Entity extraction based on the combination of information entropy and TF-IDF

Hankiz Yilahun and Askar Hamdulla*

Xinjiang Key Laboratory of Multilingual Information Technology,
School of Information Science and Engineering,
Xinjiang University,
Urumqi, 830046, China
Email: hansumuruh@xju.edu.cn
Email: Askar_H@21cn.com
*Corresponding author

Abstract: Traditional knowledge graph entity extraction methods require expert knowledge and a large number of artificial features. Furthermore, deficiencies exist in the accuracy and efficiency of keyword extraction based on methods such as TF-IDF. Thus, this study proposes a Chinese entity extraction method based on the combination of information entropy and TF-IDF. First, the text is preprocessed, which involves operations such as sentence segmentation, word segmentation, removal of stop words, and POS tagging, to detect keywords based on POS. Secondly, the word frequency is analysed to determine feature word weight, and the TF-IDF algorithm is used to compare the importance of keywords. Finally, information entropy is used to improve the TF-IDF algorithm to provide entity knowledge for the construction of the knowledge graph. The entity extraction method and optimisation scheme proposed in this study can help users extract domain entities and provide better entity resources for the construction of knowledge graphs.

Keywords: entity extraction; improved TF-IDF; information entropy; knowledge graph.

Reference to this paper should be made as follows: Yilahun, H. and Hamdulla, A. (2023) ‘Entity extraction based on the combination of information entropy and TF-IDF’, *Int. J. Reasoning-based Intelligent Systems*, Vol. 15, No. 1, pp.71–78.

Biographical notes: Hankiz Yilahun received her BS in 2002, MS in 2009, and PhD in 2020 all in Computer Science and Technology from the Xinjiang University of China, Beijing University of Technology of China and Xinjiang University of China, respectively. Currently, she is an Associate Professor working as a teacher at the School of Information Science and Engineering, Xinjiang University, China. Her research interests include knowledge graph and its applications.

Askar Hamdulla received his BE in 1996, ME in 1999 and PhD in 2003 all in Information Science and Engineering from the University of Electronic Science and Technology of China. In 2010, he was a Visiting Scholar at the Center for Signal and Image Processing, Georgia Institute of Technology, GA, USA. Currently, he is a Professor in the School of Software Engineering, Xinjiang University. He has published more than 140 technical papers on speech synthesis, natural language processing and image processing. He is a senior member of CCF and an affiliate member of IEEE.

1 Introduction

In the era of big data, tremendous amounts of information and data have drastically changed human civilisation. The rapid growth in the number of documents generated everyday means that a large amount of knowledge is proposed, improved, and used. For readers, especially newcomers to a given field, excavating suitable knowledge entities from massive documents is time-consuming and labour-consuming, negatively impacting research efficiency. The broad availability of information provides more opportunities for people, but a new challenge has risen as well; that is, how to extract and use knowledge from numerous information resources, especially how to conduct knowledge extraction and text mining (TM) from massive documents in special domains (Moqurrab et al., 2021; Hu et al., 2021). Entity extraction is a basic task in the field of natural language processing (NLP) and is the most important part of the knowledge graph construction. It realises the extraction of entity information from text data. High-precision entity extraction virtually guarantees the broad applicability of the constructed knowledge graph.

Initially, a rule-based method was used to extract entities. Domain experts and linguists manually formulated effective rules to identify entities by matching text and rules. For example, in the Chinese system, the words that follow a person’s name might be ‘do’, or ‘teacher’, and the end of the unit organisation name might be ‘company’, or ‘school’… Using these rules, the person names, place names, organisation names and other entities can be extracted.

