Scientific Literature based Big Data Analysis for Technology Insight

Hui Gao¹,², Lin Gui¹ and Wei Luo²

¹Department of Computer Science, National University of Defense Technology, Changsha China
²China Defense Science & Technology Information Center, Beijing China

¹Corresponding author: gaohui_baixiang@163.com

Abstract. The increasingly massive amount and open access of literature provide a data foundation for technology insight based on big data analysis. This paper proposed a new technology insight framework - Technology Dependency Graph (TDG). Firstly, an adversarial multitask learning and distantly-supervised learning were applied to extract the technology entity and dependency relation with limited labeled sample. Then, a TDG was constructed with the entities as vertices and the dependency relations as edges. A TDG contains rich and valuable semantic information which represents the support, contribution or relying on relationship between technologies. At the same time, the social network properties of TDG allow researchers to analyze and mine hot topics, key technologies, and technology architecture by using network theories, methods and tools. In the case study, the TDG of DSSC (dye-sensitized solar cell) was constructed. Furthermore, the technology dependency architecture for DSSC was constructed according to a spanning tree out of the TDG, which provides a global perspective for the research of DSSC.

1. Introduction

Currently, researchers have access to a huge and rapidly growing amount of scientific literature available on-line [1]. Recent estimates reported that a new paper was published every 20 seconds [2]. At the same time, during the last few years the number of scientific papers that were freely accessible online considerably grew. Sometimes between 2017 and 2021, more than half of the global papers are expected to be published as Open Access articles [3]. The massive amount and open access of literature guarantee the feasibility of big data analysis.

The scientific literature not only contain the authors' original research, but also a large number of reviews and summaries of previous research works. From the beginning of the 20th century, researchers began to use the bibliometrics method for technology insight. However, bibliometrics generally focuses on the metadata to find co-citation and co-occurrence information, lacking deep mining for the content.

The achievement of big data analysis and artificial intelligence technologies has provided new routes for literature-based technical insights, which are two aspects: the data of interest is not only the metadata, but also the unstructured full text of the literature; the representation of knowledge is no longer co-citation or co-occurrence information, but the knowledge graph containing various technology entities and semantic relationships between technology entities.
Related researches mainly come from two research areas: management and computer science. The former often uses simple text mining methods, such as: clustering [4], rules-based methods [5][6], etc., to extract keywords [7] or technical concepts; but its advantage is that it can analysis text mining results with mature technical insight methods, such as technology roadmap [5], network theory [7], technology life cycle theory [8], TRIZ [4], etc., to obtain more professional conclusions. The latter has the advantage of strong information extraction ability, and can extract richer semantic information based on various methods such as LDA [9], CRF [10], and machine learning [11]. However, its disadvantage is the lack of technical analysis theory, so only some preliminary conclusions can be drawn in technical analysis.

In this paper, we applied deep learning approaches to solve the problem of text mining with only a small number of labeled samples, and we implemented the extraction for technology entity and dependency relation between entities. A weighted directed graph was constructed with the technology entities as vertices and the dependency relations as edges, which is called as Technology Dependency Graph (TDG). A TDG contains rich and valuable semantic information which represents the support, contribution or relying on relationships between technologies. At the same time, the social network properties of TDG allow researchers to analyze and mine technologies by using network theories, methods and tools. The main contributions of this paper are: 1) we constructed a comprehensive architecture for building and defining TDG; 2) we discussed the potential applications of TDG for technology insight, and applied TDG into technology architecture generation.

2.System architecture

2.1 Basic concepts

**Technology entity**: There were many different expressions for technology in previous researches, such as technology concept, technology entity, technical terms, and key phrase. Because the TDG is a special knowledge graph, the expression of technology entity was adopted to represent the technology in paper.

**Dependency relation**: The dependency relations refer to the semantic relations of promotion, dependence, contribution, use, etc. between two different technology entities. If the technology entity $v_1$ has dependency relation with the technical entity $v_2$, it can be expressed as $<v_1, v_2>$.

