ATLANTIS PRESS
Advances in Engineering Research, volume 162
3rd International Conference on Advances in Materials, Mechatronics and Civil Engineering (ICAMMCE 2018)

Information Retrieval Technology Based on Knowledge Graph

Wang Ce[1,a], Yu Hongzhi[2,b] and Wan Fucheng[3,c,*]

[1]Key Laboratory of National language intelligent processing, China, Gansu Province, Lanzhou,730030

[2]Key Lab of China's National Linguistic Information Technology，China,Lanzhou,730030

[3]Key Laboratory of National language intelligent processing, China, Gansu Province, Lanzhou,730030

[1]852876572@qq.com, [2]wanfucheng@126.com, [3]306261663@qq.com

*Wan Fucheng

Keywords: Knowledge Graph, Map Retrieval, Automatic Construction, Smart Recommendation

Abstract. In the big data environment with the rapid development of the Internet, the dependence on information search is becoming stronger and stronger. At present, full text search based on keywords has been difficult to satisfy people's search needs. In this case, an information retrieval method based on knowledge graphs is proposed. Through a self-supervised open Chinese relation extraction method, the knowledge of knowledge graphs is extracted from large-scale unstructured data in the Internet, and the knowledge graphs are constructed based on related domain knowledge bases. Based on the knowledge graph, information retrieval is performed through the calculation of semantic similarity. Using this technology for information retrieval, the efficiency and accuracy of the retrieval results will be greatly improved, and it has a very good application value in the field of information retrieval and smart recommendation.

1. Preface

With the rapid spread of the Internet, the rapid increase in the volume of digital information data has brought us a wealth of valuable information data. Although these data have been classified and managed, effective information has been retrieved from thousands of data for search engines. It is also a great challenge.

In the age of big data, the massiveness, heterogeneity, dynamics, and diversity of Web data have become the major challenges facing information retrieval. The traditional information retrieval is to index the content of the main webpage through the keyword searched by the user and feed back the relevant webpage link to the user based on the keyword in the matching user's search request. This search mode brings great convenience to Internet information retrieval. However, this model has a big drawback. That is, the results returned by the search engine are in a single form. It is impossible to directly provide accurate information based on the user's search request. The user still needs to continue to search for the required information in the web page according to the provided link.

In response to this problem, we propose an information retrieval technology based on knowledge graphs. This technology implements entities by further performing entity information mining on Web page content and through an open-source relational extraction method based on self-supervised learning in machine learning. The extraction of the synonymous relationship, the upper-lower relationship, and the attribute relationship between (concepts), the construction of knowledge graphs through the relationship between entities. The information retrieval based on the knowledge graph is to construct the relationships between the entities and index these data and relationships. At the same time, the use of the entity-based retrieval tools in the semantic aspects is timely. The application of the knowledge graph in the information retrieval makes the search engine more efficient. A good understanding of the needs of users, and can provide users with more intelligent, accurate, humane results.

2. Based on Wikipedia knowledge graph construction

2.1 Extract concepts and entities from encyclopedias

Concepts and entities are extracted from the encyclopedia. The title of the article in the encyclopedia is generally used as a candidate for the entity, and the classification in the encyclopedia is used as a concept candidate.

(1) The category labels located in the classification system in the encyclopedia will be directly included as concepts;

(2) Other categories of labels are also candidates for concepts, but they cannot be directly selected as concepts because there are some unreasonable categories in the encyclopedia, such as :(a) empty category labels, which are directly selected by these category labels as entities; (b) Contains only its own category tags, which are considered directly entities and whose parent category is selected as an entity. After the screening of these two steps, the reliability of the remaining categories is relatively high, but it still needs further confirmation.

(3) The concept can also be extracted from the relationship between the upper and lower positions. When the resulting upper and lower positions are organized into a classification system, not the lowest level can be considered as a concept.

2.2 Entity Alignment (Synonymy Learning)

In this paper, the goal of entity alignment of the knowledge graph is to merge the entities learned from the three types of encyclopedias, and then merge the merged entities with the entities extracted from the open link data.

This section focuses on entity alignment methods based on structured data in encyclopedias and self-supervised entity alignment methods based on SVM.

2.2.1 Encyclopedia entity alignment based on encyclopedia structured data

Alignment of entities in the same encyclopedia mainly depends on two types of structured data in the encyclopedia, redirect pages and information modules.

