Research on technical architecture of knowledge extraction in vertical field

Jing Xu, Wangqun Lin*, Chengping Tian, Peng Sun, Baoyun Peng, Yu Tian
Consulting Center for Strategic Assessment, Academy of Military Sciences, Beijing 100097, China

ABSTRACT

How to extract high-value knowledge quickly, accurately and comprehensively from high volume, and high variety data in vertical field, is a major issue to be solved in the intelligent construction of various fields. Knowledge extraction, which extracts large-scale structured knowledge from multi-source heterogeneous data, can provide a strong support for many applications, such as knowledge graph construction, automatic question answering, situation awareness and intelligent decision. In this paper, we first summarize the critic problems and challenges of knowledge extraction in vertical field by surveying the key technologies involved in knowledge extraction. Then, we propose a unified framework of knowledge extraction targeting for vertical field. This framework presents a detail solution for the key problems in knowledge extraction, and indicates the potential research work to be carried out in the future.

Keywords: Vertical field, knowledge extraction, technical architecture.

1. INTRODUCTION

Many vertical fields such as finance, medicine, manufacturing, logistics, electric power, and military have accumulated a large amount of structured, semi-structured and unstructured data, such as databases, tables, format messages, text, images, audio and video. These data are massive, complex and diverse, which have brought about serious information overload problems, making users easily fall into the dilemma of "information lost" and "knowledge shortage". How to quickly, accurately and comprehensively extract high-value knowledge from industry data with a wide range of sources, large scale, scattered distribution and diverse modalities, is a major issue to be solved in the intelligent construction of various fields. This knowledge can support various users to accurately obtain data according to different tasks, and provide real-time situational awareness and auxiliary command decision-making.

2. KNOWLEDGE EXTRACTION

Knowledge extraction technology is an important means to solve the problem of information overload for huge amounts of heterogeneous data. As shown in Figure 1, knowledge extraction technology, which extracts structured knowledge such as concepts, entities, attributes, relations, events from structured, semi-structured, and unstructured data, provides a strong support for many applications, such as knowledge graph construction, semantic search, intelligent question answering, and intelligent decision. The research on knowledge extraction technology for vertical field is helpful to enhance the high usability and reusability of industry information, and realize the advantages of data-driven and knowledge-centered intelligent decision making.

Knowledge extraction technology mainly includes many methods based on rules, patterns, machine learning models, and deep learning models[1-11]. Based on manual definition of patterns and rules by analyzing the location structure, word composition, occurrence frequency and other characteristics of knowledge, it aims at precise knowledge extraction rather than wide recall. The machine learning theory-based methods, which convert a knowledge extraction task into a multi-classification or sequence labelling task, train machine learning models (such as Conditional Random Field, Hidden Markov Model, Support Vector Machine) with feature set to perform the target tasks. Due to its powerful ability to automatically capture features, deep learning technology is widely used in knowledge extraction tasks by training neural network models (such as Convolutional Neural Networks, Recurrent Neural Networks, Long-Short Term Memory Network) using a large amount of labelled data. Table 1 shows the advantages and disadvantages of the above three mainstream knowledge extraction methods.

[*] linwangqun2005@163.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250669
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2661950

