinement and evaluation concepts and describes the most significant usage of KGs. Finally, we conclude and discuss future work in Sect. 7.

[1] https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.

Table 1 Existing studies on KG Construction

Purpose	Techniques	Extraction type	Extraction method	Study
Automatic extraction of structured information from the unstructured Chinese texts	Natural Language Processing	Automatic	Entity and relation extraction	Zhao et al. (2019)
Discuss several approaches to construct knowledge graphs in a biomedical area	Machine Learning	Manual & automatic	Rule-based extraction, unsupervised & supervised machine learning	Nicholson and Greene (2020)
Presented a knowledge base on cybersecurity and deduction rules based on a quintuple model	Machine Learning, Named Entity Recognition	Not Available	Rule-based extraction	Jia et al. (2018)
Building a large-scale multi-source knowledge graph from scratch	Named Entity Recognition	Not Available	Entity Extraction	Wang et al. (2019)

2 State of the art

KG presents information in the form of entities and relationships between entities. Several formalisms are available to present the knowledge in a structured way such as Frames Minsky (1974), Logic Davis et al. (1993), and Semantic Networks Sowa (2006). Recently semantic web community using the knowledge representation formalisms to create the ”web of data” that can interpret easily by machines Berners-Lee and Hendler (2001). The linked data concept Berners-Lee (2006) Bizer et al. (2011) published and linked the web of data on the web by using W3C RDF Klyne et al. (2004) framework. RDF and Semantic Web present open-world assumption (OWA) which interprets no-existing triples as unknown. Several KGs such as Neuro Commons Ruttenberg et al. (2009), Linked Life Data Momtchev et al. (2009) and Bio2RDF Belleau et al. (2008) are used in various specific domains to integrate the different biomedical sources in life sciences.

2.1 Knowledge-base and RDF in knowledge graph

Knowledge-base is a set of rules, facts and assumptions that stores knowledge in machine understandable format. To organize these facts Knowledge-Base needs a data model and a representation format. RDF is generally designed to represent the knowledge in the triple form (subject, object, predicate). For conceptualizing of structured data, a triple set forms an RDF graph. For example ”Abdul Kalam was born in Rameshwaram” can be represented as (sub:AbdulKalam, pred:birth_place, obj:Rameshwaram). In the RDF data model, triples are considered as an atomic element. It presents the knowledge using web resources with a unique URI. For example, the URI corresponding to “Abdul Kalam” can be presented as http://example.com/AbdulKalam where http://example.com/ is the address of every entity in the Knowledge-base. Subjects and Predicates are presented by URI while objects can be URI or literal values. RDF data can be stored using different formats such as RDFa, NTriples, Turtle, XML, N-Quads, JSON-LD, and TriG. Therefore, the fact can be presented as:

http://example.com/AbdulKalam.
http://example.com/born.
http://example.com/Rameshwaram.

RDF data model is suitable for knowledge-bases generated from non-linguistic data (e.g. Espana@es, India@en), it has string literal with language tag to build multiple knowledge-bases. RDF also supports ”IRIs” that is extended versions of URI. IRIs allows non-ASCII characters in the resource names. KGs are most widely used to describe more reliable text interpretation for analysing the text. A specific concept has a link in the KG by the semantic annotation of text to present the structured data. KGs have a specific data format to interpret the meaning of data and logic to derive new facts. It is easier for KG to add new facts continuously and updating dynamically.

2.2 Knowledge graphs Construction Tasks

Several studies proposed different approach to construct a knowledge graph. This section presented tabular information of knowledge graph construction techniques, purpose, and extraction type and method. Table 1 highlighted different parameters. Nickel et. al. Nickel et al. (2015) has

presented several tasks involved in KG construction; they are Entity Resolution (ER), Link Prediction, and Link-based Clustering. ER task identifies the entities that are semantically equivalent and referring to the common real-world objects such as persons, publications, movies, or products. This task is also known as object identification Tejada et al. (2001), data de-duplication Culotta and McCallum (2005), instance matching Rahm and Bernstein (2001), or link discovery Newcombe et al. (1959). Link prediction task seeks to infer new links that can be missing edges or missing part of edges between the entities of a KG. It was generally presented for social networks as a single relation but later extended to relational data and applied to KGs Ferre (2019). In KG context, link prediction also termed as completion of KG as existing KGs are still missing several facts and sometimes, they contain some incorrect edges Angeli and Manning (2013). Link-based clustering is an extension of feature-based clustering to group entities based on similarity in relational data Nickel et al. (2015). Link-based clustering is also termed as community detection Fortunato (2010) in social networks. A scalable tool FAMER Newman (2001) Liben-Nowell and Kleinberg (2007) has been developed to integrate the parallel implementation of several clustering schemes. This tool can construct similarity graphs for entities of heterogeneous sources based on multiple linking schemes.

Li et al. (2020) has proposed a systematic procedure to construct KGs from large-scale electronic medical records (EMRs). This procedure involved 8 steps to construct KGs one by one;

Data Preparation→Entity Recognition→Entity Normalization→

Relation Extraction→Property Calculation→Graph Cleaning→Relate-entity Ranking→ Graph Embedding.

They have used probabilistic translation on hyperplanes (PrTransH) algorithm to learn graph embedding for the developed KG. This KG is mainly applied in three applications such as: medical information retrieval, clinical decision support system, and knowledge transferring with neural networks.

2.3 Reasoning in knowledge graph

Data is extracted from the web to construct the KG which has noisy facts and found error-prone and incomplete. It is essential to reduce the conflicts in the KG and it can be managed by reasoning and inferencing. There are three types of reasoning such as logical reasoning, graph-based reasoning, and entity and relation embedding-based reasoning. In logical reasoning, rules are existing between the relations in KG which are based on literals. For example:

$\text{bornInCity}(p, q) \wedge \text{cityInCountry}(q, r) \rightarrow \text{bornInCountry}(p, r)$

As rules are designed with fixed properties in several KG. Rules are easier to detect inconsistencies found in relation instances in KG. For example: bornInCity(AbdulKalam, Rameshwaram)∧cityInCountry(Rameshwaram, India)→ bornInCountry(AbdulKalam, India) will have a conflict with bornInCountry(Rameshwaram, Indonesia).

NELL Carlson et al. (2010) is a rule learner and based on FOIL while SOFIE Suchanek et al. (2009) is the first system integrates logical reasoning. Graph reasoning improves the performance to extract new relation instances. Major tasks, such as prediction of missing entities, relations and facts are involved in KG reasoning while it has different applications such as question-answering, KG completion, and recommender systems.

3 Literature review

The primary aim of this review to present the detailed features of several existing KGs such as academia, industry, and expert KG system and approaches to construct the knowledge graphs. A systematic review is planned and conducted to perform a deep analysis of existing KGs.

3.1 Planning and conducting the review

This phase generally aims to define how the review protocol will be accomplished. Some common points are suggested by Kitchenham and Charters Keele (2007) to follow: 1) selection of source and search criteria 2) research questions 3) the exclusion and inclusion criteria 4) selection criteria. In order to describe the search criteria and source selection, specialized journals, abstract and citation sources of peer-reviewed literature, have been considered. Table 2 describes the selected sources to perform the search from existing sources more precisely.

For preparing a systematic literature review, the selection of sources is an important part. In this paper, several databases are included to conduct the review. Despite databases, specific journals have also been examined because it has published several papers on the related domain. After specifying the sources, it is required to define the search term that are related with the specific domain involved in KGs. The retrieval of documents for the review can be based both on manual and electronic searches. The search criteria can be different like Keywords, Year, Author Name, and Subject Area. Some search keywords are Knowledge Graph Construction, Information Extraction in Knowledge Graphs, Knowledge Graph Techniques, and many more.

