Conceptual Extraction of Domain Knowledge Graph in Different Data Sources

Hui-zhen BIAN¹ and Si HA¹,*

¹School of Computer Science and Technology Inner Mongolia Normal University
Hohhot, China
*Corresponding author

Keywords: Data source, Domain knowledge graph, Concept extraction.

Abstract. With the continuous evolution and the Web technology development, the knowledge graphs are favored by various fields with theirs powerful semantic processing ability and open interconnection ability. All the industries and universities have built their own areas of knowledge graph, how to extract the concepts from different data sources becomes the key technology of knowledge graph construction, the more complete the concept extraction, the more comprehensive the knowledge graph, the higher the value of the use. This paper will illustrate the conceptual extraction methods of knowledge graph, in order to guide scholars to choose reasonable methods for academic analysis and enhance the application level of knowledge graph.

Introduction

The knowledge graph is first proposed by Google, and its purpose is primarily to study the various entities and concepts that exist in the real world, as well as the relationships between these entities and concepts. The research work of foreign knowledge graph has been carried out for many years, and it has achieved fruitful results. It also produces a large number of excellent knowledge graphs, including DBpedia, YAGO [1], FreeBase, and Google Knowledge Graph and so on. These knowledge graphs have played a significant role in the field of knowledge search and natural language processing. The construction of Google's knowledge Graph is based on the knowledge graph to create a new intelligent search mode. There are many teams in the country are doing this research work. Such as XLore, Zhishi.me and so are excellent Chinese knowledge graph projects currently. The difference between the knowledge graph and the ontology is that the former has enriched and expanded the latter at the entity level. The emphasis on the Ontology is the semantic relation between the concepts during the construction process, which lays the foundation for the data pattern of the knowledge graph. Knowledge graphs are widely used in data mining, intelligent question and answer, semantic search, knowledge engineering and other fields. Based on the existing data source, the most likely and accurate to extract the concepts for building domain knowledge graph [2], which is the first and the most important step.

In this paper, we summarize the method of conceptual extraction of knowledge graph from unstructured data, semi-structured data and structured data, taking advantages of the different structural features of these datum to ensure the coverage and the quality of the knowledge graph as much as possible.
**Conceptual Extraction of Unstructured Data**

Unstructured data generally refers to the data which has no fixed structure, unstructured data in this paper refers to pure document, and document information widely exists in the Internet, which is one of the most important sources for knowledge graph. Because of the non-specificity of the structure, the machine can’t interpret the semantic information it contains, it is necessary to acquire the knowledge from the unstructured text generally base on the natural language processing (NLP) technique, including the word segmentation, Partnered annotations, named entity recognition and syntactic analysis, and then rely on statistical analysis, machine learning and other methods to obtain knowledge.

Text-to-Onto is a relatively early ontology learning system, and using weighted word frequency statistical method to obtain concepts first. Shamsfard and Barforoush proposed a domain-independent concept extraction method based on a simple ontology kernel, and then use textual understanding technologies to extract the concept of ontology automatically. Lee et al. proposed an ontology-based ontology construction method to learn the domain ontology from unstructured text files, in which the clustering and classification of concepts mainly use concept similarity calculation based on fuzzy numbers.Navigli and Velardi proposed a text-based domain ontology learning tool, OntoLearn [3], which automatically acquires concepts and conceptual relationships from the website's text. The most important feature is using semantic interpretation to associate a complex concept with a complex term; then using WordNet as a universal ontology, the extracted terms are linked to the concepts in WordNet to explain the taxonomy of terms and determine which terms can be classified into concepts.

The conceptual extraction method based on text (unstructured data) mainly includes the following methods.

**Conceptual Extraction Based on Linguistics**

The conceptual extraction method based on linguistics uses the pre-established rules to extract the concepts. The method requires the lexical structure in the field corpus must be analyzed accurately, and the lexical rules can be summarized, and next create matching templates based on summarized rules. Then use these templates to match the domain text to extract the concepts. This method depends on the specific language to establish the corresponding grammar rules, it is necessary to study and deal with the language environment of domain corpus.

Because conceptual extraction based on linguistics is based on linguistic rule that is analyzed in a large number of domain texts, it is relatively simple to match the text according to the acquired rules, the calculation is small, the accuracy of the concept extraction is higher, and the words with the lower frequency can also be extracted. Meanwhile, the advantage of this method is that the extraction process does not depend on the underlying document which is used to establish the rules.