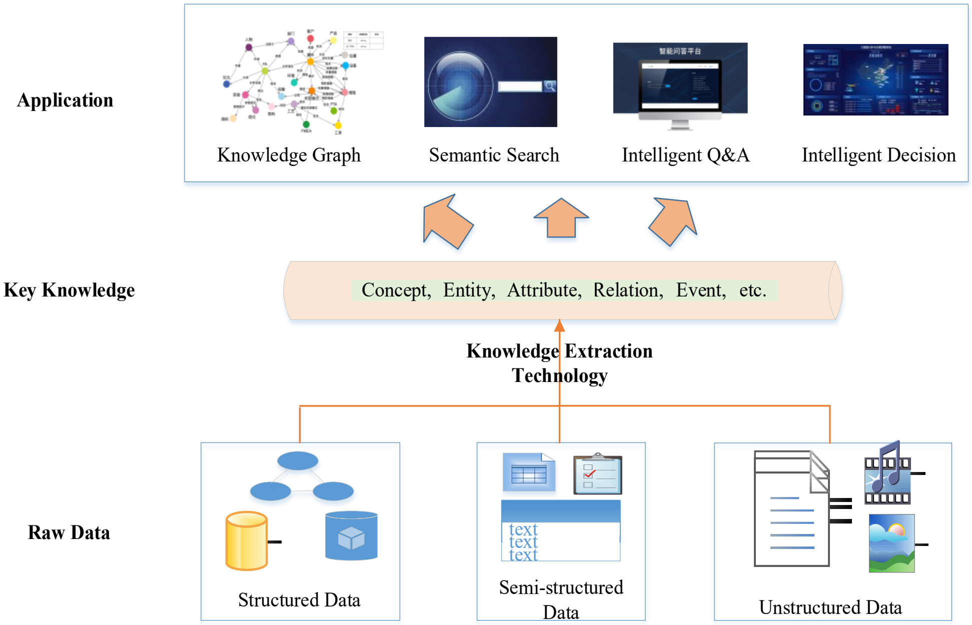

Figure 1. Knowledge extraction architecture.

Table 1. Advantages and disadvantages of mainstream methods of knowledge extraction.

Methods	Advantages	Disadvantages
Knowledge extraction methods based on rules or patterns[1-3]	The accuracy of knowledge extraction results is high on limited normalized data.	Rules and patterns are expensive to formulate and difficult to extend to different tasks and fields.
Knowledge extraction methods based on machine learning models[4-7]	The methods are flexible and robust.	Training models requires a large number of labeled data and feature sets, resulting in higher cost.
Knowledge extraction methods based on deep learning models[8-11]	The ability to automatically capture features is strong, reducing the cost of manually formulating features.	Large-scale labeled data is required to train the learning models, and the processing process is a black-box mode, and the output results lack interpretability.

3. CRITICAL PROBLEMS AND CHALLENGES OF KNOWLEDGE EXTRACTION IN VERTICAL FIELD

Structured and semi-structured data in vertical fields usually have regular structure and fixed format, from which extracting knowledge is relatively easy. But the amount of knowledge acquired is limited, and it is difficult to meet the needs of large-scale knowledge application. Therefore, it is necessary to extract knowledge from unstructured data to expand the scale. Though unstructured data contains rich knowledge, there are diverse structures, insufficient samples, unbalanced categories, concise context and other problems. These problems propose new challenge to knowledge extraction technology.

(1) Diverse data structures

There are many data sources in vertical fields, including internal industry data, external open-source data, etc. Data from different sources are different in structure, expression and quality, containing a large amount of text, image, audio, video and other unstructured data. To extract knowledge from multi-source heterogeneous data sources, it is necessary to analyze the data structures of different data sources to design corresponding extraction methods, or to standardize them into a unified data format. This undoubtedly adds difficulty to knowledge extraction technology.

(2) Insufficient data samples

The data in vertical fields contains a lot of complex knowledge with multi-granularity. In order to train the algorithm model to mine knowledge efficiently, a large number of annotated samples is needed. However, it is difficult to generate large-scale and high-quality samples, due to the contradiction between highly professional industry data and insufficient annotation power of relevant professions. Therefore, knowledge extraction technology generally faces a problem of lack of a large number of samples or only small samples, which makes it difficult to obtain various knowledge efficiently and accurately.

(3) Unbalanced data categories

The data in vertical field tends to have a long tail distribution. There are more samples for common data categories (header categories), as well as small samples for rare data categories (tail categories). The phenomenon is easy to lead to the algorithm model to describe the sparse samples insufficiently, thus resulting in the difficulty of knowledge modelling. The problem of unbalanced data categories also easily causes deviating the decision boundary of the algorithm models, which makes the training results overfit multi-sample data categories, reducing the overall performance of the algorithm models.

(4) Concise data context

There is a large amount of highly abstract and concise textual corpus, such as clause, directive, regulation. This kind of data usually omit part of the sentence, resulting in the lack of context information. It not only restricts the application of knowledge extraction methods that depend on context information, also reduces the accuracy of text parsing and pattern matching. Therefore, the extraction results are easy to be incomplete and missing.