For the literature review of knowledge graphs, we have identified 15 years of paper (2005 to 2020), as knowledge graphs term is announced in 2012. With the help of searching keywords, a total of 81 papers are selected and extracted

Table 2 Selected Sources for Review

Source	Type	URL
Google Scholar (GS)	Database	https://scholar.google.com/.
Scopus	Database	https://www.scopus.com.
Web of Science (WoS)	Database	https://apps.webofknowledge.com.
Semantic Web Journal (SWJ)	Journal	https://content.iospress.com/journals.
Applied Ontology (AO)	Journal	https://www.iospress.nl/journal/applied-ontology/.
Journal of Web Semantics (JWS)	Journal	https://www.journals.elsevier.com/journal-of-web-semantics

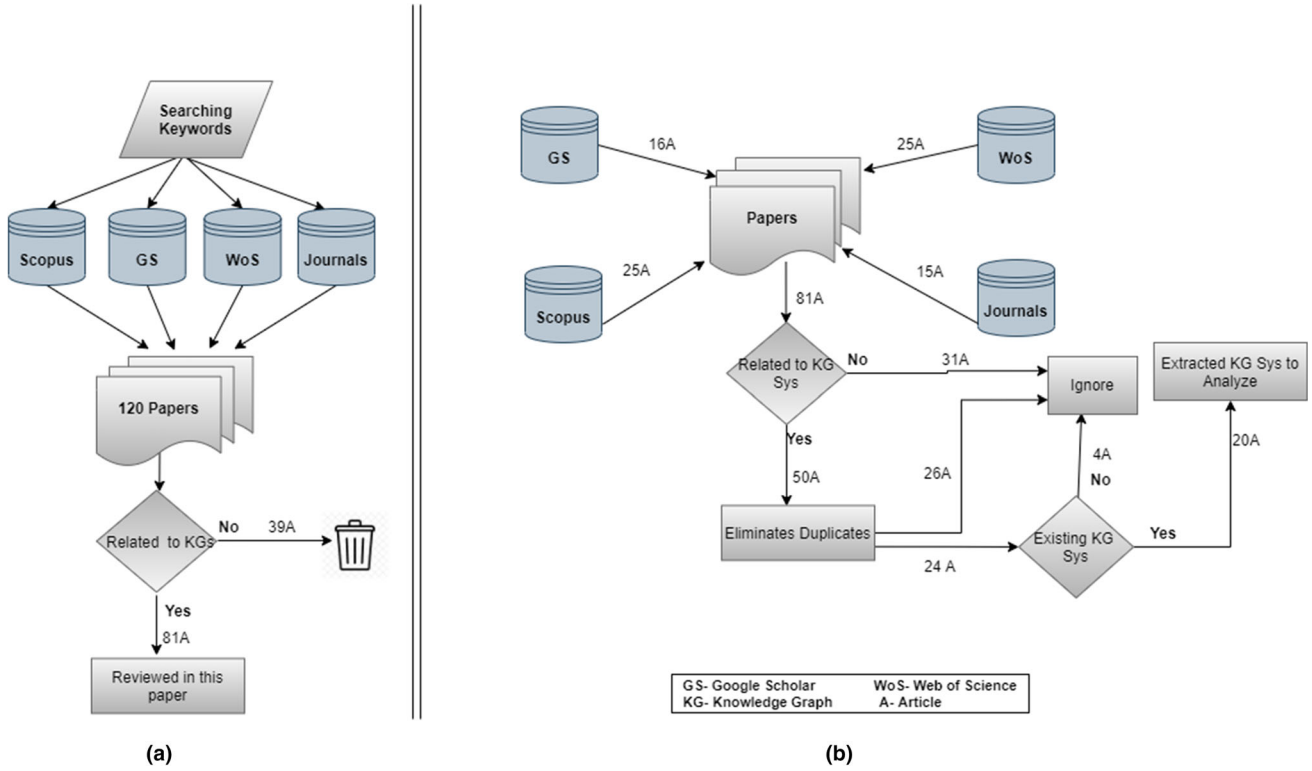

Fig. 1 Methodology to conduct review on KGs

from different sources such as databases (GS, WoS, Scopus) and Journals (SWJ, Applied Ontology, JWS) that offers significant information related to the research queries. Then the inclusion and exclusion criteria have been applied to analyse the fundamental research that only related research of existing KGs has been taken into inclusion criteria while duplicate and irrelevant study has been excluded as an exclusion criterion.

After following the inclusion criteria, we have extracted 50 papers from all sources but only 20 sources are considered to analyse the existing knowledge graphs of different categories.

Figure 1a shows the proposed methodology for conducting this review, while Fig. 1b shows the proposed methodology for conducting the review of existing KG systems.

3.2 An overview of existing knowledge graph

In this section, an overview and summarization of existing KGs have been presented. KGs Ringler and Paulheim (2017) can be classified into three varieties such as: General-purpose KGs (e.g. Yago, DBPedia, Wikidata, Freebase, OpenCyc, NELL, etc.), Expert KGs (e.g. FIBO, etc.), Industrial KGs (e.g. IBM, eBay, etc). One of the major differences between the two categories that the general-purposes KGs require a huge amount of common-sense knowledge is not necessarily strictly validated. On the other hand, experts KGs contain strict knowledge accepted in the expert community of their respective domains. Industrial KGs have the descriptions and association of places, people, things and organizations, and present general knowledge about the world. Table 3 shows the KGs and their affiliation, input dataset, used techniques,

target, type of construction, type of KG, size of KG, and their references.

Table 3 provides an extensive and well-summarized comparison (at the same time) among the most 20 prominent KGs with different characterizing parameters. The first column denotes the system name of KG as in the column's title. The second column (affiliation) denotes the development organizations or teams which mostly are large companies or notable research teams. The third column (input type) describes where the contents and data of the system are from. The fourth column (used techniques) presents the techniques that are used for extracting, acquiring and inferring the entities, relations and facts. The fifth column (target) describes the kind of knowledge that is acquired by the system as well as the format of data storage such as RDF, OWL, JSON, etc. The language of KG is English by default, except if another language is mentioned. The sixth column (construction type) describes the approaches and methods that are used to construct the KG. "Manually" denotes that the KG construction process is performed completely by a human, "cooperatively" denotes that the most tasks of KG construction are performed by a human, and "automatically" denotes that the many tasks of KG construction are performed by supervised or unsupervised techniques. The seventh column (KG type) defines what if KG is open-source or not. The eighth column (graph size) describes the content size of knowledge (entities, relations, and/or facts) of KG. Finally, the ninth column (ref.) shows the related references for mentioned KG.

Table 3 is considered a characterization table among the most commonly used KGs that have not been presented earlier. There are also some related work with different approaches that has been discussed to explore different methods for KG, such as Zhu & Iglesias Zhu and Iglesias (2015) proposed a framework for searching entity in the KG. This framework involves different types of approaches such as entity linking, natural language query processing, and semantic similarity. The proposed system interprets natural language queries into SPARQL queries. In addition, Fatiha Sais Saïs (2019) has proposed several approaches to refine the KGs to resolve different issues like Link Invalidation, correctness, and completeness, weak identity relation, and missing value prediction, etc. Some efficient methods are described to resolve the issues such as KD2R, to explore exact keys, SAKey to explore n-almost keys, and VICKEY to explore conditional keys. Moreover, Shaoxiong et. al. Ji et al. (2020) has presented a survey on knowledge graphs to explore different research areas such as representation, and learning of KGs, acquisition, and completion of KGs, knowledge-based applications, and temporal knowledge graphs. They also discussed several techniques about relational learning on KGs and neural structure of KG representation learning (KRL) and reasoning. KRL approaches are further categorized into four representation aspects: scoring function, space, auxiliary information, and encoding models.

4 Knowledge extraction for knowledge graph

KG is a Knowledge-base that represented in the graph. Knowledge-base is a technology that used to store complex information (structured and unstructured) which represents facts (knowledge) about the world (selected object). A graph is a collection of points and lines (objects) connecting and pairs some subset of them (possibly empty). We can simplify the KG definition as follows: KG is a collection of entities, attributes, relations, facts, and rules or other forms of knowledge that express connections or relationships as a paradigm rather than a specific class of things.

Fig. 2a presents an example of KG which is a collection of entities (E_i), attributes (A_i) and relations (R_i), while Fig. 2 b illustrates an example data of these entities, attributes, and relations.

4.1 Entity and relation extraction in existing knowledge graph

Supervised or unsupervised learning models deal with the formation of classification problems and can easily discriminate the problems into respective classes. Such approaches extricate several features from the sentence which for the most part incorporate words, many tags related to parts of speech, etc., and the relating labels are acquired from an enormous trained labelled corpus data. Even though such strategies acquire great exactness and consider negative samples more to extract the relation between the entities and they are neither general nor adaptable. Such techniques are over the top expensive because of the necessity of a huge amount of data. In addition, the relations learned from these strategies are of great extent subject to the field and hence, not practical for extracting the relations between entities on large web-scale data.