**Technology Dependency Graph**: The TDG is a weighted directed graph with the technology entities as vertices and the dependency relations as edges. The TDG can be expressed as $G = <V, E>$, where the vertices set is represented as $V = \{v_1, v_2, v_3, ...\}$ and edges set is represented as $E = \{e_1, e_2, e_3, ...\}$. The edge $e_i$ from $v_i$ to the $v_j$ represents that technology of $v_j$ depends on technology of $v_i$. And the weight of $e_i$ is expressed as $W(e_i)$, which equals the number of times the corresponding dependency relationship $e_i$ appears in the literature.

2.2 Architecture

The architecture of TDG consists of three modules – information extraction module, TDG construction module and technology analysis module.

2.3 Information extraction module

Entity recognition and relation extraction are both sub-tasks of information extraction. The biggest challenge for information extraction from scientific literature is the lack of labeled corpus, and large-scale manual annotation is very costly. This paper adopted semi-supervised, weak supervision and transfer learning to realize the extraction of technology entities and dependency relations based on a small amount of labeled corpus.

2.3.1 Entity extraction

A corpus for task ScienceIE [12] at Semeval2017 was built which consists of 500 journal articles distributed among the domains Computer Science, Material Sciences and Physics. Based on ScienceIE corpus, we develop a Neural Network of transfer learning to apply the knowledge gained in the three
source domains above to recognize entities in new domains (e.g., solar-energy). This model is a multi-
tasks framework which consists of a target task and two auxiliary tasks:

**Target task**: Technology entity recognition for target domain based on a small number of labeled
samples.

**Auxiliary task1**: Science keyphrase recognition for the source domain.

**Auxiliary task2**: Inspired by the work on domain adaptation [13], an adversarial task is introduced
as a domain discriminator to recognize whether source domain or target domain the input sample is
belong to. When the domain discriminator cannot distinguish the domains of the input sample, it
means that the model learns the sharable features, and ensuring better transfer learning ability.

### 2.3.2 Relationship extraction

Distantly-supervised information extraction has been proven to be effective in many researches when
the labeled data set is small. A knowledge base (KB) of relation or concept instances is used to train a
distantly-supervised IE system, but in many cases the knowledge base is lacking or incomplete: e.g.,
there is no release of a dependency relation knowledge base. Bing [14] extracted seeds from the well-
structured corpus, and then extended a complete instance-base through the semi-supervised method of
label extension, combined with the distantly-supervised method to extract the medical concepts and
relationships. Inspired by the work of Bing, we present a method which combines bootstrapping and
distantly-supervised to extracting technology dependency relation. Firstly, based on a small amount of
well annotated corpus, bootstrapping method is used to generate a technology dependency relation
instance-base that is sufficient to support distantly-supervised learning. Then, the distantly-supervised
relation extraction model is trained to achieve the generalization of Bootstrapping.

### 2.4 TDG construction module

The construction for TDG consists of two tasks: entity linking and weight assignment. Entity linking is
the task to link entity mentions in text with their corresponding entities in a knowledge base [15], which
will be used to filling the extracted technology entities and dependency relations to the correct
positions in the TDG. Because the expression in the scientific literature is rigorous, there is little
ambiguity about the full name of the technology entity. The main challenge of entity linking is to find
the correspondence between the abbreviation and the full name of the same entity, and we use a simple
pattern matching method [16] to achieve this correspondence. The weight of the directed edge in TDG is
assigned by the number of scientific literature that contains the corresponding dependency relation.

### 2.5 Technology insight based on TDG

As a weighted directed graph, the TDG has rich properties of social networks that allow researchers to
accomplish technology insight by using theories, methods and tools of graph, social network and even
traditional bibliometrics. This section will explore some technical insight scenarios of TDG.

**Path based technology architecture construction**: In graph theory, a directed path in a graph is a
directed sequence of edges which connect a sequence of vertices which are all distinct from one
another. The terminal vertex in a path is pointed to by all other vertices, which means that the
technology of the terminal vertex is depend on the technologies of all the other vertices in a path. All
the paths with a certain technology as the terminal construct a complete technical dependency
architecture which provides a global perspective for the certain technology.