At present, entity extraction can achieve significant results only in a limited text field and a limited number of entity types. In most research articles, only simple entity extraction can be performed, and the entity extraction algorithm can be improved. In addition, compared with other knowledge extraction, there are too few corpora that can be used for entity extraction, and there is no large open corpus for researchers to use, which limits the development of entity extraction. Accordingly, researchers continue to optimise the entity extraction method. In 2015, Yang proposed an improved TF-IDF algorithm to extract keywords in his dissertation. The basic content includes part-of-speech (POS) tagging, segmentation, and entity extraction. Researchers do not need to manually construct features, or construct features according to different tasks, which is very helpful for researchers who are new to the field of NLP. In 2019, Zhai et al. proposed a neural network structure entity extraction method based on the BiLSTM-CRF model, which converts the words in the text into word vectors, then used BiLSTM to process the vectors to obtain sentence features, and used CRF to mark and extract entity information. This method has broader applicability. In 2020, Gorla and others used the features related to the dictionary of geographical names that is automatically generated by the Wikipedia page to improve Telugu's named entity recognition performance, and built an entity extraction model with context, vocabulary, and corpus functions. The results that they obtained were better than support vector machine (SVM) and the conditional random field model.

In the same year, Li and He (2020) seek to overcome the limitations of traditional word segmentation algorithms and traditional keyword extraction algorithms for customer-centricity, analysis of customer reviews, and extraction of important customer needs in modern marketing activities, proposed fusion of information entropy and an algorithm for extracting multiple right TF-IDF keywords. The algorithm first uses a word segmentation algorithm that combines mutual information and left and right entropy to segment titles and user comments to generate new words. Then, it uses the TF-IDF algorithm to extract comment keywords and title keywords, and adds different feature weights according to the location factor, POS factor, and word length factor of the keywords so as not to ignore the different importance of titles and comments, while improving accuracy. Cosine similarity is then used to compare the similarity of the two keywords to determine the quality of the review.

This article uses the method of obtaining entities from unstructured data, and performs preprocessing operations such as sentence segmentation, word segmentation, stop words removal, and part-of-speech tagging. Finally, TF-IDF is combined with information entropy to improve the accuracy of the entity and provide better entity information for the construction of the knowledge graph.

2 Related algorithms

2.1 TF-IDF algorithm

The main concept behind term frequency (TF) and inverse document frequency (IDF) is that if a word appears many times in the text of the study, and only appears a few times in the entire corpus, the word is considered very important, and may become a keyword.

The TF-IDF calculation formula is

$$TF\text{-}IDF = TF \times IDF, \tag{1}$$

where *TF* is defined as

$$TF = \frac{n}{N}, \tag{2}$$

where *n* represents the number of times a word appears in the text, and *N* represents the total number of words that appear in the text. *TF* indicates the frequency of a word in the article. If a word in an article appears multiple times, the word may be a more important word. Stop words are not included here.

IDF is defined as

$$IDF = \log\left(\frac{M+1}{M(x)+1}\right), \tag{3}$$

where *M* is the total amount of text in the corpus, and *M*(*x*) is the amount of text containing the vocabulary. IDF is a measure of the 'weight' of a word. If a word has a low frequency in multiple documents, it means that it is a relatively rare word, but it appears often in a specific article. The greater the IDF value of this word, the greater its 'weight' in the article. Therefore, when a word is more common, its IDF is lower.

After calculating the values of TF and IDF, the two values can be multiplied to obtain the TF-IDF value. The higher the TF-IDF value of the word, the greater the importance in the article, and the more likely it is the keyword of the article.

2.2 Information entropy

Information entropy was proposed by Shannon to describe the uncertainty of the symbols emitted by the source. It can be measured by the probability of the occurrence of the symbols. The information entropy model is a commonly used classification method (Ding et al., 2014; Wu et al., 2019; Wang, 2005). It was first introduced by English entity Borthwick. Its characteristic is that it meets the entropy maximisation criterion under limited conditions. It is used for various types of feature training. It consists of a set of corresponding weights, which through linear combination, are integrated into a unified model (Qu and Shen, 2008).

If the source symbol has n values (x_1, x_2, x_3, …), their corresponding probability is $p(x_1)$, $p(x_2)$, $p(x_3)$, …, $p(x_n)$, and the appearance of each symbol is independent of each other. The information entropy *H*(*X*) calculation formula is then defined as (logarithm base 2)

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i). \tag{4}$$

This paper also calculates the amount of change in information entropy brought about by words. The calculation formula is

$$I(C; x_i) = H(C) - H(C') \tag{5}$$

where *H*(*C*) is the information entropy of the entity set that does not contain the word x_i, *H*(*C*′) is the information entropy of the entity set that contains the word x_i, and *I*(*C*; x_i) indicates the amount of information entropy reduction after the keyword is added (Zhang et al., 2019; Zhou et al., 2008).