4.1.1 Entity extraction

Entity extraction is also termed as entity annotation or entity linking Zhao et al. (2019). Several extraction methods have been introduced to extract the entities Lehmann et al. (2015) Hoffart et al. (2013) for Wikipedia. Named Entity Recognition (NER) is also played a significant role to identify and categorize information such as location, person, organization called named entities are entity annotation. Entities can be associated with several text described in documents; for example, 'Jaguar' can interpret the concept Animal or Jaguar Company. Entity linking links text to their corresponding rep-

Table 3 An Overview of Existing Knowledge Graphs with Different Features

KG	Affiliation	Input Type	Techniques	Target	Construction Type	KG Type	Graph Size	References
Cyc	Cycorp	Manual assertions (rules and common-sense knowledge)	Manual	Knowledge Base	Manually	Partially open source	0.5M concepts & 17000 relation & 7M assertions	Matuszek et al. (2006)
NELL	CMU	Seed ontology & web pages	Contextual pattern, POS, Inductive & Logistic regression	Populating ontology (RDF & TSV file)	Automatically	Open source	-Seed ontology (123 categories & 55 relations) -populated ontology (seed ontology with over 242,000 facts)	Carlson et al. (2010)
WordNet	PU	Expert-authored	Manual	Multi-languages ontology repository (OWL/ RDF)	Manually	Open source	117000 synset grouped as a synonym	Miller (1995)
Freebase	Metaweb	Metadata, structured data	Manual	Cross linked data	Manually	Open source	Approximately 44 million topics and 2.4 billion facts.	Bollacker et al. (2007)
YAGO	MPII	Wikipedia's category pages & WordNet	Template-based & rule-based	Ontology that is anchored in time and space (OWL & a slight extension of RDFS)	Cooperatively	Open source	>10M entities >120M facts about these entities.	Suchanek et al. (2007)
FrameNet	UC Berkeley	Text Summarization	Software tools with PERL/CGI-based	Multi-languages lexical database-asymmetric directional relation (XML & RDF)	Cooperatively	Open source	10,000-word senses >170000 sentences	Baker et al. (1998)
DBPedia	DBPedia	Wikipedia pages	Wiki parser, Infobox templates & TF-IDF	LOD Cloud & RDF	Cooperatively	Open source	1.86 B facts that describe 13.7M things in 111 languages	Lehmann et al. (2015)

Table 3 continued

KG	Affiliation	Input Type	Techniques	Target	Construction Type	KG Type	Graph Size	References
KnowItAll	UW	Web Pages	POS & Feature-based classification	Domain independent extraction	Automatically	Open source	54 753 facts	Etzioni et al. (2005)
ConceptNet	MIT	Wiktionary	NLP techniques, graph-structured knowledge	Multilingual KG(JSON file)	Automatically	Open source	21M edges and over 8 M nodes	Liu and Singh (2004)
HowNet	Keenage	Chinese Character	NLP techniques, graph-structured knowledge	Chinese & English Bilingual Knowledge	Cooperatively	Open source	800 sememes in HowNet	Dong and Dong (2003)
Probase	MSRA	Probabilistic models	NER, Word Sense Disam-biguation	Probabilistic Ontology	Automatically	NA	2.7M concepts	Wu et al. (2012)
GKV	Google	Web Data	Classification techniques	HTML DOM trees, HTML Web tables	Automatically	Open source	1.6B triples	Dong et al. (2014)
GKG	Google	Ontology, Facts from Factbook, Freebase, Wikipedia	Crowdsourced	Multi-languages semantic search functionality (RDF triples)	Automatically	Open source	3.5B facts over 500 M objects	Singhal (2012)
Facebook KG	Facebook	Post and comment data from Facebook blogs	NLP techniques & Template-based	Semantic Search Service (GraphML)	Automatically	Open source	50 M primary entities, 500 M assertions	Sengupta (2013)
Satori	Microsoft	Bing search streams	Feature and Pattern based	Structured database & RDF	Automatically	NA	300 M entities 800M relations	[A]

Table 3 continued

KG	Affiliation	Input Type	Techniques	Target	Construction Type	KG Type	Graph Size	References
Prospera	Prospera	Semi-Structured text	Hearst-pattern & Pattern-based	Web-scale knowledge graph (RDF)	Cooperatively	NA	2M typed entities	Nakashole et al. (2011)
DeepDive	DeepDive	Diverse Data Resources	Markov logic, Supervised Approach	Factor Graphs	Automatically	Open Source	NA	De Sa et al. (2016)
IBM	IBM	Structured and Unstructured text	Watson Discovery Services	Customer oriented knowledge graph	Automatically	NA*	>100M documents, 5B relations, 100M entities	Noy et al. (2019)
HowNet	Keenage	Chinese Character	NLP techniques, graph-structured knowledge	Chinese & English Bilingual Knowledge	Cooperatively	Open source	800 sememes in HowNet	Dong and Dong (2003)
eBAY	Beam	Unstructured text	ML Techniques and replicated log	Flattened document store	Automatically	NA	around 100M products, >1B triples	Noy et al. (2019)
FIBO	FIBO	Structured text	Machine Learning Techniques Business conceptual model (OWL	RDF)	Automatically	Open Source	NA	[B], [C]

*GKV- Google Knowledge Vault, GKG- Google Knowledge Graph, M- Million, B- Billion, NA- Not Available,
[*A*]- http://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/,
[*B*]- https://spec.edmcouncil.org/fibo/doc/FIBO_BTDM.pdf,
[*C*]- https://edmcouncil.org/page/aboutfiboreview

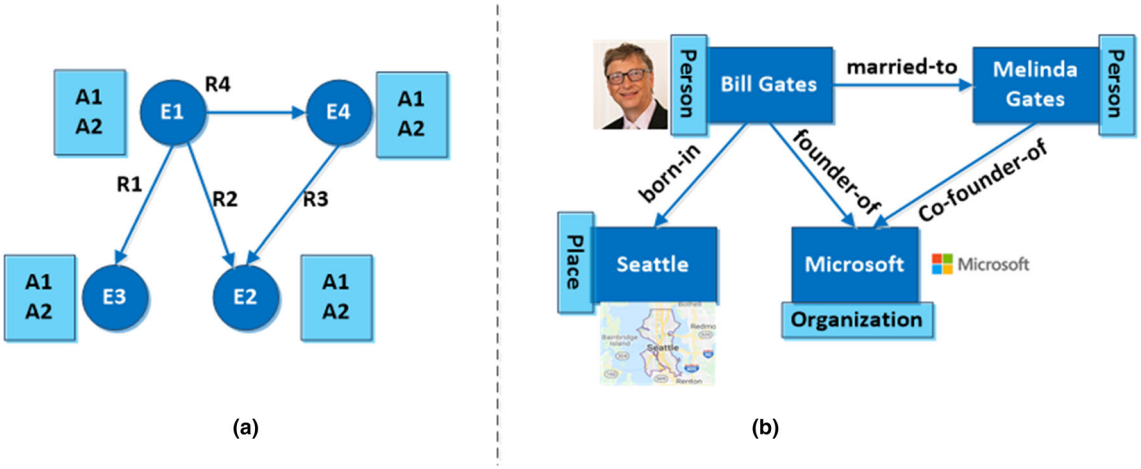

Fig. 2 An Example of Knowledge Graph

resentation in KGs. It links structural and textual knowledge together. Yan J. and colleagues Yan et al. (2018) presented a function to learn the entity:

$function\ f : m_i \rightarrow e_j,\ for\ m_i \in M\ and\ e_j \in E$ This function presents a mapping between mentioned entities m_i and entities e_j in KG.

4.1.2 Relation extraction

RE is a process to collect facts about entities and extract the binary relation as a triple (subject, predicate, object). Here predicate presents the semantic links between subject and object. For example, "Abdul Kalam created Missile" can present in a tuple: (Abdul Kalam, created, Missile). Supervised learning is efficient for RE but recently it is observed that deep learning is more suitable for RE Yan et al. (2018).

RE process also referred as a binary classification and categorized in to three sub types: relation classification based on supervised approach; pattern-based extraction follows semi-supervised approach; open relation extraction follows unsupervised approach.