**Centrality based evaluation of technical importance and maturity**: In social network analysis,
indicators of degree centrality and betweenness centrality can identify the importance of vertices
within a graph. The degree increases as the frequency of occurrences of corresponding technology
increases in the literature, which means that we can obtain the hottest technology by finding the vertex
of the highest degree. As the maturity of a technology increases, its application becomes more
extensive, and the outdegree of corresponding vertex will increases. Therefore, the ratio of the
outdegree and the intdegree can reflect the maturity of a technology. Betweenness centrality quantifies
the number of times a node acts as a bridge along the shortest path between two other nodes. In a TDG, betweenness can be used to find the key technologies.

**Structural equivalence based discovery for Alternative technology:** In a network, if two actors do not change the structure of the entire network after replacing each other, the two actors are structurally equivalent. The structural equivalence can measure the similarity of relational patterns or network positions between two vertices. In a TDG, alternative technologies can be discovered according to the structural equivalence between technologies, which help researchers to bypass insurmountable technical difficulties in technological innovation activities.

**Community Evolution based technology trend prediction:** In the TDG, technologies similar in disciplines, fields, applications, etc. are grouped into communities. By studying the birth, merger, decomposition, and extinction of these communities in the time dimension, we can predict the technology trend. For example, it is indicated the emergence of a new interdisciplinary when two different communities are merging together.

3. Case study

This empirical study applied the proposed framework to literature of related to DSSC (dye-sensitized solar cell). We retrieved 473935 related articles with keyword of 'dye-sensitized solar cell' from an aggregation digital library (consisting of INSPEC, WPI, Elsvier, Springer, etc.). All were published in between 1999 and 2015.

3.1 Construction for TDG of DSSC

We employed a framework of adversarial multitask learning introduced in 3.3.1 to extract technology entity. The corpus of source domain was released by Semeval2017 [12]. The corpus of target domain was about DSSC built by manual annotation from 20 articles. We implemented the extraction based on the work of Chen [17]. For the dependency relation extraction, we also labeled 38 relations instances from 20 articles, and expanded to 573 instances by bootstrapping. After that, we applied the distantly-supervised learning to extract the dependency relations between technologies. At last, we obtained a TDG of DSSC which consists of 472 vertices and 1027 edges. Being rendered by Gephi, the graph is shown in figure 1.

![Figure 1. The TDG of DSSC. Every node represents a technology, and the directed edge represents the dependency relation between technologies.](image-url)
3.2 Technology insight

Figure 2. The dependency architecture of DSSC. For example, DSSC depends on anodes, which in turn depend on nanocrystalline technology (shown in the red wireframes).

In this section, we applied the TDG to analyze the technology architecture of DSSC. Firstly, we got a spanning tree with DSSC as the terminal out of the TDG using Breadth First Search. In this spanning tree, all other technologies point to DSSC with a hierarchical structure (Figure 2.), which means the technology of DSSC is directly or indirectly depended on the other technologies. So, the spanning tree represents the technology architecture of DSSC. In order to make a clear description, we converted the spanning tree into the table 1., which contains the dependency relations whose weight bigger than 10. Based on expert knowledge and literature verification, it was confirmed that the technology architecture of DSSC obtained by this analysis method was close to the actual situation.

