Keyword Extraction and Technology Entity Extraction for Disruptive Technology Policy Texts

Aofei Chang, Bolin Hua and Dahai Yu

October 19, 2021

Keyword Extraction and Technology Entity Extraction for Disruptive Technology Policy Texts

Aofei Chang †
Department of Information Management
Peking University
Beijing, China
chaf@pku.edu.cn

Bolin Hua †
Department of Information Management
Peking University
Beijing, China
huabolin@pku.edu.cn

Dahai Yu
Department of Information Management
Peking University
Beijing, China
1900016644@pku.edu.cn

ABSTRACT

The rapid development of disruptive technologies has attracted the attention of major countries in the world in recent years, and the mining and research on the texts of disruptive technology policies of these countries can reveal the key layout, focus areas, and development pattern of each countries disruptive technology.

This article first crawls the texts of disruptive technologies from the science and technology policy websites of major countries.

Then, the text is segmented by Spacy, the segment result is filtered by a word list to construct an applicable TF*IDF matrix, and finally the matrix weights are optimized with manually collected domain core words and important words. After these, extraction and statistics of technical entity are performed according to a specified word list. Through comprehensive analysis, it can be found that the keyword hotspots of the experimental texts are focused on artificial intelligence, information security, new energy, etc.

The key areas of specific disruptive technologies are artificial intelligence, air and space, and new generation communication technologies. The result reflects the current situation and policy focus of disruptive technology development in these countries.

CCS CONCEPTS

• Computing methodologies • Artificial intelligence • Natural language processing • Information extraction

KEYWORDS

disruptive technology, science and technology policy, keywords extraction, technology entity extraction

1 Introduction

Science and technology field are changing rapidly, especially under the wave of big data and artificial intelligence. Some countries have introduced many S&T policies to promote the development of S&T. 2. Unlike academic papers, S&T policy texts do not carry keywords, so keywords need to be extracted from a large number of long texts of S&T policies. The text of S&T policy will involve both S&T projects, innovation mechanisms, transformation of results, industrial support, S&T rewards, S&T management and configuration, but also other specific S&T content, which are the research frontier, high-tech, industry common technology, disruptive technologies, and so on.

Disruptive technology was proposed by Harvard University Professor Christensen in 1997 [1] and have become a hot topic of interest for international institutions and researchers in recent years. It is generally believed that disruptive technologies are strategic innovative technologies that open up new technological tracks based on new principles, combinations and applications of S&T, and produce an overall or fundamental replacement for traditional or mainstream technologies. Disruptive technologies have strong application capabilities, can enhance the scientific and technological competitiveness of enterprises and even countries, promote the renewal of scientific and technological products, improve social production efficiency, and are expected to have far-reaching impact in many fields. Disruptive technology policies can stimulate technological innovation and provide corresponding support and guarantee, so it is necessary to study the mining of disruptive technology policies text.

Named entity extraction (NER) is a hot topic in the field of natural language processing (NLP), which specifically includes unsupervised or supervised recognition of specialized domain vocabulary such as names of people, places, time, products, and organization names, and textual keyword extraction on this basis is also an important application. Named entity extraction is usually the first step in intelligent information retrieval, relationship extraction, and even knowledge graph construction. Domain-specific entity extraction is also the key to explore the trends and hotspots in the field and to build dynamic knowledge graphs. Keyword extraction is pivotal in NLP, and a few keywords can effectively reveal the main contents and topics of long documents.

The study of S&T policy texts has also been paid attention by some scholars in recent years, Alan Porter proposed that Tech

Mining makes exploitation of text databases meaningful to those who can gain from derived knowledge about emerging technologies. [2] And he came up with “science overlay maps” as a new tool for research policy and library management. [3] Scholars such as Wen Zeng argue that China's S&T policies play an important role in promoting economic and social development, so they make use of semantic technologies to extract and analyze the relatively important information from massive S&T policies in China, but they mainly focus on exploratory terms and sentences extraction. [4] T Dmitrievna focused on the construction of a corpus in the field of S&T, he suggests that corpus linguistics tools can be applied to solve the problem of aerospace terminology which presents a good idea to better study the scientific and technical policy text. [5] Michael T. Gorczyca and other scholars stress value in leveraging language representation models (LRMs) on domain-specific text corpora for domain-specific tasks, which helps solve the problem that algorithms for developing text mining models require a large amount of training data. [6] SV Podolkova considers classification of scientific and technical texts based on the criterium of text communicative purport, he focused on structure and composition peculiarities, exact definitions and clear organization of text representation. [7]

At present, although there has been a lot of research on S&T policy, there is no specialist algorithm and thesaurus for disruptive technology policies, much less extracting keywords from these policies documents.