Relation classification follows the supervised learning methods which further classified into feature-based and kernel-based extraction. In feature-based relation extraction, different set of features are applied such as word features for entities headwords, Entity features for named entities (location, person) and Parse features. On the other hand, kernel-based extraction methods are generally used in machine learning area. String kernel Kambhatla (2004) is an example to compare text documents by their substrings. In relation extraction, string kernel can be extended to the complex structure such as word sequences or parse tree. Pattern-based extraction is in trend for textual analysis in relation extraction. Patterns can be identified as noun phrases like $NP_0, NP_1, NP_2, ..., NP_n$ which follows the Hearst pattern Hearst (1992). Pattern-based relation extraction followed a bootstrapping approach to identify the patterns and explore the relation instances. It uses seeds to populate the relations. Generally, relation extraction system follows supervised method or semi-supervised methods. Open relation extraction extracts all relation instances from web data. It has a framework to extract all relation facts from the web. These systems can generate a large amount of knowledge automatically but errors can be frequently arising.

4.2 Knowledge graph construction based on data resources

KG data can be collected from different resources and types of data. KG can be gathered and extracted from the structured text (e.g. Wikipedia, tables, databases, and social nets), unstructured text (e.g. web pages, news, and reference articles), images, and videos. We can divide the studies and techniques regarding extracting and constructing KG under three groups of approaches that can be categorized by the type of data sources, as in the following sub-sections.

4.2.1 Knowledge graph construction from repositories and encyclopedias

There are many repositories (cross-domain) that contain millions of entities and facts for KG, such as YAGO (Yet Another Great Ontology) Freebase[2], BabelNet[3], and DBpedia[4] that their prime ontologies and knowledge rely on structured contents of Wikipedia pages. YAGO Suchanek et al. (2007) is

[2] https://developers.google.com/freebase.

[3] https://babelnet.org/.

[4] https://wiki.dbpedia.org/develop/datasets.

additionally using WordNet[5] to build the ontologies. However, Wikipedia categories are often quite fuzzy and irregular this is considered one of the disadvantages of these repositories Arnold and Rahm (2014). In addition, due to the sparse of non-English data in Wikipedia, there are many studies that tried to construct KG from multiple non-English encyclopedias such as CN-DBpedia Xu et al. (2017), XLore Wang et al. (2013), Zhishi.me and Zhishi.me2 Niu et al. (2011), which used the Chinese online encyclopedias, i.e., Baidu Baike, HudongBaike and Chinese Wikipedia as the main data sources to extract and construct the KG.

In the Wu et al. (2019) study, the authors proposed a framework that aimed to construct KG from multiple online encyclopedias. They have considered knowledge linking and knowledge extraction as core modules to design the KG. Knowledge extraction consists of regular extraction techniques, while knowledge linking applies a semi-supervised learning method and heuristic lightweight entity matching strategies to detect the common properties and entities from multiple online encyclopedias.

In the Abouenour et al. (2014) study, the authors established a new Arabic ontology repository, for question answering (QA) application, in the Arabic language. This ontology combines the lexical information and hyponymy relations in Arabic WordNet (AWN)[6] with the syntactic and semantic frames of verb classes in Arabic VerbNet (AVN)[7]. Then the authors transformed the AVN frames into the conceptual graph formalism, in order to ensure the usability of their ontology in QA application.

All of the above studies are good studies, but as we know, the data, which is available in a semi-structured format on the web, is a part of the information but there still a huge part of the information that is available in an unstructured format Heist (2018). That leads to a need for extracting techniques to extract knowledge from unstructured data.

4.2.2 Knowledge graph construction by using or constructing ontology

There are many approaches and studies that use an existing ontology and/or constructing taxonomies for constructing the KG. In study Carlson et al. (2010), the authors proposed a prototype called Never-Ending Language Learner (NELL).The goal of this approach is to grow the knowledge-base by continuously reading. Firstly, the initial seed ontology and a handful of seed examples for each predicate in this ontology are defined the knowledge-base. Then, populating this ontology by continuously reading from the web for 67 days. Knowledge Vault (KV) Dong et al. (2014) is an automatic constructing Web-scale probabilistic knowledge-base that aims to fuse together multiple extraction sources with prior knowledge derived from an existing knowledge-base. Heist Heist (2018) study focus on entity co-occurrence for constructing KG, it seeks to discover patterns that indicate relationships between the included entities. This study used an existing ontology as seed (KGs), then patterns including groups of entities that are related to two folds, are identified (connected on the document surface and have a semantic connection). Contrast the many studies such as in Carlson et al. (2010) Dong et al. (2014), that focused on relations between two entities at a time, this study focuses on relations among multiple entities.

The main shortcoming of this study is its inability to extend and populate the seed ontology. All the above studies in this section used the semi-structured data (web-pages) for constructing KG.

4.2.3 Knowledge graph construction by information extraction techniques

In recent years, there are many research and techniques that try to extract the entities, attributes, relations, facts, and rules from unstructured data for KG. These studies used information extraction (IE) techniques to extract and construct knowledge. IE is the process of automatically extracting information from corpora, it is linking references to named entities with stated relationships between such entities. IE process consists of three main steps to extract KG as follows:

1. processing and tagging the parts of speech;
2. detecting and identifying the named entities;
3. annotating the text by using the annotation methods such as dependency parser and paths.

For further explanation, suppose we have this text "Bill Gates was born in Seattle", Fig. 3 shows an example of IE process to extract KG from this sentence. In the last two decades, numerous researches that use IE to extract knowledge have emerged. TEXTRUNNER Banko and Etzioni (2008) is one of IE system which used a Naive Bayes with un-lexicalized POS and NP-chunk features to extract the information. It focuses on extracting binary relations from the text by the form (arg1, relation phrase, arg2).

There are other IE systems that using (a linear-chain) Conditional Random Fields (CRFs) such as in Banko and Etzioni (2008) Tiwari et al. (2018) or using Markov Logic Network (Markov-chain) such as in Al-Aswadi et al. (2019) to extract the knowledge. Using linear-chain or Markov-chain can lead to improving knowledge extraction, but it still a challenging process Zhu et al. (2009). In the study Banko

[5] https://wordnet.princeton.edu/.

[6] http://www.globalwordnet.org/AWN/.

[7] http://ling.uni-konstanz.de/pages/home/mousser/files/Arabicverbnet.php.

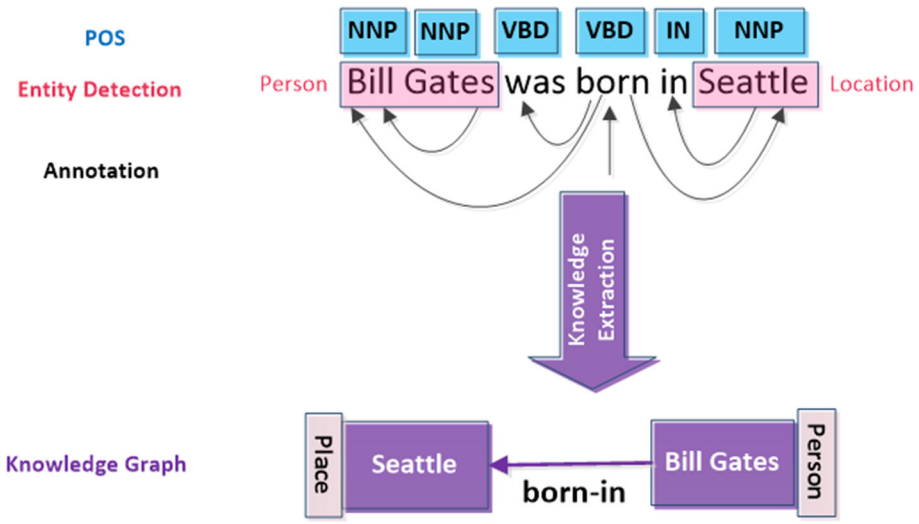

Fig. 3 Information Extraction

and Etzioni (2008), the authors used CRF to develop TextRunner system for extracting knowledge, while in the Banko and Etzioni (2008) Zhang et al. (2016) studies, the authors develop new IE systems by using CRF and Etzion Etzioni et al. (2008) developed an IE system called O-CRF, this system used CRF as well as POS tagging and phrase-chunking to extract the knowledge. While Zhang, et al. Al-Aswadi et al. (2019) developed Simultaneously Entity and Relationship Extraction (SERE) IE system based on CRFs to extract binary relationships from the text. There are other studies that use deep learning techniques to extract knowledge such as in Wang et al. (2017) Chen et al. (2010) Zhong et al. (2016). The Chen et al. (2010) Zhong et al. (2016) studies used Deep Belief Networks (DBNs) for extracting the knowledge. While the Wang et al. (2017) study used Convolutional Neural Networks (CNNs) to classify the text then using this classified text as well as the TF-IDF matrix to construct knowledge. As well, there are other reviews and research that showed and discussed the ontology construction (entities or concepts and their relations) as in Tiwari et al. (2018) Mishra and Jain (2019) Tiwari and Abraham (2020), and deep learning for ontology construction and KG as in Al-Aswadi et al. (2019) Liu and Han (2018).