Table 1. Table of the dependency architecture for DSSC

| dssc             | window layer | electrolyte                   | photocatalyst            | buffer layer | interlayer | antireflection coating | absorber layer | anodes | active layer |
|------------------|--------------|-------------------------------|--------------------------|--------------|------------|------------------------|----------------|--------|-------------|
|                  | thin-film cds, flexible plastic substrates, gallium arsenide, microcrystalline silicon | potassium iodide, electron donor mediator, polysulfide, ionic liquid | nanoparticles, titanium oxide, optical radiation, quantum dots, titanium dioxide, znin2s4 | znse, zn-doped cigs layer, znte, porous silicon, znpc, znpo particles, azo films, bathocuproine, cds nano-layer, | copolymer encapsulant, transparent multiwalled carbon nanotube sheets, carrier transportation process | pcvd-grown sinx film, silicon nitride films, thin nanoporous layer, moth-eye antireflection nanostructures | ternary chalcopyrite semiconductor cu, single-phase sns films, colloidal nanocrystals, evaporated sns | titania, ruthenium, nanocrystalline, nanowire | novel cyanine-fullerene dyad, vhf-gd technique, bulk heterojunction structure |
4. Conclusion
This paper introduced Technology Dependency Graph (TDG), a big data analysis framework could facilitate the technology insight from scientific literature. We applied deep learning methods to implement the construction of TDG, and discussed the potential applications of TDG for technology insight. Several extensions of the framework we presented can be performed in future works. The types of technology (such as PROCESS, ATTRIBUTE, and MATERIAL) should be identified further to enrich the semantic information in TDG. And we will do more case studies to evaluate the application of TDG for technical insights.

References
[1] Ronzano F, Saggion H. Knowledge Extraction and Modeling from Scientific Publications// Semantics, Analytics, Visualization. Enhancing Scholarly Data, pp.11-25, 2016.
[2] Munroe R. News: the rise of Open Access. Science, 342(6154), pp.58-59, 2013
[3] Lewis, D. W. The inevitability of open access. College & Research Libraries, 73(5), pp.493-506, 2012.
[4] Zhang Y, Zhou X, Porter A L, et al. Triple Helix innovation in China’s dye-sensitized solar cell industry: hybrid methods with semantic TRIZ and technology roadmapping. Scientometrics, 99(1), pp.55-75, 2014.
[5] Yang C, Zhu D, Zhang G. Semantic-Based Technology Trend Analysis// International Conference on Intelligent Systems and Knowledge Engineering. IEEE, pp.222-228, 2015.
[6] Guo J, Wang X, Li Q, et al. Subject–action–object-based morphology analysis for determining the direction of technological change. Technological Forecasting & Social Change, 105(4), pp.27-40, 2016.
[7] Silva F N, Amancio D R, Bardosova M, et al. Using network science and text analytics to produce surveys in a scientific topic. Journal of Informetrics, 10(2), pp.487-502, 2016.
[8] Durmuşoğlu A. A pre-assessment of past research on the topic of environmental-friendly electronics. Journal of Cleaner Production, 129, pp.305-314, 2016.
[9] Canobasave A E, Osborne F, Salatino A A. Ontology Forecasting in Scientific Literature: Semantic Concepts Prediction Based on Innovation-Adoption Priors. EKAW 2016, pp.51-67, 2016.
[10] Prabhakaran V, Hamilton W L, Dan M F, et al. Predicting the Rise and Fall of Scientific Topics from Trends in their Rhetorical Framing[C]// Meeting of the Association for Computational Linguistics. pp.1170-1180, 2016.
[11] Mckeown K, Daume H, Chaturvedi S, et al. Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science & Technology, 67(11), pp.2684-2696, 2016.
[12] Augenstein I, Das M, Riedel S, et al. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. CoRR abs/1704.02853, 2017.
[13] Ganin Y, Ustinova E, Ajakan H, et al. Domain-Adversarial Training of Neural Networks. Journal of Machine Learning Research, 17(1), pp.2096-2030, 2017.
[14] Bing L, Ling M, Wang R C, et al. Distant IE by Bootstrapping Using Lists and Document Structure, AAAI'16, 2016.
[15] Shen W, Wang J, Han J. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. Knowledge & Data Engineering IEEE Transactions on, 27(2), pp.443-460, 2015.
[16] Adar E, Datta S, Building a Scientific Concept Hierarchy Database (SCHBASE), In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, China, pp. 606-615, 2015.
[17] Chen X, Shi Z, Qiu X, et al. Adversarial Multi-Criteria Learning for Chinese Word Segmentation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1193–1203, 2017.