Keyword extraction and ranking based on crawler and natural language processing

Enbo Zhang[1], Changmao Li[2], Li Liu(✉)
{ 248290248@qq.com[1], 2538001101@qq.com[2], link_liuli@hotmail.com (✉) }

Department of Information Science and Engineering Dalian Polytechnic University Dalian, P. R. China

Abstract. This paper adopts crawler, Hidden Markov Model, Viterbi algorithm to make a segmentation of text data on Internet, and adopt TF-IDF algorithm to extract and sort the keywords. Secondly, an experiment was carried out to extract and sort keywords from analyzing online recruitment text data. Through the experience the authors come to the conclusions: The method described in this paper can analyze the keywords of the online text and apply to various situations.

Keywords: crawler, Hidden Markov Model, Viterbi algorithm, natural language processing

1 Introduction

With the rapid development of the Internet, due to the rapid growth of the amount of data on the Internet and the reduction of the difficulty of data acquisition, people are often confused and induced by the text information on the Internet in work, study and life, and it is difficult to quickly and efficiently obtain the information they really need. In addition, due to the huge amount of data, only a small part of the data can be obtained and read by human beings, which results in a very one-sided and subjective view of things. If can analyze large amounts of data, to extract the keywords and returned to the user,this problem will be solved. Therefore, it is important to extract and sort keywords, In browsing the recruitment information. For example, when browsing the recruitment information, extracting keywords can make people quickly understand the job requirements, so as to make overall judgment.

Data acquisition is the first step,if a large amount of data is analyzed. Crawler is a common technology to get generous data. For example, there are studies[1] based on crawlers to obtain the behavior data of online forums. There are also researches[2] on price index calculation based on crawler. Processing language text data belongs to the field of natural language processing.

EAI GreeNets 2021, June 06-07, Dalian, People's Republic of China
Copyright © 2021 EAI
DOI 10.4108/eai.6-6-2021.2307707

Table 1. Comparison between crawler and traditional data collection methods.

Comparision target	Crawler	Traditional method
Acquisition efficiency	High	Low
Energy expended	Low	High
Amount of data obtained	High	Low

For example, some studies[3] use natural language processing to identify enterprise names, and some studies[4] identify professional vocabulary.

Therefore, based on crawler and natural language processing, this paper attempts to use HMM, Viterbi algorithm and TF-IDF algorithm to extract keywords from text and sort them according to weight. By extracting keywords and weights, people can not only quickly understand the main content of the text, but also understand the problem from a macro perspective.

2 Crawler

2.1 Definition of crawlers

Crawler is a program or script that automatically grabs the World Wide Web information according to certain rules. The crawler used in this paper is focused crawler, that is, according to a certain web page analysis algorithm to filter links irrelevant to the topic, keep useful links and put them into the URL queue waiting to be crawled until a certain condition of the system is reached[5].

2.2 Advantages of crawlers

With the explosive growth of data, it is more and more difficult to obtain a large number of online target information quickly and accurately. In the past, if people want to obtain data, the usual way is to use human, artificial to search all kinds of websites on the Internet, looking for the information they want. This method usually takes a lot of time and energy. As people step into the era of big data, the amount of data people need is more and more huge, which highlights the advantages of crawler.This method usually cost a lot of time and energy. As people step into the era of big data, the amount of data people need is more and more huge, which highlights the advantages of crawler.

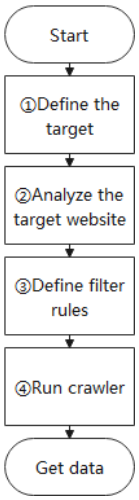

Fig. 1. Flow chart of a Crawler program

Compared with the traditional method, the main advantages of crawler are shown in Table 1, high efficiency, low energy consumption, and large amount of data acquisition. In terms of data acquisition efficiency, crawlers can obtain a large amount of information in a limited time in seconds (the specific efficiency depends on the type of target data, network speed and response speed of the target website). However, traditional methods need to search, judge and screen manually, which takes a long time and has low acquisition efficiency. In terms of energy consumption, after finishing the crawler program, let the computer run the program without supervision. The traditional method is to use manpower all the time. In terms of the amount of information, crawler can obtain a huge amount of data, but traditional methods are difficult to achieve this.

Crawler can obtain huge amount of data in a short time, which is efficient and reliable.