The maximum entropy model can integrate various language-related or unrelated factors, automatically assign them different weights according to the training corpus, and make extremely accurate predictions for situations that have not been seen before. The maximum entropy model constructs the model by statistically learning relevant feature information from actual data, and uses the feature information to predict the change trend of related parameters in a random process (Zhou et al., 2008). The maximum entropy method uses statistical methods to learn rules and then uses the learned rules to classify the input text. It can be seen as the unity of rules and methods.

2.3 Improved TF-IDF model based on information entropy

To overcome some of the shortcomings of the TF-IDF algorithm, this article combines it with information entropy to optimise the calculation function of TF-IDF, thereby improving the accuracy of the extracted entities. The formula is

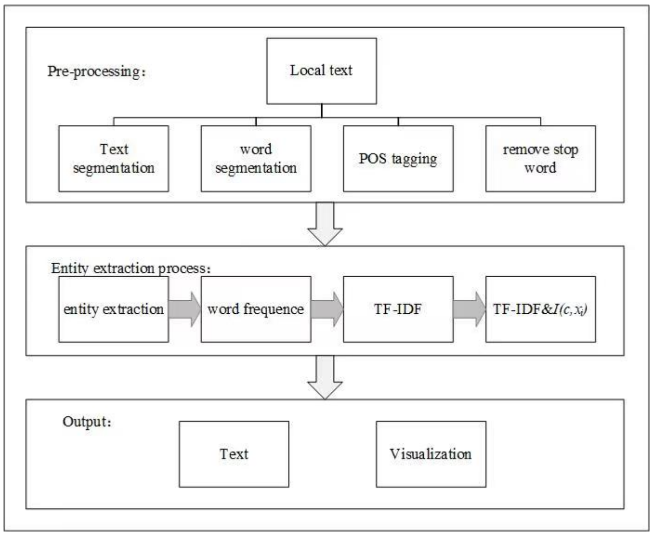
$$w_i = tf_{idf(x_i)} + I(C; x_i), \tag{6}$$

where w_i is the weight value of the word x_i, *tf_idf*(x_i) is the TF-IDF value calculated in the previous stage of the word, and *I*(*C*; x_i) is the information entropy reduction of the entity set after the word x_i appears. The more chaotic a system, the higher the entropy is. After the keyword appears, the entity set is more ordered, and the corresponding entropy value is lower. The addition of *tf_idf*(x_i) and the reduction of information entropy can increase its weight value. Conversely, for words that are definitely not keywords, the entropy value increases, and where the information entropy decrease is a negative value, *tf_idf*(x_i) is added to the information entropy decrease to reduce its weight value.

3 Realisation and analysis

In this study, entity extraction is divided into three steps: a preprocessing stage, an entity extraction and result processing stage, and an output stage. Figure 1 is a technical line of entity extraction.

Figure 1 Flowchart of entity extraction



3.1 Text preprocessing

3.1.1 Text segmentation, word segmentation, and POS tagging

Before processing the text formally, it is necessary to segment it to prepare for the subsequent data cleaning. The principle of the segmentation is to find the sentence ending characters such as '.', '?', '!', ';'. At the same time, we must consider the compound symbols contained in the character dialog, such as periods (.) and single quotation marks ('), so as to block.

In this study, we use regular expressions and *Jieba* libraries to perform string operations, to retrieve text that meets a certain rule, and to implement sentence clauses. Table 1 lists the processing modes and results of the *Jieba* tool for processing example sentences.

1 The precise mode separates the sentences most accurately, and the result can be connected into sentences; the text is accurately separated, and there are no redundant words.

2 The full mode scans out all possible words in the text. If there is redundancy, all possible word results in the sentence structure are displayed, but the ambiguity cannot be resolved.

3 The search engine mode continues to segment long words in the precise mode to improve the recall rate and is suitable for search engines.