However, this method requires a lot of resources in the construction of the linguistic rules. The rules are very easy to affect the results of extraction. The maintenance of the rules is difficult, the language is highly dependent, and the established rules can’t be used for any fields, for different fields need to build different grammar rules, poor portability.
Conceptual Extraction Based on Statistical

Based on the statistical extraction method [4] according to different statistical characteristics between the domain concepts and the general vocabularies in the document set to identify the concept of the target field. Statistical-based extraction is easier to implement than linguistic methods, and is not limited to a particular domain, which is the dominant conceptual extraction method currently. In general, relatively high-frequency words within the target domain document represent words and phrases in the domain features.

The method of statistical-based extraction does not need to be based on specific semantic information, not limited to a specific field, not limited to language type and sentence structure. It is easy to expand. But this method needs a lot of field documents to support, large amount of calculation. The extraction effect of the low frequency term is poor, and the accuracy of concept extraction is generally not up to the expected value.

The conceptual extraction method based on the relevance semantic and semantic consistency is a typical concept-based concept extraction method, which uses a non-target document set to filter out a large number of terms that are independent of the target domain.

Conceptual Extraction Based on Mixing Method

In the actual knowledge graph construction process, the statistical method is used to combine with the linguistic method to carry out the concepts extraction, the rule matching is adopted to get the candidate concepts and then the statistical method is used to filter, or the statistical method is used to obtain the candidate concepts, linguistic method to establish rules to filter the concept of the domain, in order to the extraction of concepts with high accuracy and high coverage.

Conceptual Extraction of Structured Data

Structured data is an important data source of knowledge graphs. Through careful study, it is found that a large number of datum in the information system are stored in the relational database in the form of structured data sources. Therefore, the concepts needed for domain knowledge graphs can be automatically extracted from the database.

Structured Information Extraction in Web Pages

With explosive growth of internet, huge amounts of information is presented by structured data, which are generated from background databases, and displayed in a semi-structured form attributes and their associated values for an entity[5].These online structured data make the internet a big pool of information, which provides new opportunities for revolution of some crucial applications, including automatic semantic annotations, knowledge database establishments and further information utilizations[6].

There are many existing methods for structured data extraction from web pages. They are manually tailored wrappers in early days, and later wrapper learning techniques. Complex generative models, bootstrapping, interactive learning and probabilistic graphical model are also developed. There are also unsupervised template detection approaches, such as RoadRunner [7] and EXALG. Recently, a unified solution [8] and a Bayesian approach [9] are proposed.
Conceptual Extraction Method of Artificially Constructed Ontology or Dictionary

Artificially constructed ontologies or dictionaries are machine-readable and are generally regarded as structured data which have good practicality and reliability. WordNet [10], Cyc and OpenCy are the most common ontologies used in ontology construction. These generic ontologies sometimes are used for direct data sources for conceptual extraction, sometimes combined with other data sources, sometimes as the initial data for the startup process. In some cases, it is also used to evaluate the learning outcomes.

Here, we take WordNet [11] as an example to illustrate conceptual extraction of constructing knowledge graphs based on open ontologies or dictionaries.

WordNet, as a general ontology, covers almost all concepts of all different aspects and the relationship between the concepts, extracting the required concepts and relationships from WordNet make the full use of resources, and compared to technologies that extract concepts[12] and relationships from domain documents or the internet, reducing the information retrieval, part of the annotation and other related operations, saving a lot of time and improve the efficiency of the knowledge graph construction. It is generally believed that the concepts extracted from WordNet is divided into the following steps:

- The first step: the domain to be built as the “target domain”, because the same domain concepts in WordNet semantic tree are not continuous, the target domain subdivided out of the field known as the" sub-domain. "For the target domain, according to its actual situation, combined with domain books, Wikipedia and field-related workers, divides it into a number of sub-domain (based on manual completion).
- The second step: Find the core concepts of each the sub-domain.
- The third step: For each sub-domain, according to the core concept, through a certain semantic similarity algorithm, access to the top concept of each sub-domain.
- The fourth Step: Based on the top-level concept, from WordNet extracts the concepts contained in each sub-domain from the WordNet, so get the sub-domain "initial concept set."
- The fifth step: remove the inappropriate concepts from initial concepts of each sub-domain, to obtain the concepts needed for each sub-domain, and to obtain the “final conceptual set of sub-domains”.