2.3 Crawler programming steps

Definition of regular expression: regular expression is a logic formula for string operation, that is, some specific characters defined in advance and the combination of these specific characters are used to form a "regular string", which is used to express a filtering logic for string[6].

Definition of XPath expression: XPath uses path expressions to select nodes or node sets in XML documents. Nodes are selected by following paths or steps. Reference[7] uses XPath expression to extract agent information of web pages.

As shown in Figure 1, there are four steps in using crawlers to collect data:

① Describe or define the crawling target: it is to determine the URL of the website to be crawled.

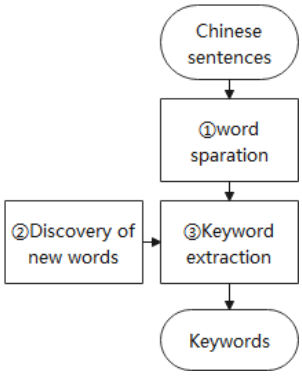

Fig. 2. Flow chart of a keyword extraction

② Write out the filter rules: generally, the filter rules are described by programming Regular Expressions or XPath Expressions.

③ Run: run the crawler program, get the data on the target website and save it according to the written filtering rules.

3 natural language processing

3.1 Definition of natural language processing

Natural language processing is an important direction in the field of computer science and artificial intelligence. It studies various theories and methods that can realize effective communication between human and computer with natural language. Basic natural language processing technologies include stop word removal, word segmentation, root extraction and part of speech tagging[8].

3.2 Keyword extraction algorithm flow

The field of natural language processing involved in this paper is syntactic semantic analysis. The main algorithm is shown in Figure 2. Firstly, the Chinese text crawled down by the crawler is segmented; secondly, new words are found to prevent word segmentation errors when new words do not exist in the dictionary appear in the text; finally, TF-IDF algorithm is used to calculate the keywords.

3.3 Process and principle of word segmentation

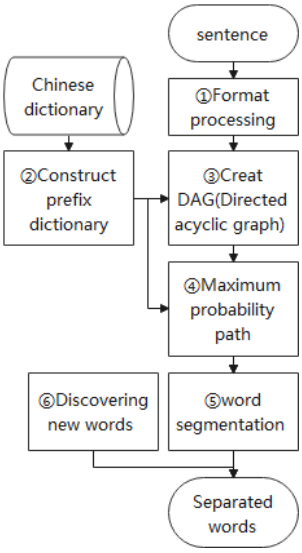

Fig. 3. Flow chart of a word segmentation

（1）Word segmentation process：

As shown in Figure 3, firstly, format the Chinese text, that is, use non Chinese characters to segment the Chinese text; secondly, construct a prefix dictionary based on the Chinese dictionary to prepare for the subsequent construction of directed acyclic graph; thirdly, construct a directed acyclic graph of sentences according to the prefix dictionary; fourthly, find the maximum probability path according to the probability in the prefix dictionary; fifthly, find the maximum probability path according to the maximum probability Probability path is used to segment sentences;

finally, new words are found, and the words that do not appear in the dictionary are recorded; the new words and the divided words are regarded as the final segmentation results, and the segmentation process ends. The Chinese dictionary and non Chinese character set files are obtained from GitHub.

（2）Principle of word segmentation

①Using the non Chinese character set obtained from GitHub, the Chinese text to be segmented can be segmented by Regular Expression or string operation.

②The prefix dictionary is constructed based on Chinese dictionary, and a trie tree is constructed to store the prefix dictionary. A prefix dictionary is a dictionary where words with the same prefix are stored together. For example: The prefixes of "Gong Ye Da Xue" are "Gong", "Gong Ye" and "Gong Ye Da". Prefix dictionaries store words with the same prefix together. The construction of prefix dictionary is to facilitate the subsequent construction of directed acyclic graph.