The redirection mechanism is used when the user accesses the same article using different input conditions, the system will automatically locate the unique article representing the corresponding entity of the article, so that the different articles representing the same thing in the current encyclopedia are merged, that is, The so-called article alignment; use the redirection for entity alignment, you need to traverse all pages, if it includes a redirection tag, the corresponding article title and the target article title that is redirected are recorded and marked as the same entity.

In addition, the information module also contains some synonymous information. Although these same information are in the same page, if the synonym entities described therein correspond to multiple articles, the entities corresponding to these articles also need to be merged.

2.2.2 Self-supervised encyclopedia entity alignment based on SVM algorithm

After the candidate entity is determined, we need to determine the entity. For two candidate entities, they may or may not be synonymous entities, which is a simple classification problem. Therefore, we use the machine learning classification algorithm to solve this problem. The model used is Support Vector Machine.

The SVM model, like the perceptual model, separates the positive and negative classes for finding the optimal hyperplane in the n-dimensional space. The best attained here is that the closest distances from the hyperplanes of the two types of sample points are the maximum, the interval is the largest and it is different from the perceptual model, and the SVM also has the nuclear skills, so the SVM is the actual nonlinear classifier function. Assuming a training data set given on a feature space

$$\mathrm{T} = \{(x_1, y_1),\ (x_2, y_2),\ (x_3, y_3),\ \ldots,\ (x_i, y_i)\}\ . \tag{1}$$

$x_i \in \mathrm{X} = R^n$, $y_i \in \mathrm{Y} = \{+1,\ -1\}$, $\mathrm{i} = 1,2,3,\ldots,\mathrm{N}$, representing N sample instances, x_i For the ith feature vector, y_i for x_i Class tag.

The goal is to find a separate hyperplane, dividing the positive and negative classes on both sides of the plane. Separation of hyperplane corresponding equations $\mathrm{w} \cdot \mathrm{x} + \mathrm{b} = 0$ When the data set is

linearly separable, there are multiple such hyperplanes. The perceptron uses the misclassified points to solve, and there are an infinite number of solutions. The SVM uses the interval maximization to find the optimal hyperplane and the solution is unique. Let the classification decision function be

$$f(x) = \text{sign}(\text{w} \cdot \text{x}) + b \ . \tag{2}$$

As shown in the figure, a linear classifier is needed to separate the black and white points. Obviously, straight lines H2 and H3 can complete this partitioning task, but H3 is the most suitable line, because according to this boundary, it is closest to the boundary. The black points and white points (called Support Vectors) are the sum of the distances from the boundary to the minimum. This partitioning process uses a mathematical expression as shown in Equation (3). In a training set D,

$$\text{D} = \{(x_i, y_i) | x_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^{n} \ . \tag{3}$$

X_i is a p-dimensional vector,y_iIt can be 1 or -1 and represents two different categories. If this data set is linearly separable, then it is certain that two categories of linear edges can be found.

$\text{w} \cdot \text{x} - \text{b} = 1$ with $\text{w} \cdot \text{x} - \text{b} = -1$ The data points in the two categories are completely separated; thus, a classifier equidistant from both edges can be defined based on the two edges.$\text{w} \cdot \text{x} - \text{b} = 0$, as shown on the right, where w represents a new vector,$\frac{b}{\|w\|}$ represents the distance from the edge to the splitter.