Keyword extraction algorithms mainly include unsupervised methods and supervised methods. Unsupervised algorithms are based on statistical features of the text, such as the TF*IDF algorithm based on the word frequency of the document collection, and Campos, Ricardo et al. had proposed the YAKE algorithm, which makes full use of the basic features of the text, it uses 5 features: “Position of Word in the text”,” Word frequency”,” Term Relatedness to Context” and “Term Different Sentence”. [8] Based on the algorithmic idea of PageRank, the graph-based keyword extraction algorithm TextRank emerged, and later Xiaojun Wan and Jianguo Xiao proposed the Expand Rank algorithm, which extended the TextRank algorithm from independent documents to a collection of similar documents. Based on TextRank.[9] Corina Florescu assign larger weights to words that are found early in a document, which is the core thought of the Position Rank. [10] The algorithm based on text features and document graph structure does not require a large amount of data annotation, but it does not completely consider the semantic relationship between words and documents, and it is difficult to make a breakthrough in accuracy.

Deep learning, as an emerging supervised approach, provides new ideas in keyword extraction. The basic approach is to vectorize the candidate words by pre-training the model with word embedding and then calculate the similarity to the document. Based on this, the Embed Rank algorithm was proposed by Bennani-Smires et al. it uses both Sen2Vec and Doc2Vec, then a score will be calculated by the Maximal Marginal Relevance (MMR) formula which take into account similarity between text and phrase as well as diversity of the keywords set.[11] After the powerful BERT model is proposed, Grootendorst proposed KeyBERT's keyword extraction algorithm in 2020, it extracts phrases that have better cosine similarity to the document vector which is produced by using pre-trained domain-specific BERT model.[12]

Although deep learning helps improve extraction precision, training such models often requires large amounts of annotated data, which is expensive to gather. In the case of keyword research in policy texts, especially in S&T policy texts, the data are mostly semi-structured long texts, and large-scale annotation training would be very difficult. In addition, for the traditional unsupervised algorithm, it will not have a high accuracy rate in a specific field, especially in our focus on S&T, and the extracted keywords are often not close to the topic of S&T.

To address these issues, we combine automatic word separation and technology domain word lists to achieve simple and fast entity extraction in the technology domain, including the extraction of named entities, and based on this, we achieve keyword extraction by the weighted optimized TF*IDF algorithm, and also perform simple statistical analysis of disruptive technology entities. Our approach combines supervised and unsupervised algorithms, and since it focuses only on the technology domain, it is relatively easy to build our corpus, and the keyword extraction based on the technology domain word list itself focuses on the thematic and semantic relationships between keywords and documents.

2 Method design

By researching major global S&T policy websites, we determined the search strategy and key websites, followed by automatic crawling of S&T policies using python to form a database of disruptive technology policy texts. Afterwards, we judged the initial data for disruptive technology relevance and screened out 1005 policy texts with high relevance. For these experimental texts, we performed automatic keyword extraction and specific disruptive technology extraction, respectively. For keyword extraction, we performed automatic word segmentation by self- designed automatic word segmentation based on spacy, then determined the TF*IDF matrix of candidate words and the overall text collection by screening the domain keyword list, and finally optimized the weights based on the self- designed core words and technology. The final results and rankings are calculated based on the self-designed core words and technology related word list for weight optimization.

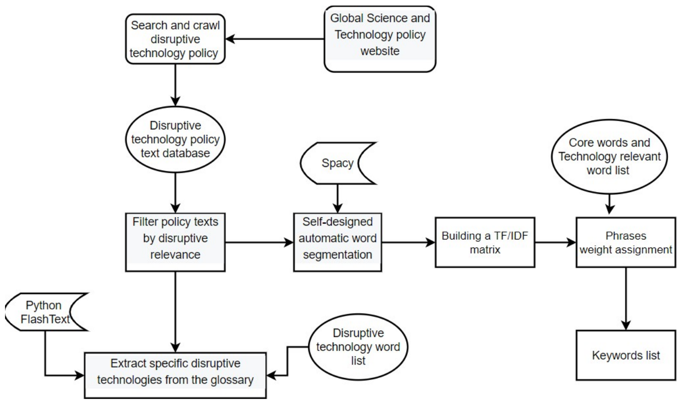

Figure 1: Main method flow chart

2.1 Text collection and relevance judgement

The main large websites crawled are the UK government website, the EU publications website, the US Center for Strategic and International Studies, etc.