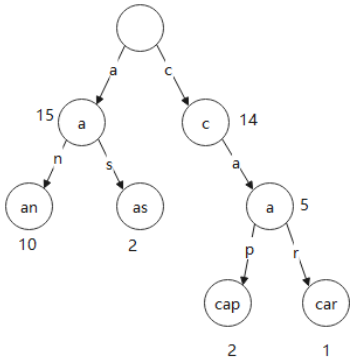

Fig. 4. Trie tree

Trie tree, also known as prefix tree[9], is an ordered tree, which stores words with the same prefix together and has the advantage of fast search speed. All descendants of a node have the same prefix, that is, the string corresponding to the node, while the root node corresponds to an empty string. The above feature is just used to store the prefix dictionary. As shown in Figure 4, key is labeled in the node and value is labeled outside the node. Each complete word corresponds to a specific integer. In this paper, the value of the node is the word frequency of the word. An empty string is saved in
the root node. The left child node is "a", and the saved value is 15, which means that the word frequency of "a" is 15 times; "the left child node of" a "is" an ", and the value is 10, which means that the word frequency of" an "is 10 times;" the right child node of "a" is "as", and the value is 2, which means that the word frequency of "as" is 2; and "an" and "as" have the same prefix "a" .

③According to the prefix dictionary, the sentence is divided into words, and the directed acyclic graph is constructed. In this paper, we use the method of calculating the optimal path through directed acyclic graph, as described in reference[10], which is one of the common applications of DAG.

④Calculate the maximum probability path: in the directed acyclic graph, each vertex is weighted, and the corresponding weight is the word frequency of the word.

Let the *route* = (w_1,w_2,w_3,...,w_n) we want to demand make $\sum weight(w_i)$ the largest

As shown in formula (1), the weight $\{R_{i\to j}\}$ of any path from W_i to W_j is equal to the sum of the weight of the path from wi to W_j.

$$\{R_{i\to j}\} = \{R_I + weight(j)\} \tag{1}$$

As shown in formula (2), the weight $\{R_{i\to k}\}$ of any path from W_i to W_k is equal to the sum of the weight of the path from W_i to W_k.

$$\{R_{i\to k}\} = \{R_I + weight(k)\} \tag{2}$$

Therefore, for nodes W_j and W_k with common precursor node W_i, it is necessary to repeatedly calculate the probability of the path to W_i, which is a repetitive subproblem.

The optimal path R_{max} of the whole graph and a terminal node W_x, for its possible existence of multiple precursors W_i, W_j, $W_{k....}$ Let the maximum probability paths to W_i, W_j and W_k be R_{maxi}, R_{maxj} and R_{maxk} respectively, then:

$$R_{max} = \max\left(R_{maxi}, R_{maxJ,}\ R_{maxK} \ldots\right) + weight\ \ (W_x) \tag{3}$$

Therefore, the problem can be transformed into solving R_{maxi}, R_{maxj} and R_{maxk}... The optimal solution in the substructure is a part of the global optimal solution, forming the optimal substructure problem.

It satisfies the repeated subproblem and the optimal substructure problem, so it turns into a dynamic programming problem. Similar to the method described in reference[11], the optimal path (maximum probability path) in the directed acyclic graph is calculated by dynamic programming. Finally, Chinese text segmentation is realized according to the maximum probability path.

⑤Find new words based on HMM.

HMM: Hidden Markov model, is a statistical model based on Markov hypothesis[12]. Hidden Markov model (HMM) can be described by five elements, including two state sets and three probability matrices[13]:

（1）Implicit state S

（2）Observable state O

（3）Initial state probability matrix π

（4）Implicit state transition probability matrix A

（5）Observation state transition probability matrix B

As described in reference[14], the process of Chinese word segmentation can be transformed into marking the position of each word in a string of Chinese characters sequence, which is brought in as a parameter of the model. In this paper, the position B (begin), M (middle), E (end) and S (single) of the word in the word are regarded as the hidden state, and the word is the observed state. and the remaining parameters in the dictionary file are used, which is a standard decoding problem. According to the probability, Viterbi algorithm is used to solve the maximum possible hidden state. We can find new words. In this paper, the trained HMM is used.

In fact, the essence of Viterbi algorithm is to use dynamic programming to solve HMM prediction problem[15], that is, to use dynamic programming to find the

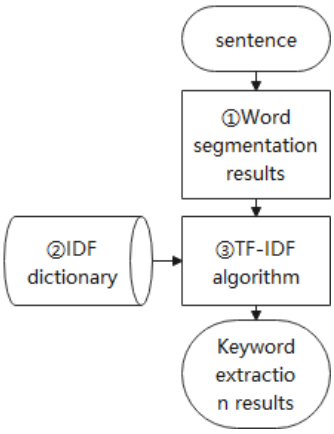

Fig. 5. Flow chart of a keyword extraction (TF-IDF part)

maximum probability path (optimal path). In this case, a path corresponds to a sequence of states.