4.3 Supervised and unsupervised learning for knowledge extraction

To check whether the learning algorithm performs in a better way, certain questions need to be addressed here: (1) Are enough features used in the training time during classification? (2) Does the size of the training set is very huge? (3) Does the classifier carry out the learning process in a better way? (4) Up to what extent validation turns are required to minimize the error? To answer such questions, a feature matrix is formed with only relevant information. The primary goal of the learning algorithm is to make the classifier learn in a better way such that either of training error or test error is reduced. This is only possible if and only if the relevant information is extracted from the dataset Abualigah et al. (2018, 2019).

The information needed by the learning algorithm can be easily represented in the form of an undirected graph G=(V,E), where V is a set of user entity and E is another set of entities. Correspondingly, $V = v_1, v_2, ..., v_n$ and $e = e_1, e_2, ..., e_n$. Further, to extract such relevant entities, a candidate pair (p) of (V, E) is formed and grouped together in the matrix. These candidate pairs are called features and the matrix thus formed further is called a feature matrix. The feature matrix contains a mapping relation X between these entities $X \in R^{|V|\times|E|}$. Most of the methods have been used to optimize this mapping in an unsupervised manner, with the information of V and E.

However, the method of supervised learning is when the model is getting trained on a labelled dataset. The statistical systems that are being used to extract the keywords from the knowledge graph are totally based on manual methods. These made the designers compulsory to devote a substantial amount of time to make the method to work in an automatic way.

4.3.1 Supervised learning for knowledge graph

Supervised models utilized in the field of data extraction include how classification problems are formed and how the given classifier can classify the extracted features from the sentence in an accurate positive and negative sample? These features include part of incorporate words, many tags related to parts of speech, etc., along with their corresponding labels. Even though such strategies acquire great exactness and consider negative samples more to extract the relation between

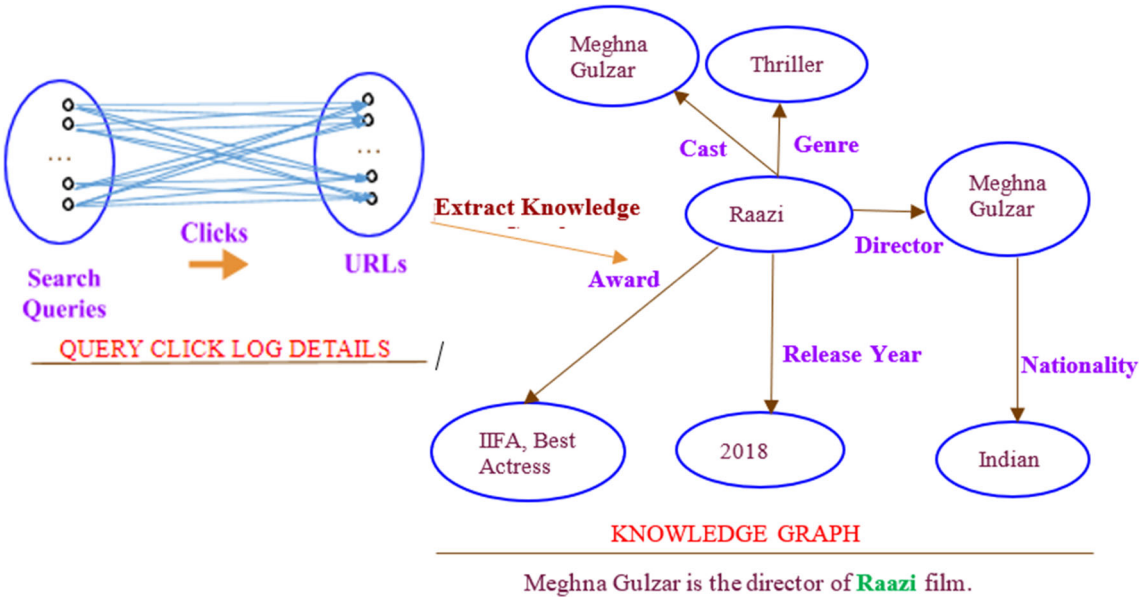

Fig. 4 Example of query click logs

the entities and they are neither general nor adaptable. Such techniques are over the top expensive because of the necessity of a huge amount of data. In addition, the relations learned from these strategies are of great extent subject to the field and hence, not practical for extracting the relations between entities on a large Web-scale data Tiwari et al. (2018) Gaurav et al. (2020) Rahul et al. (2019).

To carry out such a process in a rigorous manner, some kinds of features are extracted first from the entity relations. The features may be related to the word, entity, and parsed data. For extracting such entity features, parts of speech (POS) tags are enough but when word related features are extracted then, POS tags are not enough. The better solution for such word related feature extraction, parsing is required. Parsing is done throughout the feature dependency tree. A kind of rule is defined to carry out the extraction process as:-

$$P = \sum_{i=1}^{n} Feat\,(p_i) \tag{1}$$

The function Feat() involves different functions like hyponym, concatenation, etc. Parameter P is a noun phrase that is taken as entities in the graph. These are produced by POS tags. The features are called patterns in terms of a graph. The extraction of patterns generally follows a bootstrap procedure. This procedure uses seeds for making the algorithm to learn about the relation of entities. In the first phase, candidate pair relations are gathered together as seed pairs. After that, pre-processing is done for cleaning the seed pairs. Then, patterns or features are extracted in the context of seed pair. After using a pattern matching technique, relations are established. The training set is formed with a pair of relations and entities. This training set is passed to the classifier and the performance of this model is calculated. This process is repeated iteratively until the error is minimized.

4.3.2 Unsupervised learning for knowledge graph

Heck et. al. Heck et al. (2013) presented a new approach for unsupervised semantic parsing with semantic KGs with no requirement for semantic schema design, no data collection, and no manual annotations. We develop a graph crawling algorithm for data mining, and two entity extraction approaches a CRF-based method with unsupervised MAP (maximum a posteriori) adaptation and a relation model with induced entity extraction grammars. DilekHakkani-Tur Hakkani-Tür et al. (2013) presented a novel statistical language which helps to understand the pattern of evolving semantic web. The motive is to give more emphasis on the knowledge

being colonized for extracting the data related to the KG domain instead of just making the models/learning algorithm. The first way to precede this is to form the training relationship between the semantic web entities as a graph. This works in an unsupervised manner. This process also leads to form a relation between a set of connected entities over the graph, and then the parried entities are searched over the web correspondingly. These spotted relations are further used to extract the natural language queries with the help of knowledge-based theories. From these relations, certain observations are carried out using KG, and iterations are done using a bootstrap approach.

KG is drawn by the mining process from the WWW and the click logs related to the search query, Fig. 4 shows an example of the outlines of the query click logs and KG. The