In terms of relevance evaluation, considering that the occurrence of keywords is not limited to fixed pairings, for example, *disruptive* and *technology* may not appear next to each other, but if the keywords are split and then matched statistically, the accuracy rate will be reduced, so we took a compromise approach, that is, to detect whether keyword pairs appear in a sentence.

In general, the keywords in the title and abstract of an article are relatively more important, so the content of the article is divided into three parts according to the position: title, abstract and body, and the corresponding weights are 5,3,1. In addition, only considering the word frequency will ignore the influence of the length of the article on the score, so a balancing factor of the average article length is added to the formula.

Here is the formula for calculating the correlation score:

$$score = \sum_{i=1}^{N} \frac{(k+1)w_i * p_i}{k(1-b+b*l_d/l_{avg})+w_i}$$

The meaning of each of these symbols is as follows.

1. w_i: keyword weight.
2. p_i: Position weighting.
3. N: Total number of keywords in a document
4. b: A free constant that specifies how much the document length affects the score
5. k: Free constants to specify the upper limit of the impact of a single word on the rating
6. l_d: Length of the document
7. l_{avg}: Average document length

The total number of texts crawled was over 10,000, and 1005 of them were identified as experimental texts after relevance screening.

2.2 The specific process of keyword extraction

2.2.1 Words segmentation and entity extraction. Specific process: document reading and sentence slicing, using Spacy natural language processing tools, specifically using Spacy's dependent syntactic analysis to determine the predicate of each sentence, that is, the root of a sentence, as one of the basis for slicing, using Spacy's deactivation table to mark the deactivation words, and then mark some specific lexical words, specifically including'ADV','AUX','CONJ','INTJ','NUM','PRON','SYM','SCONJ','PREP'. The above words are used as cut nodes to split the words and get the preliminary splitting results.

*2.2.2 Disruptive technology lexicon construction and weighted TF*IDF calculation.* The first step is the identification of core terms, key terms and domain word lists. Considering that the topics of the texts we collected are all about technology policy and disruptive technologies, when extracting keywords we need to prioritize the words that directly match this topic, so we first defined 27 core terms as follows: *disruptive innovation, radical innovation, innovation, disruptive technology, incremental innovation, open innovation, new product development, business model, absorptive capacity, technological innovation, developing technology, advanced technology, integrated technology, future technology, promising technology, next generation technology, evolving technology, radical technology, Next Big Thing, radical technology, breakthrough technology, game changer, gaming changing technology, emerging technology, revolutionary technology, transformative technology*.

A crawler program on Web of Science was used to obtain the search results of the advanced search formula "TS = (disruptive technology OR disruptive innovation)", then we extracted all the keyword fields and performed a simple word separation process and lemmatization. Then the keyword frequency statistics were arranged in descending order. The keyword statistics were then used as the base keyword list, followed by manual identification to construct a keyword list of specific disruptive technologies. Finally, 7400 base keywords and more than 300 non-repetitive technology-specific words was collected.

Since Web of Science integrates academic journals, invention patents, academic conferences, academic websites and various other high-quality information resources to provide academic information in multiple fields, it is accurate and reasonable to grasp the dynamics of disruptive technologies and keywords in academia through Web of Science.

Finally, we constructed a TF*IDF matrix of 1005*5246 based on the results of word separation to determine the words and word frequencies for each text. For the core words, we multiplied the TF*IDF results by 20 as weights, and the words appearing in the specific technical word list were multiplied by 10 as weights. Finally, the top ten words in weight for each document were filtered as candidate keywords.

2.3 The simple process of technology entity extraction

Through the previous collection of more than 300 disruptive technology entity words, we used python's FastText tool to match these entity words to get the results of each text technology entity extraction. The process of the entity extraction part does not have much detail and focuses on the analysis of the results

3 Analysis and measurement of results

3.1 Keyword extraction algorithm measurement

In order to effectively measure keyword extraction algorithms, we will compare them with some mainstream unsupervised keyword extraction algorithms. These algorithms include: Yake, TextRank, KeyBert.