3.4 Keyword extraction and its principle

Keywords are mainly obtained by TF-IDF algorithm. As shown in Figure 5, after the sentence is divided into words, each divided word is calculated based on the IDF dictionary by using TF-IDF algorithm, and the result is calculated to get the $TF\text{-}IDF_{ij}$ The larger the value of ij, the more important the word is relative to the text.

Definition of TF-IDF algorithm: TF-IDF (term frequency – inverse document frequency) is a common weighting technology for information retrieval and data mining. TF is term frequency and IDF is inverse document frequency. TF-IDF is a statistical method to evaluate the importance of a word to a file set or one of the files in a corpus. The importance of a word increases with the frequency of its appearance in the document, but decreases inversely with the frequency of its appearance in the corpus[16].

The principle of TF-IDF algorithm is as follows

$$\mathrm{TF} - IDF_{ij} = tf_{ij} \times idf_i \tag{4}$$

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{5}$$

As shown in formula 4, the importance of a word to a file set $TF\text{-}IDF_{ij}$ is equal to the product of word frequency *TFij* and inverse text frequency index idf_i.

As shown in formula 5, word frequency tf_{ij} is equal to the number of times that *i* word appears in *j* document nij divided by the sum of the number of times that all words appear in *j* document $\sum_k n_{kj}$. The inverse text frequency index *idf* in this paper is directly read into the dictionary.

4 Keywrd sorting

4.1 All Chinese Texts

When the text is all Chinese, according to the weight of each word calculated by TF-IDF algorithm after word segmentation, all the words are arranged in descending order, and a threshold *U* is specified. The words whose weight is greater than the threshold *U* are the keywords of the text.

4.2 The main body is Chinese text mixed with English

When the text is a mixture of Chinese and English (such as the recruitment information on the recruitment website), most English words are separated by spaces except some special words (such as New York), so the difficulty of word segmentation is relatively simple. First of all, we need to screen out and save the English text separately with regular expression; second, we need to get the information like "new" on the Internet The third step is to filter the mixed English text, screen out and save the special English words separately; the fourth step is to divide the English words by spaces; finally, the selected special English word set and the separated ordinary English word set are combined to complete the English text segmentation.

After word segmentation in the English part, to filter the second time, we need to delete "an", "and" and other meaningless function words from the English word set, and finally calculate the frequency of each English word, that is, the frequency of each English word.

In the Chinese text part, the word frequency is calculated by the method described in this paper.

According to word frequency, Chinese and English words are sorted in descending order. Take the first *N* words as the keywords of the mixed language text.

5 Experiment

Taking the information text of Chinese recruitment website as an example, based on Python and MySQL, this paper uses the method described in this paper to analyze the keywords of "Python development engineer" on a certain website.

5.1 Crawler section

Fig. 6. The composition rule of URL

First, determine the URL of the website to be crawled and its naming rules. The data to be crawled in this paper is the detailed recruitment data of "Python development engineer" in Dalian area on a recruitment website. After manual login, the naming rules of URL are analyzed, and the specific format of URL naming is shown in Figure 6, which is the combination of "main URL", "region number", "position number" and "number of pages". Through manual search and comparison, you can get the region number of Dalian and the position number of Python development engineer, so you can make your own URL to crawl.

The second step is to manually log on the target web page, watch the front-end code of the web page, and formulate the crawling strategy according to the target data. For example, observe the format of data in the front-end code, storage rules, etc. For example, the recruitment data to be captured in this paper is stored in the P tag under div with multiple category attributes of position number.

The third step is to start to write the crawler program and write the filtering rules. In order to improve the efficiency of filtering, this experiment uses Regular Expression and XPath Expression according to the situation. In this step, we use XPath Expression and Regular Expression to filter out the text data, and then use regular expression to clean the filtered text data again. This paper mainly crawls the text data for each recruitment data.

Finally, we start to run the crawler program and save the text data in the database to prepare for the natural language processing steps.

5.2 Natural language processing

First of all, the crawler program crawled down the text format processing, to prepare for word segmentation. In this paper, the target site in Dalian area accurate search "Python development engineer" recruitment data, a total of 28, this article put 28 data into a text, start word segmentation.