The specific steps are shown in Figure 1:

![Figure 1. Flow chart for extracting concepts from WordNet.](image-url)
Conceptual Extraction of Semi-structured Data

Semi-structured data is characterized by a certain implicit structure, but lack of strict and fixed data model. A large number of Web documents in the HTML format and some documents in the form of XML are the most representative semi-structured data. For this kind of semi-structured data should make full use of the structured information, for which the unstructured data, a text-based concept extraction method can be used. Sanchez and Moreno have developed a new technique, which use search engines to obtain web pages related to the initial keywords. Then, based on the frequency of the concept extract the new concepts and the relationships between concepts. Papatheodorou et al. proposed a taxonomic relationship between the concepts in the XML format. First, the keywords that represent the contents of the document are extracted from the document, and then the clustering technique is used on the basis of these keywords. Divide the documents into different groups according to the similarity of the contents, and then extract the concept and the relationship between the concepts.

Semi-structured data mainly exists in the encyclopedia [13], it is generally believed that from the encyclopedia to obtain the concepts as follow:

Classification in Encyclopedia as a Candidate for Concept

Firstly, the class labels in the classification system are directly regarded as concepts; Encyclopedia classification system is basically checked by encyclopedia administrators and senior editors, so it has very high reliability.

Secondly, other categories of labels are also candidates for the concepts, but can’t be directly selected as a concepts, because there are unreasonable categories in the encyclopedia, including: (a) empty category labels, an empty category label refers to a category that has neither a subcategory nor an article that belongs to it; for such category labels, they are treated directly as entities (b) contain only their own category labels, which do not have subcategories and contain article pages that are authored by themselves; Likewise, these labels are considered to be entities directly, and their parent categories are set to the concept of entities.

Thirdly, the concepts can also be extracted from the hypernym-hyponym relationships, when the final relationship between the hypernym-hyponym relationships organized into a classification system, not at the bottom (leaf nodes) are regarded as the concept.

The Title as the Primary Concept, Relevant Concepts Serve as Candidate Concepts for Concept Libraries

Download the latest corpus collection, the explanatory documentation for all Encyclopedia concepts are included in the <doc> </doc> tags. The document fragments in the <doc> </doc> tag contains information about the abstracts and link information.

Before the concept is extracted, the title (the main concept) should be searched and scored in the Encyclopedia page to improve the quality of the knowledge graph. A large number of studies show that the relevant concepts of Encyclopedia are presented through the main concept’s hyperlinks. The linking concepts that appear in the main concepts pages’ body as the source of relevant concepts. The semantic similarity calculation method is used to calculate the similarity between the title (the main concept) and the relevant concept respectively. If the candidate concept does not have the explanatory fragment in the corpus, there is no corresponding abstract, think the similarity is zero.
After the logical judgment of the main concept of the (title), the relevant concepts are extracted by hyperlinks, and a series of relevant candidate concepts can be obtained in the preprocessing module. However, there are still some concepts related to the topic in the extracted candidate concepts, which have little relevance, for these weak correlation concepts, the human brain can judge by thinking ability, but the machine can not automatically make strong or weak judgments, it is generally believed that there are two matching patterns for the deep screening to candidate relevant concepts.

1) Conceptual screening based on linguistic features (part of speech and word frequency)

The weak correlation concepts are graded according to the part of the relevant concepts and the frequency in the text to determine whether it can be added to the domain concept library. We can also construct a noun phrase analysis method to determine whether the candidate concept can be used as a concept. The linguistic feature-based approach can improve the accuracy, but in the face of massive data, the computational complexity is high and the portability is poor.

2) Conceptual screening based on existing knowledge base

For the weak correlation concepts, we can calculate the correlation and similarity between the main concepts respectively, through the design algorithm in the existing knowledge base, and add the value to the domain concept base if the correlation and similarity reach the set threshold.

Summary

Knowledge graphs are widely used in semantic retrieval, data mining, artificial intelligence, knowledge organization and intelligent question and answer [14]. How to effectively organize and utilize various types of datum such as structured data, semi-structured data and unstructured data in the background of the current Internet big data, making these datum better services for knowledge graphs have become a new challenge in the era of big data.

This paper explores the concept extraction of knowledge graphs, summarizes and proposes some extraction methods of concepts, and discusses the method of extracting concepts from different data sources.

The construction of the knowledge graph is a huge and difficult task. The work of this paper only discusses the method of concept extraction. There are still many research work, such as the design method of the algorithm [15].

Acknowledgment

This paper was financially supported by National Science Foundation of China: "Research on Mongolian vocabulary and WordNet (No.61363035)". At the same time, I would also like to thank the scholars cited in this paper.

References