4. TECHNICAL ARCHITECTURE FOR KNOWLEDGE EXTRACTION IN VERTICAL FIELD

Focusing on the above problems and challenges, and fully considering the industry data structure, category, scale, quality and other factors, as well as the advantages of the current mainstream technology, the technical architecture of knowledge extraction in vertical field is designed as shown in Figure 2. The solution is as follows.

(1) Data preprocessing

In order to standardize data expression and improve data quality in vertical field, it can be considered to add data cleaning, data integration, text conversion, filtering stop words to the industry data from different sources and different structures in the data preprocessing stage. For example, removing noise and repeated data, converting images, audio and video data into corresponding text, and removing some useless or meaningless words in the text corpus. Through data preprocessing, the burden of knowledge extraction technology can be reduced and the accuracy of extraction results can be improved

(2) Data labelling

In order to solve the problem of insufficient data samples and category imbalance, the high-quality, reliable and high-precision industry knowledge is extracted as the seeds, from structured and semi-structured data sources (such as databases, tables) in vertical fields. Moreover, combining with expert experience and knowledge, the methods based on small sample, transfer learning or adversarial learning are studied to guide data annotation and train sample generation. For example, the common category knowledge is transferred to the corresponding rare category, and the rich prior knowledge is used to help them learn, so as to generate virtual rare category samples that conform to the distribution and realize the transferability of small sample knowledge. Adversarial learning method is used to add noise data in generating training data, to simulate the situation of training data adding random noise. On the one hand, it is helpful to solve the imbalance of data categories in vertical fields. On the other hand, it can improve the generalization and robustness of the algorithmic models.

(3) Feature Extraction and Vector Representation

In order to solve the problem of concise data context and improve the performance of knowledge extraction models, the shallow language features and statistical features independent of context information are extracted from corpus. In addition, the features and corpus are represented as numerical vectors with the discrete representation and distributed representation methods. They are used for the input data of machine learning and deep learning models, to alleviate the limitation of the models caused by missing context information.