The recruitment data is a mixture of Chinese and English text, so the flow chart is shown in Figure 7. First, program regular Expressions to separate the Chinese and English in the text. Second, segment words separately.

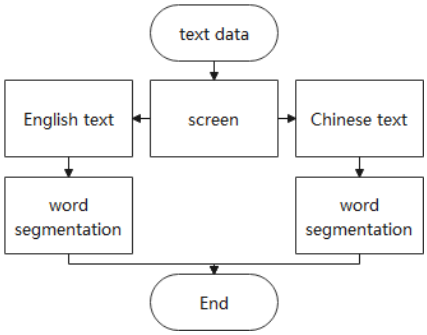

Fig. 7. Flow chart of Chinese English mixed text

Key	Value	
0:	[0,2,5]	“有”，“有经验”，“有经验者”
1:	[1,2]	“经”，“经验”
2:	[2]	“验”
3:	[3]	“者”
4:	[4,5]	“优”，“优先”
5:	[5]	“先”

Fig. 8. Key-Value diagram

In the process of English text segmentation, the special English words are separated based on the special English word dictionary, and then the space and punctuation are used as the segmentation position. After the word segmentation, it is combined with the separated Special English words to form a new array. And calculate the frequency of each English word.In the process of Chinese text segmentation, we first download the non Chinese character set from GitHub and use regular expressions to segment Chinese text.

Second, we need to build a trie tree to process the downloaded Chinese dictionary into prefix dictionary.

In the third step, according to the prefix dictionary, the result of the first step is further divided into words, and the directed acyclic graph is constructed.

When constructing a directed acyclic graph, we need to create a python list or dictionary to store the mapping relationship between vertices. For example, The sentence “有经验者优先。” is segmented according to the prefix dictionary, it can be divided into eight words: “有” “经” “验” “者” “优” “先” “有经验” “经验” “优先” As shown in Figure 8, this paper marks each word with the position in the text, the position of the first word is set to 0, and the position of the last word is set to 5, which is used as the key; the list formed by the end position of the divided words is used as the value. As shown in Figure 8, there are six words “有经验者优先”, so the key label is 0-5; Because there are three words related to “有” after word segmentation: “有”, “有经验” and “有经验者”, the key "0" Representing “有”, "the key" 2 "Representing “经” and the key " 3 "Representing “者” are written into the value of key0, and the value is [0,2,5]. The rest of the key value pairs are the same.

Construct a directed acyclic graph with the above words and words, as shown in Figure 9. The directed acyclic graph has six vertices, representing six words in the text respectively. The serial number above the vertex represents key, and the vertices are connected according to the order of appearance in the text. After word segmentation, ，“有经验”、“经验” “有经验者” and “优先” are divided together, so that “有” points to “验” and “者”, “经” points to “验” and “优” points to “先”, and the representative is divided into one word.

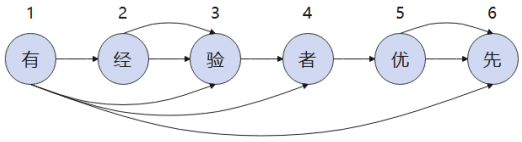

Fig. 9. DAG（Directed acyclic graph）

In the fourth step, by writing a bottom-up dynamic programming algorithm, the maximum probability path of the directed acyclic graph is calculated, and the Chinese text segmentation is realized according to the maximum probability path.

In the fifth step, new words are recognized based on HMM and Viterbi algorithm, and the results are added to the word segmentation results in the fourth step.

Finally, the TF-IDF algorithm is used to calculate the Chinese keywords, and the Chinese and English keywords and the word frequency of each word are saved in the database.

5.3 Keyword sorting and display

The Chinese and English keywords in the database are sorted in descending order according to word frequency.

In order to better show the conclusions (keywords and weight), data visualization technology is used to display the results in the word cloud and figures.

The final result is shown in Figure 10, keywords are displayed in the figure of word cloud. All the words in the figure are the key words analyzed by the system. The larger the font of the word, the greater the weight of the word. That is to say, the more important the word is to the position of "Python development engineer". For example, the biggest word is "Python" and the second is "MySQL". It is proved that the two most important skills for the position of "Python development engineer" are "Python" and "MySQL".

Font color is generated randomly, just to distinguish keywords, does not have special meaning.