Table 1 Processing modes of Jieba

Word segmentation mode	*Results*
Precise mode	京北/输油管/油输/北京
Full mode	京/北/输油/输油管/油管/油/输/北京
Search engine mode	京北/输油/油管/输油管/油输/北京
POS tagging mode	京北_ns 输油管_n 油输_n 北京_ns

In addition, users can import a custom domain dictionary, which stipulates that certain domain words need to be disassembled, and certain words need to be combined and displayed, so as to achieve a better word segmentation effect.

For example, in Table 2, ‘information science and engineering’ represents an organisation name and cannot be separated, and ‘student union’ in the original sentence represents a noun plus a verb, not an organisation name. It needs to be broken down into two words: ‘student’ and ‘meeting’.

Table 2 Examples of word segmentation directly and after adding to the dictionary (see online version for colours)

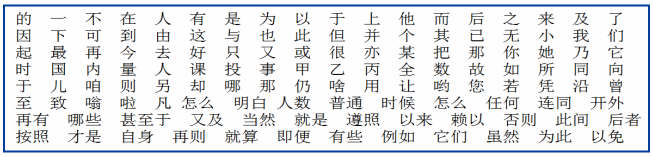
Original sentence	新疆大学信息科学与工程学院的学生会主动学习
Direct word segmentation	新疆大学 信息 科学 与 工程 学院 的 学生会 主动 学习
After adding to the dictionary	新疆大学 信息科学与工程学院 的 学生 会 主动 学习

Note: ‘Information science and engineering’ is the part marked in red. ‘Student’ and ‘meeting’ is the part marked in blue.

3.1.2 *Remove stop words*

In this study, we chose the *Jieba* word segmentation tool to perform word segmentation, and it turns out that some words are correctly identified, but there are also some words not related to the domain entity information, such as ‘的’, ‘上’, ‘and’, ‘时’, ‘所’ ‘wait’. Therefore, these stop words need to be removed to make the remaining text have as much domain-meaning as possible.

Figure 2 Stop words document (see online version for colours)



的 一 不 在 人 有 是 为 以 于 上 他 而 后 之 来 及 了
因 下 可 到 由 这 与 也 此 但 并 个 其 已 无 小 我 们
起 最 再 今 去 好 只 又 或 很 亦 某 把 那 你 她 乃 它
时 国 内 量 人 课 投 事 甲 乙 丙 全 数 故 如 所 同 向
于 儿 咱 则 另 却 哪 那 仍 啥 用 让 哟 您 若 凭 沿 曾
至 致 嗡 啦 凡 怎么 明白 人数 普通 时候 怎么 任何 连同 开外
再有 哪些 甚至于 又及 当然 就是 遵照 以来 赖以 否则 此间 后者
按照 才是 自身 再则 就算 即便 有些 例如 它们 虽然 为此 以免

Table 3 Remove stop words and some punctuation results

Original word segmentation result	劳动 生产力 上 的 增进 以及 运用 劳动 时 表现 的 熟练 技巧 和 判断力
After removing the stop words	劳动 生产力 增进 运用 劳动 表现 熟练 技巧 判断力

First of all, we need to have a ‘.txt’ file containing meaningless words, that is, a stop word database. Figure 2 shows the stop word document used in this experiment, which lists some meaningless words. Adding the usual punctuation marks in the stop word document can also achieve the effect of removing the punctuation marks, which greatly aids in data cleaning. As shown in Table 3, the results are compared after the stop words and some punctuations are removed. The words and commas marked in red in the first line of the original sentence are considered the stop words removed.

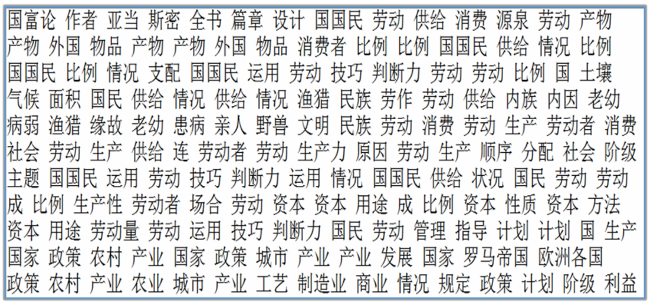
3.2 *Entity extraction and result processing*

Entity extraction focuses on extracting certain nouns in the text, including person names, place names, organisation names, and proper nouns. In the entity extraction process, regular expressions can be used to find the corresponding symbols to filter, so as to achieve entity extraction (Bai, 2008). Table 4 lists the noun types of stammering. The results of word segmentation by Jieba are shown in Figure 3.