3.1.1 P/R/F-Score comparison without considering keyword order.

Table 1 P/R/F-Score comparison

Algorithm	RECA LL	PRECISIO N	F-Score
Optimized TF*IDF	0.3921	0.3229	0.3542
YAKE	0.2763	0.1855	0.2219
Text Rank	0.2159	0.1395	0.1695
KeyBERT	0.3631	0.3010	0.3291

The data show that our designed word list optimized TF*IDF algorithm has a recall rate of 0.3921, which is higher than other algorithms and has the highest recall rate, indicating that the algorithm optimized by the word list in the field of S&T has a great advantage in keyword extraction of S&T texts, which somehow makes up for the deficiency of TF*IDF in not reflecting the relationship between words and texts.

3.1.2 Comparison of result sequences considering keyword order. Considering that the measurement criteria have a certain weight order, and that the calculation results of our algorithm and other mainstream algorithms calculate scores and obtain ordered candidate keywords, the sequential order comparison can evaluate the algorithm results more comprehensively and completely. The method choice is MAP, Average Precision and Mean Average Precision.

Table 2 MAP comparison

Algorithm	MAP
Optimized TF*IDF	0.7856
YAKE	0.5164
TextRank	0.4493
KeyBERT	0.5937

In the case of considering the keyword score sequence order of the algorithm results, the TF*IDF of our word list optimization still reached the highest 0.7856, indicating that the extracted correct keywords are generally ranked high and the extracted results are more desirable.

*3.1.3 Evaluation of other unsupervised algorithms.*These unsupervised algorithms are advantageous to get high accuracy and recall without relying on domain-specific word lists, which means they are domain-independent and more applicative. However, since our proposed algorithm is supervised algorithm based on domain word lists, it has some degree of advantage in terms of accuracy and recall. If these unsupervised algorithms are optimized in combination with domain word lists, they will be more effective.

3.2 Keyword and technical entity extraction results

We conducted statistical analysis on the keyword extraction results, selected the top 120 keywords in terms of word frequency, and generated the word cloud map by python program as follows

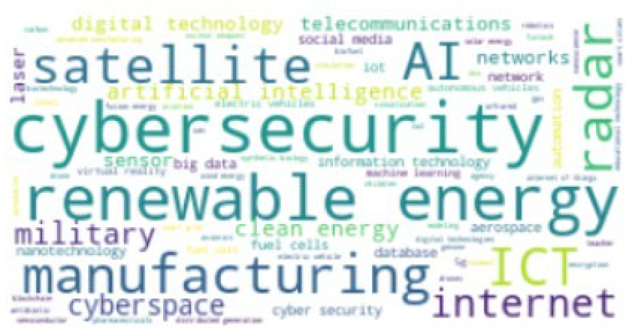

Figure 2 WordCloud of extraction result

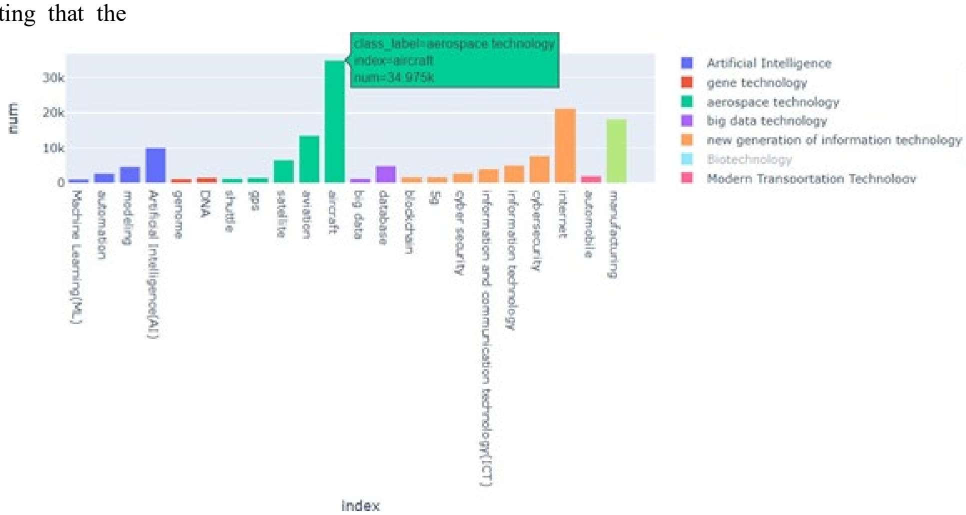

Figure 3 Classification statistics of hot technical words

Observing the word cloud map, it can be seen that the experimental text has numerous disruptive technology hotspots, involving information security, artificial intelligence, Internet of Things, new energy and even military technology fields.