Further data are shown in figures 11, 12 and 13. As shown in Figure 11, the x-axis represents all the analyzed keywords, and the y-axis represents the word frequency of the keyword, that is, the number of times the keyword is required to be mastered by the recruiter.

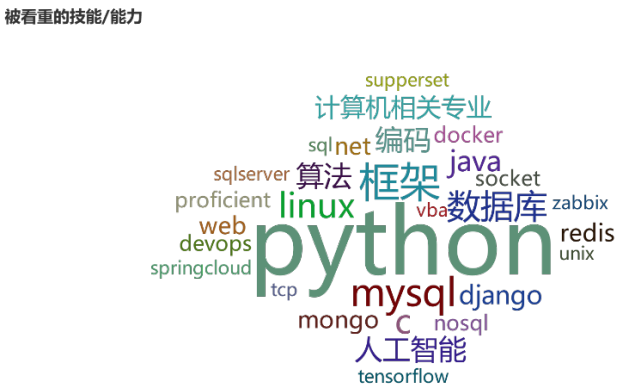

Fig. 10. The word cloud of keywords

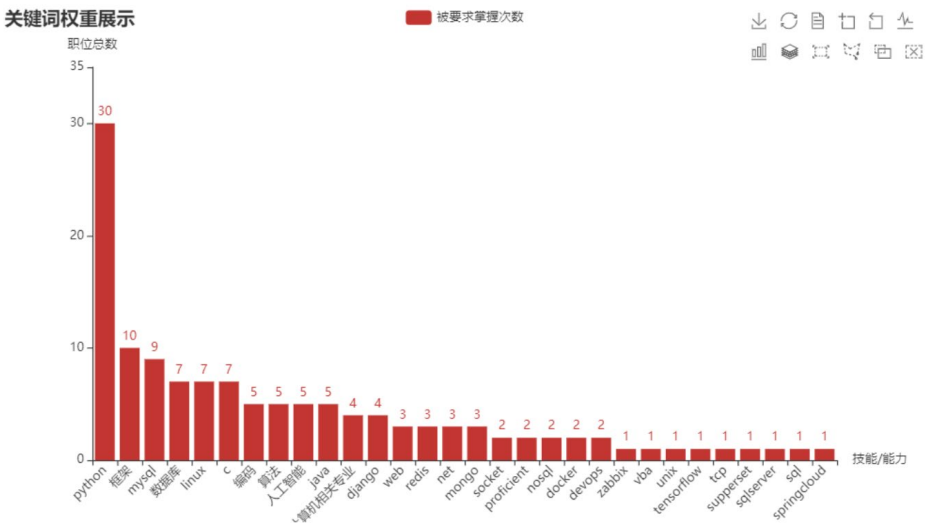

Fig. 11. Keywords overview

As shown in Figure 12, each pie chart represents a keyword, and the percentage at the center of the pie chart represents the probability of the keyword appearing in the recruitment requirements.

Figure 13 shows the degree to which keywords are required to be mastered by the recruiter. The degree is divided into four categories: "expert","Familiar", "know" and "master". As shown in Figure 13, the x-axis represents the keyword, and the y-axis represents the probability that the keyword is required by the recruiter in the four standards.

