ISSN: 2664-9640

# Research on Search Engine based on Knowledge Graph

Zixian Gong

Xiamen University Malaysia. Kuala Lumpur, Malaysia

MAT1909400@xmu.edu.my

## Abstract

**This paper mainly introduces the current popular search engine technology based on knowledge graphs, and covers the development of search engines and knowledge graphs. It focuses on the main structure of the search engine based on the knowledge graph, namely the three modules, and introduces the main functions of the three modules and the application of related technologies.**

## Keywords

**Search Engine, Knowledge Graph, Web Crawler.**

## 1. Introduction

### 1.1. Background and Significance of the Search Engine

While the explosive growth of digital information brings convenience to users, it also brings a lot of trouble. How to accurately find the exact information needed is as difficult as finding a needle in a haystack. The search engine returns to the user related web pages based on the user's search content, which brings great convenience to the retrieval of Internet information. However, the results returned by search engines cannot provide accurate information. Therefore, if search engines can retrieve the most important information faster and more conveniently based on the semantic relationship between related entities, it can bring convenience to users to a large extent. Therefore, search engines based on knowledge graphs have become a research trend.

### 1.2. Research Status and Development Trends

1.2.1 Research Status and Development Trends Of Search Engine

The first-generation of search engines is mainly based on the World Wide Web. The search efficiency of this search engine is based on the number of search results as the only assessment standard, but it has a low efficiency.

The second-generation search engine is mainly based on the data capture of web crawler robots and the establishment of hyperlink analysis, so that the updated content of web pages can be fed back faster, which also improves the retrieval efficiency of search engines.

A third-generation search engine that can comprehensively analyze the web pages of the World Wide Web and more in-depth data mining is proposed.

After this, Google was the first to propose a new generation of search engine based on knowledge graphs. The main idea of this is: grab data and extract knowledge fragments and integrate knowledge fragments to generate knowledge that can represent entities, the semantic relationships between entities constitute a knowledge network.

1.2.2 Research Status and Development Trends of Knowledge Graph

Building a knowledge graph in a field is a very important task. At present, many information extraction systems such as NELL[1], Open-IE are used in related research. Google uses a lot of technology to extract knowledge fragments from the web, and merge them to form facts,

ISSN: 2664-9640

concepts are entity knowledge. Different entities and complex semantic relationships constitute the knowledge graph.

Knowledge graphs have been widely used in many fields, the most important of which is related to search engines. It is very common for an enterprise to have massive physical data, and how to use this information efficiently has also become a problem. The generation of knowledge graphs and the maturity of related technologies have also brought a turning point to the above-mentioned problem.

## 2. Related Technologies about Search Engine based on Knowledge Graph

### 2.1. The Structure of the Search Engine based on Knowledge Graph

The structure of the Search engine based on Knowledge graph is mainly divided into 3 modules: Search Module, Crawler and Index Module, Knowledge Graph Module. Among them, Search Module combines the ontology library Query to parse the searched sentence to get the keywords, and can get the result through the Indexer and return it to the user; The Crawler and Index Module is mainly responsible for crawling data from the network, and can parse the entity-related information obtained by the Search Module and use it to build an index; The remaining Knowledge Graph Module can be based on entities obtained by Crawler and Index Module to merge and align the entities to obtain all related entities, and then the Knowledge Graph Module can give the final ranking result of related information. In this part, we will introduce these three modules.

### 2.2. Search Module

2.2.1 Query Analysis

Query Analysis is responsible for extracting key information from the user's Query in the entire search system to understand what the user hopes to search through this Query. This requires the following tasks: Error correction and rewriting, it is necessary to correct and supplement the user's input errors or not input complete content. This is also because the underlying database can only support accurate search, so the content of Query needs to be changed and rewritten into complete and correct sentences; Intention recognition, to identify the main search intent, which is equivalent to guiding the downstream in which database should be searched; Entity recognition, which is similar to the previous intention recognition, but its granularity is more fine-grained, specific to the analysis of specific words in Query, and extract key entities from Query. As mentioned above, intention recognition is to tell downstream which database should be used to do a search, then entity recognition is to tell the downstream which key words should be searched in this database.