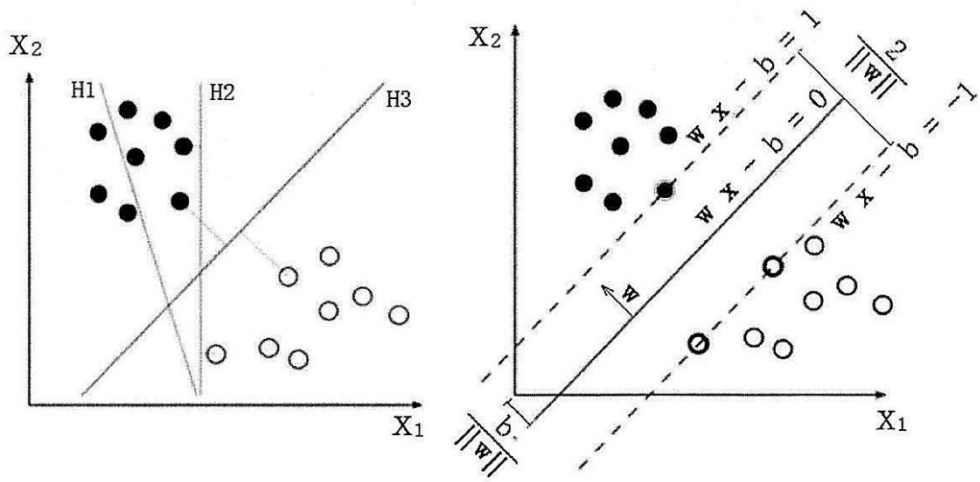

Fig. 1. Support vector machine intuitive illustration

In practical applications, there may not be a linearly separable boundary, which can be mathematically transformed from a low-dimensional space to a high-dimensional space, and then it becomes a linearly separable boundary. This process takes a technique called kernel, and the functions used in the conversion process are called kernel functions.

2.3 Upper and lower relationship learning

(1) Extract concepts between categories - subconcept relations

Encyclopedia has a relatively complete classification system, from which you can directly obtain an original classification system; Encyclopedia's classification system is usually edited or modified by the classification administrator or high-level, so their reliability is relatively high.

(2) Extracting concepts between categories and articles - subconcepts and concepts - instance relationships

Each article generally has more than one category tag, these tags can be seen as a hypernym of the current article, and then we from the Wikipedia category system to extract the relationship between the upper and lower. For articles with multiple category labels, we need to filter these tags and

choose the most reliable ones. The basis of the heuristic rule is: if the same article has more than one category label inside the upper and lower bits, those who are not The lowest level of the upper and lower relationships will be filtered.

2.4 Property Learning

The information module obtained in the encyclopedia contains a large amount of information of "attribute-value". Although the information module in the encyclopedia is defined based on the concept, it is not directly displayed in the concept-related encyclopedia, but in these concepts. In the entity. Therefore, to determine the concept of the property, you need to bottom-up, first determine the entity attributes it contains, and then make a convention to get the concept of the appropriate attributes.

Extracting attributes from the page corresponding to the entity is very simple, and it only needs to parse the format-specific write adapter of the information module in the page to complete. All entity-level attributes and values do not need to be manually checked. This is mainly because the number of entities is too large, and even if there are a small number of entity attributes that are unreasonable, the scope of the impact is only these attributes themselves, and will not affect other entities or attributes.

2.5 Conflict Resolution in Learning Process

The construction of knowledge graphs will inevitably lead to some conflicts, mainly due to the differences in construction methods and conflicts caused by different data sources.
To solve these problems, this paper mainly adopts two methods to solve the conflicts in the learning process:

(1) Conflict resolution method for scoring data sources

This method mainly uses the statistical method and the artificial method to score the reliability of the structured and semi-structured data for constructing the knowledge graph and selects the data source with higher score as the final selected value.

(2) Artificial conflict resolution

Based on the artificial conflict resolution method, that is, when the automatic conflict resolution method fails to complete the conflict resolution, manual intervention is required, or in the case that certain reliability requirements for the knowledge graph are high, the manual solution to the automatic solution is required. The results of the final determination and modification.

2.6 Knowledge update

Logically, the update of the knowledge base includes the updating of the concept layer and the updating of the data layer. The conceptual layer update refers to new concepts acquired after adding data, and new concepts need to be automatically added to the conceptual layer of the knowledge base. The update of the data layer is mainly the addition or updating of entities, relationships, and attribute values. The updating of the data layer needs to take into account various factors such as the reliability of the data source, the consistency of the data (whether there is a contradiction or redundancy, etc.) and other factors. The current popular method is to select reliable data sources, such as web sites, etc., and select the entity and attribute relationships that occur at high frequencies in each data source to add knowledge graphs. The update of knowledge can also use the crowdsourcing model (such as Freebase), while for the conceptual layer update, it requires a manual review by a professional team.

There are two ways to update the content of the knowledge graph: data-driven comprehensive update and incremental update. The so-called full update refers to the construction of a knowledge graph from scratch based on all the updated data. This method is relatively simple, but it consumes a lot of resources and requires a lot of human resources to maintain the system; incremental update is based on the current new data, and new knowledge is added to the existing knowledge graph. This method consumes a small amount of resources, but it still requires a lot of manual intervention (defined rules, etc.), so it is very difficult to implement.