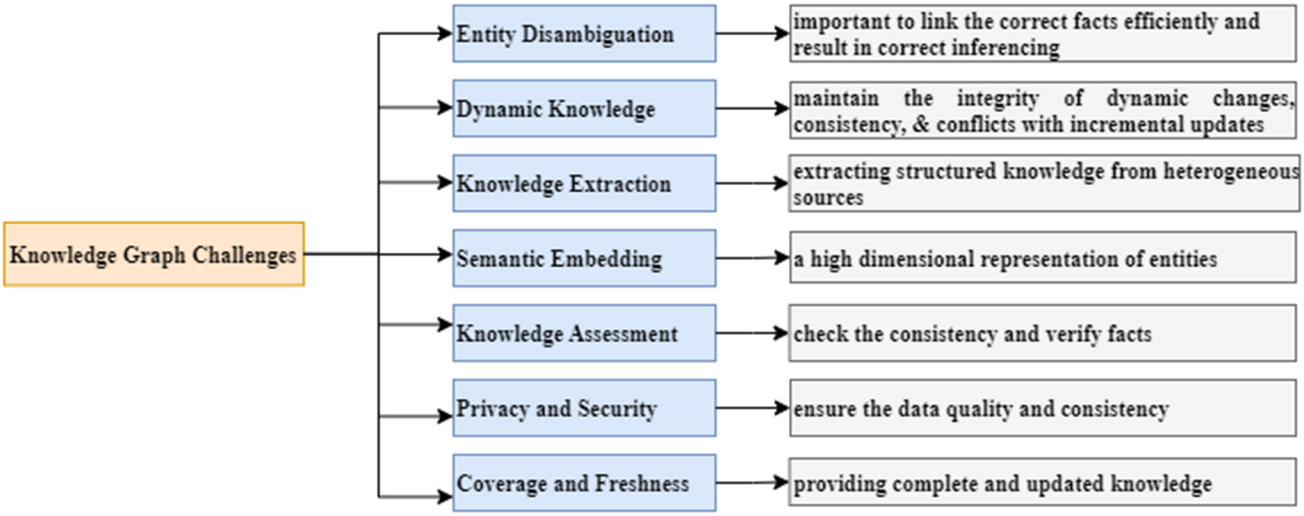

Fig. 5 Knowledge Graph Challenges

mining and refining of the annotations are done by the bootstrap method. This method iteratively uses the pre-trained model for finding out the relations related to the mined data in a more efficient way. In relation to these, patterns are mined first from the search results and training data are extracted in the form of queries obtained from clicked websites along with the entity pairs. It happens so because some of the entities may have more than one relation or may be connected to other entities as well. To present this in a more lucent way, some snippets are required for that. For example, Raazi is a thriller film which is directed by Meghna Gulzar and Alia Bhatt is a starring actress in this movie. This mined information is drawn from the training examples for the "Director" relation along with two other relations "Cast" and "Genre". For refining such annotated examples, the bootstrap method is used to draw a relation among a classifier, extracted data, and their corresponding annotations. The purpose of this classifier is to pick up the unlabelled data and label them properly with their corresponding relations. Only those relations (say v) are included further which have a higher chance of

being appeared in the expression (say e). At last, the threshold T is optimized for finding v in terms of probability P(v/e).

5 Knowledge graphs challenges and problems

KG represents knowledge in a graph, based on the extracted entities and their relationships. The need for KG comes as a result of the Internet development that has been coupled with the continuous growth of data on it. Despite the many advantages of existing KGs, but they still face many problems. As well there are many challenges that can determine the directions of future research on KG construction or refinement. The following sub-sections present these challenges and problems.

5.1 Challenges in knowledge graphs

Researchers and engineers addressed several challenges for the implementation, coverage, requirements in KGs. These challenges include knowledge extraction, disambiguation, managing variable knowledge, and many more. In this section, most prominent challenges[8] are discussed in Fig. 5 that appear consistently

5.2 Problems in knowledge graphs

KGs are successfully used in most of the search engines as well as social network sites (LinkedIn, Facebook), e-commerce (eBay, IBM) sites, they are continuously growing due to its dynamic nature and sometimes generating problems[9] such as Data Insufficiency, Explainability, Integration of Knowledge[9], Inconsistencies and many more. Figure 6 has presented all these problems.

When users make a query to find some knowledge stored in KG, sometimes information is missing and it returns either incomplete or incorrect knowledge. For example, in Fig. 7 a query is written to find the Google CEO in 2013. It gives the right answer with the wrong picture. It is showing the picture of the current CEO, Sundar Pichai as Google is taking it from Wikipedia's page of Sundar Pichai. Somehow it is incomplete and incorrect knowledge represented by Google Knowledge

[8] https://queue.acm.org/detail.cfm?id=3332266.

[9] https://www.poolparty.biz/what-is-a-knowledge-graph.

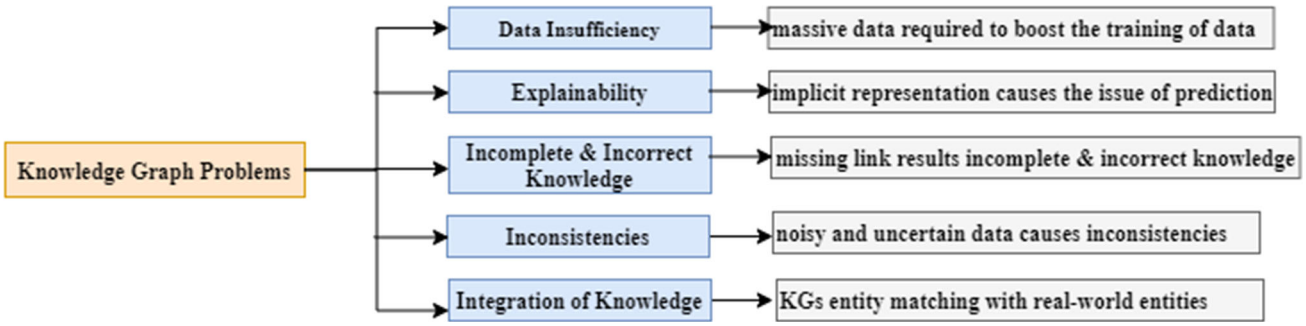

Fig. 6 Knowledge Graph Problems

Graph. These types of problems should be removed from the KGs.

It is also crucial to managing the consistency Chekol et al. (2017) of extracted knowledge when acquiring structured data from noisy and unstructured sources. Generally, KGs deal with both numeric and non-numeric facts. The rules and facts contained in existing KGs tend to be erroneous and noisy due to either uncertainty in the source data or the inconsistency of the extraction tools. This uncertainty and inconsistency cause incorrect interpretation and improper conceptualization. To overcome with this recently temporal KGs Chekol et al. (2017); Chen et al. (2019) are proposed to maintain the consistency.

6 Knowledge graph refinement, evaluation and applications

KG has many applications that are used. In addition, there are many studies that try to refine and evaluate existing KGs. The following sub-sections present

the KG refinement and evaluation concepts as well as the most used applications of KG.

6.1 Knowledge graph refinement and evaluation

KGs are freely usable and accessible for interacting with the stored and modelled knowledge. As KGs are growing day by day and modelled knowledge is also variable in nature, hence it is required to refine the stored knowledge in KGs to maintain consistency and accuracy. Refinement is a process to include missing knowledge and detecting errors. KG refinement methods Paulheim (2017) identify inconsistent knowledge and validate consistent knowledge. Several KG refinement approaches exist for instance-level (ABox) and concept level (TBox) such as error detection, completion, refinement target, external and internal methods.

Error detection and completion are two primary goals for KG refinement. Error detection method identifies the inconsistent knowledge in the KGs while the Completion method includes the missing knowledge in the KGs. Refinement can be targeted by information such as entities of KGs, relations among entities, and interconnection between multiple KGs, literal values. External approaches use additional knowledge such as text corpus while internal methods use only KGs as input.

Evaluation plays an important role to refine the stored knowledge in KGs. There are several evaluation methods for knowledge refinement such as Partial Gold Standard, Silver Standard, and Retrospective Evaluation. The Partial gold standard evaluation is a common strategy for KG evaluation Färber et al. (2018). In this evaluation, 34 metrics were defined to evaluate the KG quality and analysed on the Freebase, DBpedia, YAGO and Wikidata. These metrics are further categorized in four groups such as Contextual (completeness, relevancy, timeliness), Intrinsic (accuracy, consistency and trustworthiness), Accessibility(license, accessibility, interlinking) and Representational Data Quality(Interoperability and ease of understanding).

6.2 Knowledge graph applications

According to the different studies[10] Zou (2020), KGs are successfully applied in various applications such as Query-answering, Semantic Searching, Knowledge Sharing, Recommender System, Dashboard, Knowledge Management and Domain-Specific (Medical, Finance, Education etc.). Figure 8 has presented different applications along with specific areas. Figure 8 has shown that KGs have richer ability to provide sematic structured data of specific domains.

7 Conclusion

KG is an essential knowledge repository for both natural language understanding and for logical reasoning. It contains a wealth of knowledge about the entities, their attributes,

[10] https://www.slideshare.net/phaase/getting-started-with-knowledge-graphs.