### 2.3. Crawler and Index Module

2.3.1 Crawler Technology

Web crawler can automatically collect all the pages that can be visited, and it can capture data according to the specified scope to facilitate further processing by search engines, which also enables users to retrieve the information they want in a short time [2].

The crawler is mainly divided into four parts: Protocol processor, Content detection, URL extraction, and URL processor. The Protocol processor is responsible for processing related protocols. Content detection mainly completes the extraction of the required information from the webpage obtained by the URL and performs detection. URL extraction is mainly responsible for extracting URLs that are still crawling, so that the crawler can continue; URL processor is responsible for sorting the URLs that have been extracted.

In addition, the processing flow of the web crawler is to use one or several URLs as the initial URL, and then start web page crawling. And continuously add the extracted new URL to the

ISSN: 2664-9640

initial URL to form a URL list, and finally, if the crawler completes all the conditions, it will get this URL list. If any of the conditions are not met, use this URL list as the initial URL to start a new round of page crawling, and loop until all the conditions are met, finally get the desired result. Among them, there are three main strategies for page crawling: Breadth first strategy, Depth first strategy, Optimal choice strategy.

Breadth first strategy: The crawler will grab all the links in the web page obtained according to the initial URL, and then compare it with the initial page to get the closest URL, and then continue the previous operation. This method occupies a larger amount of memory, because each initial URL before it is stored there and not discarded, which is why the processing speed of the Breadth first strategy will be faster.

Depth first strategy: This method will extract the URLs on the corresponding webpage based on the initial URL, select one of them as the new initial URL and perform the next round of crawling, and repeat the process. Depth first strategy can find all relevant information, and it will delete some of the duplicate nodes when it generates successor nodes, so its memory is relatively smaller. Its disadvantage is that the importance of links quickly decreases every time it goes deep, so in the case of great depth, the efficiency of Breadth first strategy is not high.

Optimal choice strategy: This strategy is to calculate the similarity between the candidate URL and the target webpage based on a specific analysis algorithm. Then select one or several URL with higher similarity. This method is a locally optimal strategy [3], so some important web pages may be ignored.

## 2.4. Knowledge Graph Module

2.4.1 Knowledge Graph Model

Knowledge graph mainly includes Web entity mining and entity data processing, entity expression, knowledge storage, data analysis interface [4] and other modules. Web entity mining refers to the fusion of relevant knowledge fragments obtained by extraction into knowledge; Entity data processing is mainly responsible for entity alignment by merging the knowledge fragments that describe the same content into the whole knowledge; Entity expression module, this part is mainly to conduct structured representation of entity knowledge. Knowledge storage module is to complete the storage of the entity. The data analysis interface module is responsible for searching, filtering, sorting, and other important work of the content.

# 3. Conclusion

It is very important to search for information quickly and accurately. After Google proposed the knowledge graph and began to apply the knowledge graph to search engines, more and more companies began to use this technology. Moreover, under multi-party research, this technology will become more mature.

# References

[1] Carlson A, Betteridge J, Kisiel B, et al. Toward an Architecture for Never-Ending Language Learning[C]//AAAI. 2010, 5: 3

[2] Xu Zenglin, Sheng Yongpan, He Lirong, et al. Overview of Knowledge Graph Technology [J]. Electronics Journal of University of Science and Technology, 2016, 45(4): 589-606

[3] Liu Jinhong, Lu Yuliang. A Survey of Topic Web Crawler Research[J]. Journal of Computer Research, 2007, 24(10): 26-29

[4] Ding Yishan, Du Yanhui, Zhu Yancheng, etc. Research on keyword extraction technology based on knowledge graph [J]. Software Guide, 2020,19(2):273-277.