We performed a simple count of high frequency words from the results of disruptive technology entity extraction and categorized them by domain.

As Figure 5 shows, the hot high frequency words in the chart mainly belong to the fields of artificial intelligence, air and space technology and new generation communication technology, such as AI and Machine Learning in the field of artificial intelligence, aircraft and satellite in the field of air and space technology and internet, ICT and 5G in the new generation communication technology, which all reflect the technology hotspots in these fields .

4 Conclusion and Discussion

As for the disruptive technology entity extraction part, the method is mainly depend on a manually designed words list, so it has great scope for improvement. There are existing excellent named entity recognition (NER) algorithms that can be utilized, for example the currently popular named entity recognition model of BILSTM-CRF via BERT pre-training. And Emma Strubell et al. first used IDCNN for entity recognition, reducing the time complexity of the former. [13]

In the keyword extraction part, considering the special characteristics of scientific and technical texts, i.e., the focus topic of the text is often the current development of a scientific and technical field or technology, we enlarge the weight of scientific and technical terms so that the keyword extraction procedure can be more sensitive to these field terms and often extract the less frequent but important scientific and technical field terms. The results of the automatic word screening are intuitive and measured well because of the use of a manually developed word list and the increased weight of technical terms in the calculation of TF*IDF.

However, considering the subjective nature of the manually developed word list, the fact that the content of the word list limits the determination of candidate keywords and phrases, and the fact that the field of S&T is developing rapidly and the nouns of S&T are changing rapidly, it is not possible to extract technical noun entities and candidate keywords by relying on the manual word list alone.

The keyword algorithm is insufficient: it does not have the ability to extract new words, and the extraction results are overly dependent on the word list, so it needs a great degree of optimization. Optimization direction: Later on, neural networks and deep learning algorithms can be used to extract a wider range of noun entities in the field of S&T in combination with the word list, and the ability to extract words outside the word list is enhanced.

In specific disruptive technology extraction part, the extraction in this paper relies entirely on the manually developed word list, and later we can also rely on deep learning algorithms for optimization. Through stronger autonomous extraction, statistical analysis, we will optimize discovery of hotspots in S&T fields and make preparation for domain knowledge extraction.

ACKNOWLEDGMENTS

This work was supported in part by The National Social Science Fundation of China (Number: 17BTQ066).

REFERENCES

[1] Christensen, Clayton. "M.(1997) The Innovators Dilemma: When New Technologies Cause Great Firms to Fail." Harvard Business Review Press (1997).

[2] Porter, Alan L., and Scott W. Cunningham. Tech mining: exploiting new technologies for competitive advantage. Vol. 29. John Wiley & Sons, 2004.

[3] Rafols, Ismael, Alan L. Porter, and Loet Leydesdorff. "Science overlay maps: A new tool for research policy and library management." Journal of the American Society for information Science and Technology 61.9 (2010): 1871-1887.

[4] Zeng, Wen, Changqing Yao, and Hui Li. "The exploration of information extraction and analysis about science and technology policy in China." The Electronic Library (2017).

[4] Butenko, Iuliia Ivanovna, Tatiana Dmitrievna Margaryan, and Elizaveta Evgeneevna Bolotova. "Scientific and Technical Text Corpus as the Basis for Aerospace Terminology Standardization." Applied Linguistics Research Journal 5.3: 113-119.

[6] Gorczyca, Michael T., et al. "A comparison of language representation models on small text corpora of scientific and technical documents." Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II. Vol. 11413. International Society for Optics and Photonics, 2020.

[7] Podolkova, S. V. "Principles of Scientific and Technical Text Analysis." Філол о г і ч н і трактати 10,№ 2 (2018): 90-100.

[8] Campos, Ricardo, et al. "YAKE! Keyword extraction from single documents using multiple local features." Information Sciences 509 (2020): 257-289.

[9] Wan, Xiaojun, and Jianguo Xiao. "Single Document Keyphrase Extraction Using Neighborhood Knowledge." AAAI. Vol. 8. 2008.

[10] Florescu, Corina, and Cornelia Caragea. "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.

[11] Bennani-Smires, Kamil, et al. "Simple unsupervised keyphrase extraction using sentence embeddings." arXiv preprint arXiv:1801.04470 (2018).

[12] Grootendorst, M. "KeyBERT: minimal keyword extraction with BERT, v0. 1.3." (2020).

[13] Strubell, Emma, et al. "Fast and accurate entity recognition with iterated dilated convolutions." arXiv preprint arXiv:1702.02098 (2017).