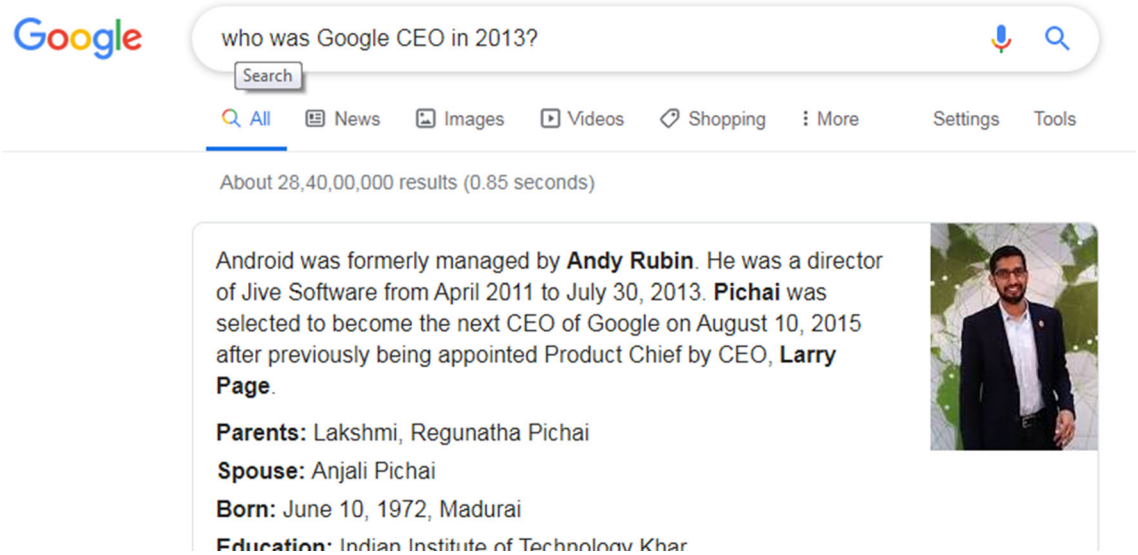

Fig. 7 Incomplete and Incorrect Knowledge ("screen taken in January 2020")

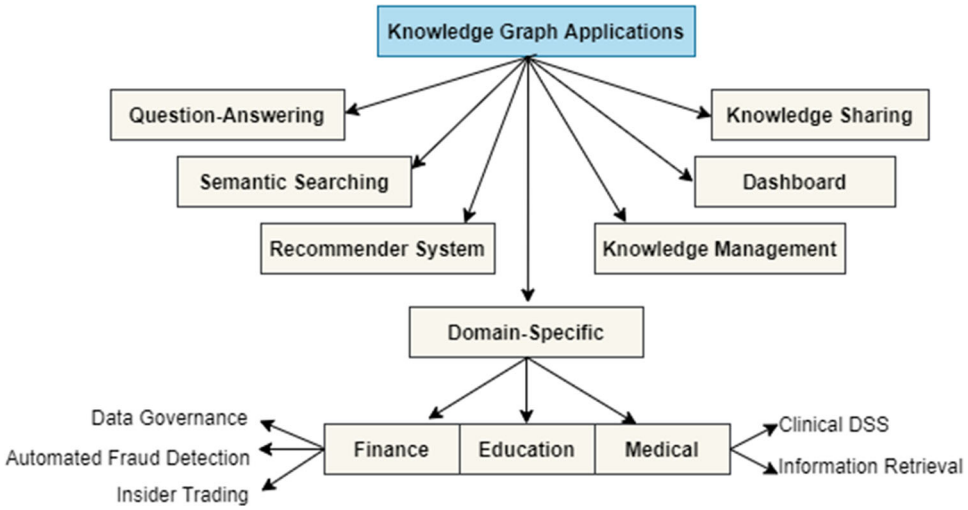

Fig. 8 Applications of Knowledge Graphs

and their relations. In this research, we have conducted a systematic review to present the characterization of existing KGs. This review provided the exploration and analysis of the state-of-the-art of KGs. Several sources are analysed to collect the related study and filtered by the relevant criteria. The comparison among prominent existing academia and industry KG systems are conducted, and then the discussion in detail about the KGs features, KGs constructing techniques, knowledge extraction methods. In addition, this review presents a close look at the challenges and problems of KGs. Moreover, this review presents the concepts of refinement and evaluation of KGs and the most significant used applications of KGs. Based on all the above, we can conclude that despite all advantages of existing KGs that have emerged as one of the most important systems for knowledge representation, organization, and understanding, but many KGs research issues, applications, and challenges require more efforts. In brief, we can present the key issues that may define the research directions of KGs in the near future, as follows: entity disambiguation, dynamic extraction, and management for knowledge, dealing with heterogeneous sources, discovering the semantic relations, inferring and verifying the facts and rules, protecting the privacy and security of KG, and converging and updating (freshing) in KG.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Compliance with Ethical Standards This study was not funded by any grant. No animals were involved. This article does not contain any studies with human participants or animals performed by any of the authors. Informed consent was obtained from all individual participants included in the study.

References

Abouenour L, Nasri M, Bouzoubaa K, Kabbaj A, Rosso P (2014) Construction of an ontology for intelligent Arabic QA systems leveraging the conceptual graphs representation. J Intell Fuzzy Syst 27(6):2869–2881

Abualigah LM, Khader AT, Hanandeh ES (2018) A novel weighting scheme applied to improve the text document clustering techniques. In: Innovative computing, optimization and its applications. Springer, Cham, pp 305–320

Abualigah LMQ (2019) Feature selection and enhanced krill herd algorithm for text document clustering. Springer, Berlin, pp 1–165

Al-Aswadi FN, Chan HY, Gan KH (2019) Automatic ontology construction from text: a review from shallow to deep learning trend. Artificial Intelligence Review 1–28

Angeli G, Manning CD (2013) Philosophers are mortal: Inferring the truth of unseen facts. In Proceedings of the seventeenth conference on computational natural language learning (pp. 133-142)

Arnold P, Rahm E (2014) Extracting semantic concept relations from wikipedia. In Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14) (pp. 1-11)

Baker CF, Fillmore CJ, Lowe JB (1998) The berkeley framenet project. In 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1 (pp. 86-90)

Banko M, Etzioni O (2008) The tradeoffs between open and traditional relation extraction. In Proceedings of ACL-08: HLT (pp. 28-36)

Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inf 41(5):706–716

Berners-Lee T (2006). Linked Data http://www.w3.org/DesignIssues.LinkedData.html

Berners-Lee T, Hendler J (2001) Publishing on the semantic web. Nature 410(6832):1023–1024

Bizer C, Heath T, Berners-Lee T (2011) Linked data: The story so far. In Semantic services, interoperability and web applications: emerging concepts (pp. 205-227). IGI Global

Bollacker K, Cook R, Tufts P (2007) Freebase: A shared database of structured general human knowledge. In AAAI (Vol. 7, pp. 1962-1963)

Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250)

Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka E, Mitchell T (2010) Toward an architecture for never-ending language learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 24, No. 1)

Chekol MW, Pirrò G, Schoenfisch J, Stuckenschmidt H (2017) Marrying uncertainty and time in knowledge graphs. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (pp. 88-94)

Chen Y, Kuang J, Cheng D, Zheng J, Gao M, Zhou A (2019) AgriKG: an agricultural knowledge graph and its applications. In International Conference on Database Systems for Advanced Applications. Springer, Cham, pp. 533–537

Chen Y, Li W, Liu Y, Zheng D, Zhao T (2010) Exploring deep belief network for chinese relation extraction. In: CIPS-SIGHAN Joint Conference on Chinese Language Processing

Culotta A, McCallum A (2005) Joint deduplication of multiple record types in relational data. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 257-258)

Davis R, Shrobe H, Szolovits P (1993) What is a knowledge representation? AI Mag 14(1):17

De Sa C, Ratner A, Ré C, Shin J, Wang F, Wu S, Zhang C (2016) Deepdive: declarative knowledge base construction. ACM SIGMOD Record 45(1):60–67

Dong Z, Dong Q (2003) HowNet-a hybrid language and knowledge resource. In International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 (pp. 820-824). IEEE

Dong X, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, Strohmann T, Sun S, Zhang W (2014) Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 601-610)

Etzioni O, Cafarella M, Downey D, Popescu AM, Shaked T, Soderland S, Weld DS, Yates A (2005) Unsupervised named-entity extraction from the web: an experimental study. Artif Intell 165(1):91–134

Etzioni O, Banko M, Soderland S, Weld DS (2008) Open information extraction from the web. Commun ACM 51(12):68–74

Färber M, Bartscherer F, Menne C, Rettinger A (2018) Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. Sem Web 9(1):77–129

Ferre S (2019, June). Link prediction in knowledge graphs with concepts of nearest neighbours. In European Semantic Web Conference (pp. 84-100). Springer, Cham

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174

Gaurav D, Tiwari SM, Goyal A, Gandhi N, Abraham A (2020) Machine intelligence-based algorithms for spam filtering on document labeling. Soft Comput 24(13):9625–9638

Hakkani-Tür D, Heck L, Tur G (2013) Using a knowledge graph and query click logs for unsupervised learning of relation detection. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 8327-8331). IEEE

Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. In Coling 1992 volume 2: The 15th international conference on computational linguistics

Heck L, Hakkani-Tür D, Tur G (2013) Leveraging knowledge graphs for web-scale unsupervised semantic parsing

Heist N (2018) Towards knowledge graph construction from entity Co-occurrence. In EKAW (Doctoral Consortium)

Hoffart J, Suchanek FM, Berberich K, Weikum G (2013) YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. Artif Intell 194:28–61

Jia Y, Qi Y, Shang H, Jiang R, Li A (2018) A practical approach to constructing a knowledge graph for cybersecurity. Engineering 4(1):53–60

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

Kambhatla N (2004) Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (pp. 22-es)

Keele S (2007) Guidelines for performing systematic literature reviews in software engineering (Vol. 5). Technical report, Ver. 2.3 EBSE Technical Report. EBSE

Klyne G, Carroll JJ, McBride B (2004) Resource description framework (RDF): concepts and abstract syntax. W3C Recommendation, Feb. 2004

Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, Van Kleef P, Auer S, Bizer C (2015) DBpedia-a large-scale, multilingual knowledge base extracted from Wikipedia. Sem Web 6(2):167–195

Li L, Wang P, Yan J, Wang Y, Li S, Jiang J, Sun Z, Tang B, Chang TH, Wang S, Liu Y (2020) Real-world data medical knowledge graph: construction and applications. Artif Intell Med 103:101817

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019–1031

Lin Y, Han X, Xie R, Liu Z, Sun M (2018) Knowledge representation learning: A quantitative review. arXiv preprint arXiv:1812.10901

Liu Z, Han X (2018) Deep learning in knowledge graph. Springer, Singapore

Liu H, Singh P (2004) ConceptNet-a practical commonsense reasoning tool-kit. BT Technol J 22(4):211–226

Matuszek C, Witbrock M, Cabral J, DeOliveira J (2006) An introduction to the syntax and content of Cyc. UMBC Computer Science and Electrical Engineering Department Collection

Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38(11):39–41

Minsky M (1974). A framework for representing knowledge

Mishra S, Jain S (2019) An intelligent knowledge treasure for military decision support. Int J Web-Based Learn Teaching Technol (IJWLTT) 14(3):55–75

Momtchev V, Peychev D, Primov T, Georgiev G (2009) Expanding the pathway and interaction knowledge in linked life data. Proc. of International Semantic Web Challenge

Nakashole N, Theobald M, Weikum G (2011) Scalable knowledge harvesting with high precision and high recall. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 227-236)

Newcombe HB, Kennedy JM, Axford SJ, James AP (1959) Automatic linkage of vital records. Science 130(3381):954–959

Newman ME (2001) The structure of scientific collaboration networks. Proc Nat Acad Sci 98(2):404–409

Nicholson DN, Greene CS (2020) Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J 18:1414

Nickel M, Murphy K, Tresp V, Gabrilovich E (2015) A review of relational machine learning for knowledge graphs. Proc IEEE 104(1):11–33

Niu X, Sun X, Wang H, Rong S, Qi G, Yu Y (2011) Zhishi. me-weaving chinese linking open data. In International Semantic Web Conference (pp. 205-220). Springer, Berlin, Heidelberg

Noy N, Gao Y, Jain A, Narayanan A, Patterson A, Taylor J (2019) Industry-scale knowledge graphs: lessons and challenges. Queue 17(2):48–75

Paulheim H (2017) Knowledge graph refinement: a survey of approaches and evaluation methods. Sem Web 8(3):489–508

Rahm E, Bernstein PA (2001) A survey of approaches to automatic schema matching. VLDB J 10(4):334–350

Rahul M, Kohli N, Agarwal R, Mishra S (2019) Facial expression recognition using geometric features and modified hidden Markov model. Int J Grid Util Comput 10(5):488–496

Ringler D, Paulheim H (2017) One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & co. In Joint GermanAustrian Conference on Artificial Intelligence (Künstliche Intelligenz) (pp. 366-372). Springer, Cham

Ruttenberg A, Rees JA, Samwald M, Marshall MS (2009) Life sciences on the Semantic Web: the Neurocommons and beyond. Brief Bioinf 10(2):193–204

Saïs F (2019). Knowledge Graph Refinement: Link Detection, Link Invalidation, Key Discovery and Data Enrichment (Doctoral dissertation, Université Paris Sud)

Sengupta S (2013) Facebook unveils a new search tool. NY Times, New York

Singhal A (2012) Introducing the knowledge graph: things, not strings. Official google blog, 5

Sowa JF (2006) Semantic Networks [Electronic resource]. Access mode: http://www.jfsowa.com/pubs/semnet.htm

Suchanek FM, Kasneci G, Weikum G (2007) Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (pp. 697-706)

Suchanek FM, Sozio M, Weikum G (2009) SOFIE: a self-organizing framework for information extraction. In Proceedings of the 18th international conference on World wide web (pp. 631-640)

Suh B, Convertino G, Chi EH, Pirolli P (2009) The singularity is not near: slowing growth of Wikipedia. In Proceedings of the 5th International Symposium on Wikis and Open Collaboration (pp. 1-10)

Sun Y, Han J (2012) Mining heterogeneous information networks: principles and methodologies. Synth Lect Data Mining Knowl Discov 3(2):1–159

Tejada S, Knoblock CA, Minton S (2001) Learning object identification rules for information integration. Inf Syst 26(8):607–633

Tiwari SM, Jain S, Abraham A, Shandilya S (2018) Secure Semantic Smart HealthCare (S3HC). J Web Eng 17(8):617–646

Tiwari S, Abraham A (2020) Semantic assessment of smart healthcare ontology. International Journal of Web Information Systems

Vrandecic D (2012) Wikidata: a new platform for collaborative data collection. In Proceedings of the 21st international conference on world wide web (pp. 1063-1064)

Wang J, Liu J, Kong L (2017) Ontology construction based on deep learning. In Advances in Computer Science and Ubiquitous Computing Springer, Singapore

Wang P, Jiang H, Xu J, Zhang Q (2019) Knowledge graph construction and applications for Web search and beyond. Data Intell 1(4):333–349

Wang Z, Li J, Wang Z, Li S, Li M, Zhang D, Shi Y, Liu Y, Zhang P, Tang J (2013) XLore: A Large-scale English-Chinese Bilingual Knowledge Graph. In International semantic web conference (Posters & Demos) (Vol. 1035, pp. 121-124)

Wu T, Qi G, Li C, Wang M (2018) A survey of techniques for constructing Chinese knowledge graphs and their applications. Sustainability 10(9):3245

Wu W, Li H, Wang H, Zhu KQ (2012) Probase: A probabilistic taxonomy for text understanding. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 481-492)

Wu T, Wang H, Li C, Qi G, Niu X, Wang M, Li L, Shi C (2019) Knowledge graph construction from multiple online encyclopedias. World Wide Web 1–28

Xu B, Xu Y, Liang J, Xie C, Liang B, Cui W, Xiao Y (2017) CN-DBpedia: a never-ending Chinese knowledge extraction system. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 428-438). Springer, Cham

Yan J, Wang C, Cheng W, Gao M, Zhou A (2018) A retrospective of knowledge graphs. Front Comput Sci 12(1):55–74

Zhang J, Liu J, Wang X (2016) Simultaneous entities and relationship extraction from unstructured text. Int J Database Theory Appl 9(6):151–160

Zhang Z, Zhuang F, Qu M, Lin F, He Q (2018) Knowledge graph embedding with hierarchical relation structure. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3198-3207)

Zhao M, Wang H, Guo J, Liu D, Xie C, Liu Q, Cheng Z (2019) Construction of an industrial knowledge graph for unstructured Chinese text learning. Appl Sci 9(13):2720

Zhong B, Liu J, Du Y, Liaozheng Y, Pu J (2016) Extracting attributes of named entity from unstructured text with deep belief network. Int J Database Theory Appl 9(5):187–196

Zhu G, Iglesias CA (2015) Sematch: Semantic Entity Search from Knowledge Graph. In SumPre-HSWI@ ESWC

Zhu J, Nie Z, Liu X, Zhang B, Wen JR (2009) Statsnowball: a statistical approach to extracting entity relationships. In Proceedings of the 18th international conference on World wide web (pp. 101-110)

Zou X (2020) A survey on application of knowledge graph. JPhCS 1487(1):012016

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.