Table 4 Types of nouns in Jieba

No.	*Symbol*	*Type*
1	n	名词 (noun)
2	nr	人名 (person name)
3	ns	地名 (place name)
4	nt	机构团体名称 (organisation name)
5	nz	其他专有名词 (other proper nouns)
6	ng	名语素 (nominal morpheme)
7	vn	动名词 (gerund)

Figure 3 Partial results of entity extraction (see online version for colours)



国富论 作者 亚当 斯密 全书 篇章 设计 国国民 劳动 供给 消费 源泉 劳动 产物
产物 外国 物品 产物 产物 外国 物品 消费者 比例 比例 国国民 供给 情况 比例
国国民 比例 情况 支配 国国民 运用 劳动 技巧 判断力 劳动 劳动 比例 国 土壤
气候 面积 国民 供给 情况 供给 情况 渔猎 民族 劳作 劳动 供给 内族 内因 老幼
病弱 渔猎 缘故 老幼 患病 亲人 野兽 文明 民族 劳动 消费 劳动 生产 劳动者 消费
社会 劳动 生产 供给 连 劳动者 劳动 生产力 原因 劳动 生产 顺序 分配 社会 阶级
主题 国国民 运用 劳动 技巧 判断力 运用 情况 国国民 供给 状况 国民 劳动 劳动
成 比例 生产性 劳动者 场合 劳动 资本 资本 用途 成 比例 资本 性质 资本 方法
资本 用途 劳动量 劳动 运用 技巧 判断力 国民 劳动 管理 指导 计划 计划 国 生产
国家 政策 农村 产业 国家 政策 城市 产业 产业 发展 国家 罗马帝国 欧洲各国
政策 农村 产业 农业 城市 产业 工艺 制造业 商业 情况 规定 政策 计划 阶级 利益

So far, the task of entity extraction of the text has been performed, showing all the nouns in the text, but some entities appear in the previous text and will appear in the later text. For example, the ‘生产’ (means product) in the results in Figure 3 appears four times, which leads to the display of the results being overly complicated. Therefore, the entities need to be deduplicated and the more important keywords extracted.

3.2.1 *Word frequency statistics to extract keywords*

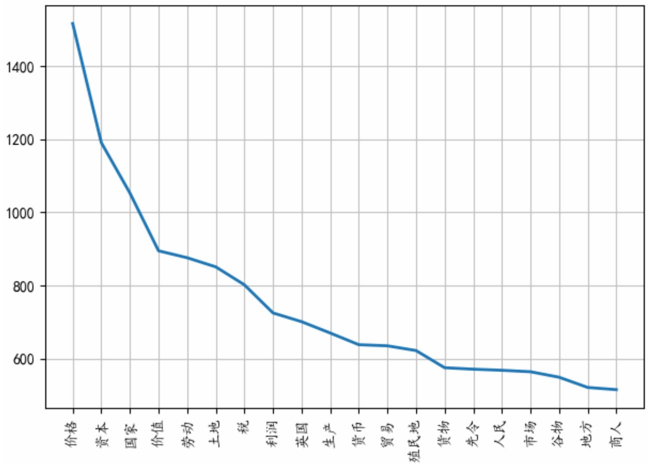
When considering keywords, it is easy to consider counting frequently occurring words. If a word is important, it should appear many times in the text (Luo et al., 2016). In this case, word frequency statistics should be conducted, duplicate words should be removed, and corresponding word frequency should be saved in a list for subsequent use.

This paper uses Adam Smith’s classic economics book, *The Wealth of Nations*, as the text resource. When the useful words in the full text are counted, there are 82,561 words in total, so the sum_of all_words = 82,561. The repeated words are combined and displayed. There are 7512 words in total, and sum_words = 7,512. Finally, the 7,512 words are sorted in descending order of word frequency, and the top 20 words and their word frequencies are selected, as shown in Table 5.