3. Knowledge graph Based Information Retrieval Model

3.1 Based on the knowledge graph Sorting results

Based on the knowledge graph, this paper presents a simple sorting model of search results, mainly to illustrate the enhancement effect of knowledge graphs on information retrieval.

$$\begin{gathered}\text{Hit set S}, \boldsymbol{S_i}\text{Property set F} \\ f_1(\mathrm{p}) = \sum_{i \in F} \lambda_i \blacksquare F_i \\ f_2(\mathrm{p}) = \sum_{i,j \in S}^{i \neq j} R(i,j) \\ \mathrm{R(i,j)} = \begin{cases} 1 & \exists r \quad i \to r \to j \\ 0 \end{cases} \\ \mathrm{f(p)} = \alpha f_1(\mathrm{p}) + \beta f_2(\mathrm{p})\end{gathered} \tag{5}$$

Formula (5) is the main steps for sorting the search results through the knowledge graph:

(1) Calculate the matching value of each attribute value and retrieval value of each entity

(2) Multiply the attribute match value in the calculation entity by the sum of the weights for that attribute.

(3) Calculate the relationship between the hit entity and other hit entities and:

a) If there is a relationship between two entities, their value is marked as 1;

b) Otherwise, it is marked as 0;

(4) The two parts of the calculated value are respectively multiplied by the weights, and the sum is obtained.

As we can be seen from the above calculation steps, the score calculation of the hit document is mainly divided into two parts, the matching value of the first part of the document and the search keyword, and the matching degree of the second part of the hit document relationship; the two parts have the score Different weights, the weights can take empirical values or cross-validation methods.

3.2 Knowledge graph Based Information Recommendation

Relevant data recommendation based on retrieval hit content is a main content of the search engine. The recommendation algorithm in the traditional search engine is mainly based on content recommendation, user-based recommendation and collaborative filtering recommendation method. The above methods have their advantages and disadvantages. This paper proposes a method of recommendation based on the knowledge graph. The main idea is to calculate the distance between the hit entity and other entities through the knowledge graph and recommend the entities closer to the entity.

4. Conclusion

This paper takes the knowledge graph-based information retrieval as the research background, and mainly introduces the technology of constructing the knowledge graph based on encyclopedia, and the application of the model of information retrieval based on the knowledge graph.

Knowledge graph can not only provide data support for information retrieval, but also represent the semantic relationship between entity knowledge in the form of graphs. The application of information retrieval based on knowledge graphs enables search engines to better understand the needs of users and provide users with a better understanding. Provide more intelligent, accurate, and efficient answers; at the same time, it can also feed back the relationship of knowledge in the map, and provide inference and related recommendation through the knowledge graph.

Acknowledgements

This work is supported by “the National Natural Science Foundation of China (Grant: 61762076)”

References

[1] Hu Fangkui. Research on construction methods of Chinese knowledge graph based on multiple data sources[D]. *East China University of Science and Technology,* 2015.

[2] SHAO Ling. Research and Application of Search Engine Technology Based on Knowledge Atlas[D]. *University of Electronic Science and Technology,* 2016.

[3] Zhao Rongying, Chen Rui. Research on Knowledge graphping of Cross-Language Information Retrieval[J]. *Information Theory and Practice,* 2011, 34(10):96-100.

[4] Zhang Zhaofeng, Zhang Junsheng, Yao Changqing. An Automatic Construction Method of Technical Power Map Based on Knowledge graph[J]. *Information Theory and Practice,* 2018, 41(03): 149-155.

[5] Liu Yu, Li Yang, Duan Hong, Liu Yao, Qin Zhiguang. A Survey of Knowledge graphping Construction Techniques[J]. *Journal of Computer Research and Development,* 2016, 53(03): 582-600.

[6] Hu Dehua, Wang Rui. Knowledge graphping of Information Retrieval Research[J]. *Library Magazine,* 2015, 34(01): 20-28.

[7] Ge Bin, Tan Zhen, Zhang Hao, Xiao Weidong. Construction of Military Knowledge Atlas[J]. *Journal of Command and Control,* 2016, 2(04): 302-308.

[8] Guo Xiyue. Extracting Entity Relationships for Open Field Texts [D]. *Central China Normal University,* 2016.

[9] Yan Fan,Chengyu Wang,Guomin Zhou,Xiaofeng He. DKG Builder: An Architecture for Building a Domain Knowledge Graph from Scratch[M]. *Springer International Publishing:*2017-06-15.

[10] Katrine Juel Vang. Ethics of Google's Knowledge Graph: some considerations[J]. *Journal of Information, Communication and Ethics in Society,*2013,11(4).

[11] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. *Semantic Web,*2016,8(3).