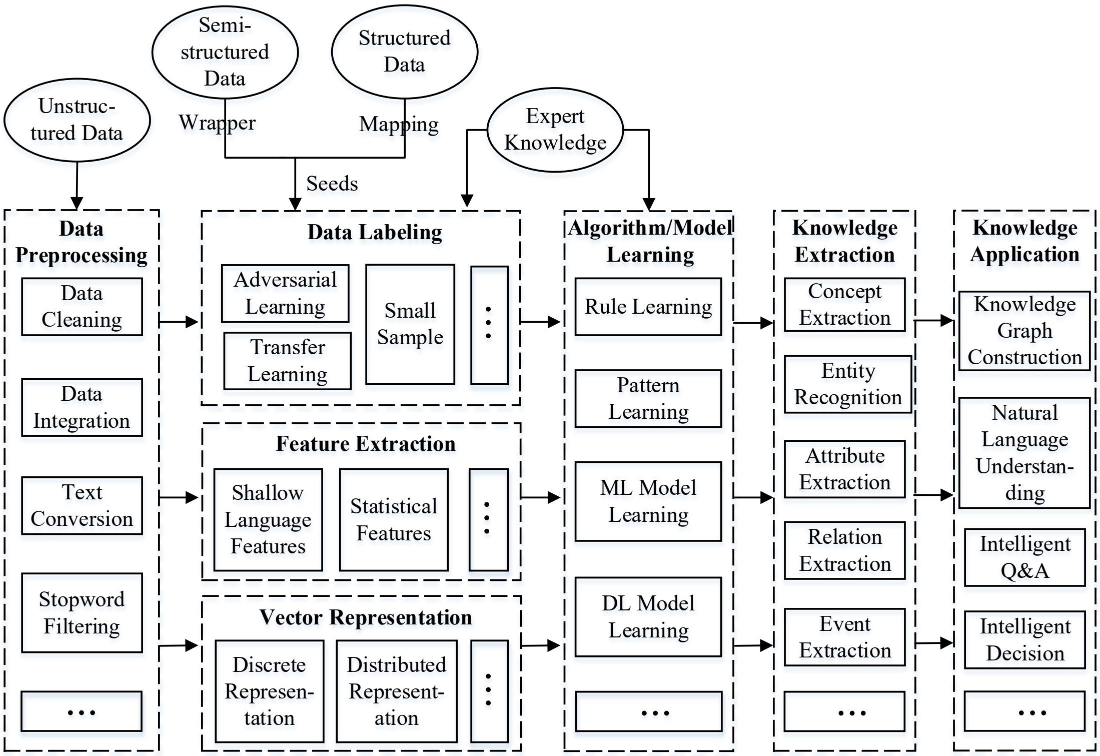

Figure 2. Technical architecture for knowledge extraction in vertical field.

(4) Algorithm/model learning

The annotation data and feature set generated in the previous stage are utilized, with the vector representation being transformed, to learn the extraction rules and patterns, as well as train the machine learning and deep learning models, such as Regular Expression, Hearst Pattern, Hidden Markov Model, Support Vector Machine, BiLSTM (Bi-directional Long-Short Term Memory)-CRF (Condition Random Field), Multi-Level Attention CNNs (Convolutional Neural Networks). These algorithms and models are used to extract knowledge by supervised or unsupervised methods.

Through the above stages, multiple knowledge such as concepts, entities, attributes, events and their associated relations can be extracted from large-scale multimodal data. These strategies can solve the problems of industry knowledge extraction under the environment of massive multi-source heterogeneous data. They can also provide support for large-scale knowledge graph construction, natural language understanding, intelligent question-answering, intelligent decision-making and other applications.

5. CONCLUSION

Knowledge extraction technology can automatically mine valuable knowledge from massive, scattered and heterogeneous industry data, which plays an important role in natural language understanding, information mining, situation awareness and intelligent command decision-making, accelerating the arrival of knowledge-based intelligent construction era. This paper analyzes the key problems and challenges of knowledge extraction technology in vertical field, and designs a technical framework of knowledge extraction targeting for vertical field. This framework presents a detail solution for the key problems in extracting industry knowledge. In vertical field, the problems such as insufficient samples and imbalance categories are common, and the knowledge extraction methods based on small samples and transfer learning will become hot topics in the future.

REFERENCES

[1] Astrakhantsev, N., "Automatic term acquisition from domain-specific text collection by using Wikipedia," *Proceedings of the Institute for System Programming of RAS,* 26(4), 7-20(2014).

[2] Sari, Y., Hassan, M. F. and Zamin, N., "Rule-based pattern extractor and named entity recognition: A hybrid approach," *2010 International Symp. on Information Technology,* 563-568(2010).

[3] Liu, X. and Yu, N., "Multi-type web relation extraction based on bootstrapping," *2010 WASE Inter. Conf. on Information Engineering,* 24-27(2010).

[4] Kambhatla, N., "Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction," *Annual Meeting of Association of Computational Linguistics,* 178-181(2004).

[5] Saha, S. K., Sarkar, S. and Mitra, P., "Feature selection techniques for maximum entropy based biomedical named entity recognition," *Journal of Biomedical Informatics,* 42(5), 905-911(2009).

[6] Culotta, A. and Sorensen, J., "Dependency tree kernels for relation extraction," *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04),* 423-429(2004).

[7] Poulymenopoulou, M., Malamateniou, F. and Vassilacopoulos, G., "Machine learning for knowledge extraction from PHR big data," *Stud. Health. Technol. Inform.,* 202(1), 36-39(2014).

[8] Zhao, H. and Wang, F., "A deep learning model and self-training algorithm for theoretical terms extraction," *Journal of the China Society for Scientific and Technical Information,* 37(9), 923-938(2018).

[9] Luo, L., Yang, Z., Yang, P., et al., "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," *Bioinformatics,* 34(8), 1381-1388(2018).

[10] Peng, N., Poon, H., Quirk, C., et al., "Cross-sentence N-ARY relation extraction with graph LSTMs," *Transactions of the Association for Computational Linguistics,* 5(1), 101-115(2017).

[11] Lin, Y., Ji, H., Huang, F., et al., "A joint neural model for information extraction with global features," *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics,* 7999-8009(2020).