Table 5 Ranking of the top 20 keywords extracted according to word frequency

No.	*Entity*	*Word frequency*
1	价格	1,517
2	资本	1,191
3	国家	1,053
4	价值	895
5	劳动	876
6	土地	851
7	税	802
8	利润	725
9	英国	701
10	生产	670
11	货币	638
12	贸易	635
13	殖民地	622
14	货物	575
15	先令	571
16	人民	568
17	市场	564
18	谷物	549
19	地方	521
20	商人	515

Figure 4 Word frequency map of the first 20 words (see online version for colours)



We can also use matplotlib to draw a word frequency map, as shown in Figure 4. An analysis of various word frequency information shows that the word frequency of 20 entities such as ‘价格: price’, ‘资本: capital’, ‘: 国家: country’, ‘价值: value’, and ‘劳动: labor’ is very high. Therefore, it can be surmised that these words are very important in this text. Even if you did not know the title of the book, you might guess that the text is about economic and production activities. However, most of the time, there are still omissions when only considering word frequency. In observing the results, we find that the domain relevance of keywords such as ‘土地: land’, ‘人民: people’, and ‘地方: place’ are not very relevant, and there are many more important keywords that are not ranked first.

3.2.2 *Keyword extraction based on TF-IDF*

To solve the problem of omissions in keyword extraction, researchers proposed the TF-IDF method. TF-IDF combines term frequency (TF) and inverse document frequency (IDF), which is now one of the commonly used methods for keyword extraction (Yuan and Yang, 2021; Jin and Huang, 2021). Comparing the TF-IDF and word frequency results, we can see that the order of many words has changed, indicating that the TF-IDF method can improve some shortcomings of keyword extraction caused by single word frequency statistics (Shi et al., 2009). The results of the first 20 keywords extracted by the TF-IDF method are shown in Table 6.

Table 6 Sequence of the top 20 words and their TF-IDF value

No.	*Entity*	*TF-IDF value*
1	价格	0.08887180539686178
2	资本	0.07391410984042042
3	劳动	0.06366402164234607
4	先令	0.06119120417159282
5	土地	0.05748757501549806
6	殖民地	0.05747915246794372
7	价值	0.05610584896215357
8	谷物	0.05474699734990813
9	国家	0.05427670044123232
10	货物	0.05296071669740436
11	利润	0.05294712657393607
12	金银	0.04723039409078612
13	贸易	0.04569042598171523
14	英国	0.04539930034384804
15	奖励金	0.04388859654742808
16	地租	0.04386516059824712
17	商人	0.04378326954360529
18	货币	0.04366339749877595
19	产物	0.04273721139460979
20	费用	0.04079760579384881

The TF-IDF method can perform keyword extraction simply and quickly, and the extraction effect is also greatly improved; however, using only TF-IDF extraction is monotonous, and it has some shortcomings. For example, words such as '生产: production' and '市场: market' with high word frequency are not successfully extracted because they appear very few times in the corpus. That is, the words in red in Table 5 do not appear in Table 6. Due to the problem of missing important keywords, it is necessary to combine other parameters to improve accuracy. Combining multiple parameters to extract keywords is also a future trend.

Table 7 Top 20 keywords extracted by information entropy

No.	*Entity*	*Information entropy*
1	价格	0.0548763267938348
2	资财	0.0429588511433838
3	国家	0.0308138579758434
4	劳动	0.0291572718111101
5	外国	0.0228962560857599
6	价值	0.0222292993604380
7	苏**格**兰	0.0176668413168883
8	生产性	0.0160413165907247
9	利润	0.0158504954787730
10	货币	0.0148302666931297
11	用途	0.0141969191680964
12	纸币	0.0120553936447933
13	贸易	0.0118073668359013
14	货物	0.0111060870412860
15	生产	0.0101431451695770
16	商人	0.0097037780646474
17	消费	0.0092694440381464
18	商品	0.0091699392222626
19	谷物	0.0085461898162773
20	地方	0.0082869476623202

3.2.3 TF-IDF combined with information entropy extraction method

When extracting keywords, it should be considered that the word frequency is greater than a certain threshold so that the scope of the keywords can be considered. There are 7,512 words in the entity set of this text, and the highest word frequency is 1,517. Comparing the results of word frequency statistics, the words with a frequency less than 100 are mostly common nouns such as '关系: relationship' and '事情: thing', so this algorithm sets the word frequency threshold to 100. For words with a word frequency greater than 100, the information entropy reduction is calculated. In this way, the calculation formula of information entropy is related to the word probability. The greater the word frequency, the greater the possibility of the word becoming a keyword, and the greater the degree of reduction in the uncertainty of the article. Next, we consider the position of the word in the text, that is, whether the word appears in the key position, assuming that it appears 10 times in the text, so that the importance of the word in the key position can be increased. The sorting results of the obtained information entropy reduction are shown in Table 7.