Fig. 12. Keywords pie chart

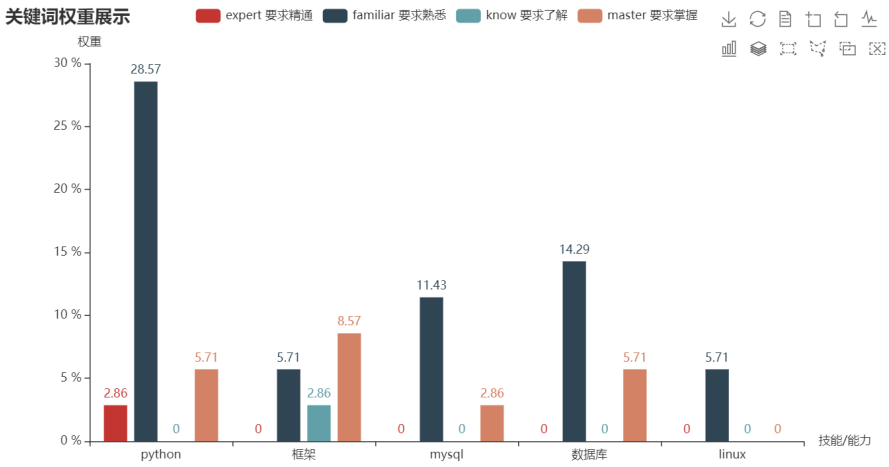

Fig. 13. Figure showing keywords in different degrees

6 Conclusion

In the era of big data, with the explosive growth of data volume and the reduction of data acquisition difficulty, people are often induced and confused by the information on the Internet, and it is difficult to obtain the real valuable information. If we use the keyword capture and sorting method described in this paper, we can produce different effects according to the size of the data and the actual application scenarios. When the amount of data is small, the key words can help users understand the central idea of multiple texts, identify the main content of the text, and improve the efficiency of learning and work; when the amount of data is large, taking the experiment in this paper as an example, we can analyze the key words through recruitment data, and summarize the requirements of most companies for a certain position or major General requirements and specific skills expectations can make users more aware of the current enterprise recruitment trends, industry hot spots and other information, so that users will not be limited to limited recruitment information, but based on the huge amount of data, have a macro understanding of the specific skills and job requirements of recruitment, so that users can understand what skills are more valued by recruiters, so as to help them Users make choices on the way to study and work. And by modifying the crawler program and analyzing the recruitment data of different regions, we can get the industry characteristics of a region, which is more region specific.

The method described in this paper is not only applicable to the above examples, but also applicable to other situations, and can be widely used in a variety of situations.

Acknowledgment

The authors acknowledge financia support from College Student Innovation and Entrepreneurship Project.

References

[1] Zheng JJ, CHENG Y, MA G, et al.: Research on evolution of public environmental protection behavior based on data crawler and dynamic social network. System Engineering-Theory ＆Practice. Vol. 2, pp. 219-229 (2020)

[2] LIU YB, ZHAO ZD, LIU HB.: A model of compiling price index based on the “Web Scraping” technology. Statistical Research. Vol. 31, pp. 74-80 (2014)

[3] HUANG GX, ZHU SX, WANG XH, et al.: Natural language processing and machine learning-based suspected soil. Chinese Journal of Environmental Engineering. Vol.14, pp. 3234-3242 (2020)

[4] ZHU TT, DU YF, LI RF, et al.: An unsupervised approach to recognizing new words in power domain. Electric Power Engineering Technology. Vol. 39, pp. 159-165 (2020)

[5] ZHOU LZ, LIN L.: Survey on the research of focused craw ling technique. ComputerApplications. Vol. 25, pp. 1965-1969 (2005)

[6] JeffreyE FF.: Introduction to regular expressions. Vol. 1, pp. 4-5. Mastering Regular Expressions, CN(2009)

[7] Antonov E, Lopatina E, Ionkina K, et al.: Agent data merging. Procedia Computer Science. Vol. 169, pp. 473-478 (2020)

[8] WANG CH, ZHANG M, MA SP.: A survey of natural language processing in information retrieval. JOURNAL OF CHINESE INFORMATION PROCESSING. Vol. 21, pp. 35-45 (2007)

[9] ZHANG M , CHEN W , WANG Q.: Data structure and algorithm. China Machine Press, CN(2010)

[10] Borenstein D.: A directed acyclic graph representation of routing manufacturing flexibility. European Journal of Operational Research. pp. 78-93 (2000)

[11] LV YJ, ZHAO TJ, YANG MY, et al.: Leveled unknown Chinese words Resolution by dynamic programming. JOURNAL OF CHINESE INFORMATION PROCESSING. Vol. 15 (2001)

[12] HU CJ, HAN ZQ.: Application study of Hidden Markov model based part-of-speech tagging. Computer Engineering and Applications. Vol. 6, pp. 62-64 (2002)

[13] Rose R C.: A hidden Markov model based keyword recognition system. Proc of Icassp Albuquerque Nm Usa. Vol. 1, pp. 129-132 (1990)

[14] WANG WF, XU HJ, YANG WZ, WU XL.: Review of Chinese word segmentation algorithms. Group Technology ＆ Production Modernization. Vol. 35, pp.1-8 (2018)

[15] CHEN K T.: Veterbi decoding method for convolutionally encoded signal. (2008)

[16] SHI CY, XU CJ, YANG XJ.: Study of TFIDF algorithm. Journal of ComputerApplications. Vol. 29, pp. 167-180 (2009)