We can see that the results obtained by the reduction of information entropy are somewhat different from the results obtained by TF-IDF. The reduction in the information entropy of the entity set after the word appears is calculated, and the word weight value can be appropriately changed to obtain a more accurate weight value so that the extracted keywords are more accurate, as shown in Table 8.

Table 8 TF-IDF combined with information entropy (top 20 words)

No.	*Entity*	*TF-IDF combined with I(c; x$_i$)*
1	价格	0.1437481321906970
2	劳动	0.0928212934534561
3	国家	0.0850905584170757
4	价值	0.0783351483225915
5	资本	0.0742116840947567
6	利润	0.0687976220527090
7	货物	0.0640668037386903
8	谷物	0.0632931871661854
9	先令	0.0600999526022224
10	殖民地	0.0588487942284573
11	货币	0.0584936641919056
12	资财	0.0583749435695012
13	贸易	0.0574977928176165
14	土地	0.0566479488016288
15	金银	0.0548466576695064
16	商人	0.0534870476082526
17	外国	0.0518553640932831
18	地租	0.0504612661020497
19	产物	0.0500704958679195
20	费用	0.0488593967456414

3.2.4 Analysis of comparative results of the three methods

So far, the goal of entity extraction has been achieved. From the initial word frequency extraction, through TF-IDF extraction, to the combination of information entropy extraction, the method of entity extraction has been improved step by step. Table 9 shows the comparison results of the three extraction methods.

Because the word frequency extraction considers only the word frequency, not the importance of the word itself, its results are not sufficiently accurate. The TF-IDF algorithm combines word frequency and the importance of the words. The selected keywords not only appear frequently in this paper but also appear less frequently in the corpus. Although the results of this combination are more accurate than the method of word frequency extraction, the algorithm still needs to be optimised. When the TF-IDF algorithm is combined with information entropy, the words in the key position can be fully considered, and the weight of the keywords can be appropriately changed so that the extracted results are more accurate. For example, in Table 9, No. 1 ‘价格: price’ ranks first among the three methods, indicating that the entity ‘价格: price’ appears more frequently and preferentially in the original text, *The Wealth of Nations*, The ranking is high, and the location is also very important. The No. 4 ‘价值: value’ ranks the same in the first and third methods, while in the second method ‘先令: Shilling’ ranks first. The results show that the entity’价值: value’ in’ has a higher priority than the entity ‘先令: Shilling’. This conclusion is obvious, and the third method proves this point. This sorting result has also been confirmed by experts in this field of study. Therefore, the optimised TF-IDF algorithm is shown to be the most effective method for extracting entities.

Table 9 First 20 entities after the three methods are combined in sequence

No.	*Word frequency*	*TF-IDF*	*TF-IDF and $I(c; x_i)$*
1	价格	价格	价格
2	资本	资本	劳动
3	国家	劳动	国家
4	价值	先令	价值
5	劳动	土地	资本
6	土地	殖民地	利润
7	税	价值	货物
8	利润	谷物	谷物
9	英国	国家	先令
10	生产	货物	殖民地
11	货币	利润	货币
12	贸易	金银	资财
13	殖民地	贸易	贸易
14	货物	英国	土地
15	先令	奖励金	金银
16	人民	地租	商人
17	市场	商人	外国
18	谷物	货币	地租
19	地方	产物	产物
20	商人	费用	费用

Therefore, the innovation of this paper is to consider the frequency and importance of the vocabulary in the corpus, and also consider the amount of information carried by the vocabulary to obtain a more accurate vocabulary weight, that is, to optimise the TF-IDF algorithm and combine the information entropy to obtain. The amount of information in *The Wealth of Nations*.

4 Conclusions

This study proposed a Chinese entity extraction method based on combination of Information entropy and TF-IDF, which resolves the problem of keyword extraction errors caused by single part-of-speech (POS) extraction. The results of experiments conducted on the classic economics book *The Wealth of Nations* show the effectiveness of this method. In the process of word segmentation, some user-defined dictionary content can be added, then we have made the word segmentation closer to the direction what the user wants. Therefore, it is necessary to expand the size of the corpus and research better entity extraction algorithms. In future research, we will focus on the following aspects: using rules and dictionaries in preprocessing, using deep learning to make computers better understand our language, it provide convenience for subsequent corpus construction and use.

Acknowledgements

This work was supported by the National Key Research and Development Plan (2017YFC0820603), Xinjiang University PhD Start-up Fund Project (620320015). We thank LetPub (http://www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

References

Bai, P. (2008) *Research on Web Information Extraction Technology Based on Frame Semantic Annotation*, Taiyuan University of Technology.

Ding, F., Yang, S. and Liu, R. (2014) ‘Extracting keywords in Chinese question based on average information entropy model’, *Journal of west Anhui University*, Vol. 30, No. 5, pp.46–49.

Gorla, S.K., Neti, L.B.M. and Malapati, A. (2020) ‘Enhancing the performance of Telugu named entity recognition using gazetteer features’, *Information*, Vol. 11, No. 2, pp.82–91.

Hu, S., Zhang, Y. and Zhang, C. (2021) ‘Keyword extraction research review’, *Data Analysis and Knowledge Discovery*, Vol. 5, No. 3, pp.45–59.

Jin, Y. and Huang, J. (2021) ‘Improved TF IDF algorithm based on information entropy and word length information’, *Journal of Zhejiang University of Technology*, Vol. 49, No. 2, pp.203–209.

Li, L. and He, L. (2020) ‘Keyword extraction algorithm integrating information entropy and multi-weight TF-IDF’, *Intelligent Computer and Applications*, Vol. 10, No. 9, pp.69–72, 76.

Luo, Y., Zhao, S., Li, X. et al. (2016) ‘Text keyword extraction method based on word frequency statistics’, *Journal of Computer Applications*, Vol. 36, No. 3, pp.718–725.

Moqurrab, A., Ayub, U., Anjum, A. et al. (2021) ‘An accurate deep learning model for clinical entity recognition from clinical notes’, *Journal of Biomedical and Health Informatics*, Vol. 25, No. 10, pp.3804–3811.

Qu, X. and Shen, X. (2008) 'Research on Chinese named entity recognition based on maximum entropy model', *Science and Technology Information (Doctoral Forum)*, Vol. 2008, No. 30, pp.15–17.

Shi, C., Xu, C. and Yang, X. (2009) 'Study of TF-IDF algorithm', *Journal of Computer Applications*, Vol. 29, No. S1, pp.167–170, +180.

Wang, J. (2005) *Chinese Named Entity Recognition Based on Maximum Entropy Model*, Nanjing University of Science and Technology.

Wu, H., Luo, S. and Sun, W. (2019) 'An extraction method for test keyword based on information entropy', *Computer and Digital Engineering*, Vol. 47, No. 3, pp.535–538.

Yang, K. (2015) *Research on the Algorithm of Automatic Extraction of Keywords Based on TFIDF*, Xiangtan University.

Yuan, Q. and Yang, F. (2021) 'Survey of named entity recognition', *Research and Development*, Vol. 4, pp.329–340.

Zhai, S.., Duan, H. and Li, Z. (2019) 'Knowledge graph entity extraction based on BILSTM_CRF', *Computer Applications and Software*, Vol. 36, No. 5, pp.269–274, 280.

Zhang, X., Wang, Y. and Wu, L. (2019) 'Research on cross language text keyword extraction based on information entropy and TextRank', *Information Technology, Networking, Electronic and Automation Control Conference*, Internet Information Research Institute, Communication Uni.

Zhou, R., Liu, S. and Qiu, W. (2008) 'Survey of applications of entropy in decision analysis', *Control and Decision*, No. 4, pp.361–366, 371.