ABSTRACT

We present a deep learning-based technology fitness landscape premised on a neural embedding space of 1,757 technology domains and their respective improvement rates. The technology embedding space is a high-dimensional vector space trained via applying neural embedding techniques to patent data. The improvement rates of respective technology domains are drawn from a prior study. The technology fitness landscape exhibits a high hill related to information and communication technologies (ICT) and a vast low plain of the remaining domains. The technology fitness landscape presents a bird’s eye view of the structure of the total technology space, a new way to interpret technology evolution with a biological analogy, and a biologically-inspired inference to next innovation.

Keywords Technology landscape · Technology improvement · Innovation management · Deep Learning

1 Motivation

In the past decade, artificial intelligence, cloud quantum computing and 5G communication technologies undergo rapid advances. Meanwhile, tremendous innovations also emerge and gain momentum in traditional technological domains, such as autonomous vehicles, drug discovery and protein structure prediction. Such contemporary innovation phenomena call for new theories and frameworks to explain them, understand the driving forces, and inform future innovation.

Many contemporary innovations share one characteristic in common: they are based on the synthesis and fusion of different technological domains, which used to be unrelated and separately developed, e.g., artificial intelligence and automobile. The rise of such innovations has ambiguated the boundaries of technological domains and industries. For instance, should an autonomous vehicle be classified as an automotive or artificial intelligence product? Should DNA data storage be defined as a biological or information technology? More holistic assessment of technological domains in one integrated space is demanded to help interpret their evolution and find next innovation directions.

Recently, Singh et al. [1] statistically estimated the improvement rates for each of the 1,757 technological domains based on patent classifications (Hereafter, we call this paper STM). Using STM’s technological domain definitions [1], we train a high-dimensional embedding vector space of the 1,757 domains based on multimodal patent data and further integrate the domains’ improvement rates (as indicators of respective domains’ fitness in the total technological system) to generate a rugged technology fitness landscape. The shape of the technology fitness landscape allows for a holistic understanding of the structure and evolution dynamics of the total technology space and a context-aware understanding of the evolution prospects of individual technological domains. This was not previously possible as the technology domains were then analyzed discretely or associated in a single-or low-dimensional space (e.g., citation information only).

Our technology fitness landscape creation draws analogical inspiration from Kauffman’s NK biological fitness landscape for assessing the evolution of genotypes [2]. It inspired us to link the domain structure of the entire technology space to...
the genetic structure of the organism, and the adaptive evolution process of technology to the mutation of the genotypes. The NK model has also been successfully used to study the evolution of firms [3], innovation networks [4,5], and technological systems [6]. In our study, a specific technological domain was found to be analogous to a genotype. The high-dimensional embedding vector of a technological domain is analogous to a DNA sequence. The improvement rate of a technology domain is analogous to the fitness (or replication rate) of a genotype in an organism. Therefore, the technology fitness landscape is analogous to the genome fitness landscape.

Within the technology fitness landscape, we find that the fastest domains of today’s technological world gather closely in one region of the space. That is, the technology fitness landscape shows a single high hill, which is related to information, electronics, and electrical technologies. The hill is steep. The improvement rates (or fitness) drop rapidly from the global peak to the low plain. Most regions of the total technology embedding space are in a vast low-flat plain in terms of improvement rates. The technology fitness landscape map that we have created can inform innovators within specific domains about their embeddedness in the total space and guide them to mutate their technologies for innovation toward the direction of interest.

2 Method

To construct the “map of technologies,” previous studies leveraged the information in patent data to capture the relationships among different technological domains in the form of networks [7,8,9]. However, these existing maps only analyze the citation information of patents to capture the interdependencies or interactions among technological domains while ignoring the intrinsic features of technologies within them. In this research, we leverage deep learning techniques to create a compact, dense, and continuous vector space in high dimensions to represent all technologies based on multimodal information (both texts and citations of patents). Both the internal (semantic) and external (connective) information of technological domains are utilized for neural embedding model training, aiming to capture both the explicit and implicit features of technological domains. The trained technology embedding space is more holistic and informative than the previously created network maps of technological domains because such a space is high-dimensional; the representation learning process based on the neural network is flexible and robust [10].

Specifically, our proposed neural embedding training method builds on the GraphSAGE model [11], a framework for inductive representation learning on large graphs. GraphSAGE generates vector representations for nodes and is especially useful for graphs with rich node attribute information. The model can learn both intrinsic and connective information from a given graph with node features by performing a link prediction task, in which the training data are generated through an unsupervised sampler. It is noteworthy that this neural embedding method is an unsupervised learning algorithm that does not require any node class label data for training.

Algorithm 1 describes the GraphSAGE neural embedding method. The spirit behind this algorithm is that as the nodes learn features from their local neighbors, they would incrementally obtain an increasing amount of information from further reaches of the whole graph.

Algorithm 1: GraphSAGE neural embedding algorithm from [11]

\[
\begin{align*}
\text{Input:} & \quad \text{Graph } G(V, E); \text{ input features } \{x_v, \forall v \in V\}; \text{ depth } K; \text{ weight matrices } W^k, \forall k \in \{1, ..., K\}; \\
& \quad \text{non-linearity } \sigma; \text{ differentiable aggregator functions } AGGREGATE_k, \forall k \in \{1, ..., K\}; \\
& \quad \text{neighborhood function } \mathcal{N}: v \rightarrow 2^V \\
\text{Output:} & \quad \text{Vector representations } z_v \text{ for all } v \in V \\
\end{align*}
\]

\[
\begin{align*}
h_v^0 & \leftarrow x_v, \forall v \in V \\
\text{for } k = 1, ..., K \text{ do} \\
& \quad \text{for } v \in V \text{ do} \\
& \quad \quad h_v^k \leftarrow \text{AGGREGATE}_k(h_v^{k-1}, \forall u \in \mathcal{N}(v)); \\
& \quad \quad h_v^k \leftarrow \sigma(W^k \cdot \text{CONCAT}(h_v^{k-1}, h_u^k)) \\
& \quad \text{end} \\
& \quad h_v^K \leftarrow h_v^K/\|h_v^K\|_2, \forall v \in V \\
\text{end} \\
z_v & \leftarrow h_v^K, \forall v \in V \\
\end{align*}
\]

In the experimental setting, we use the domain-level citation network with normalized weights as the input graph \(G\). Formally, an entry in matrix \(W_{N \times N}\) from a citing domain \(n\) (row) to a cited domain \(n'\) (column) is:
$$w_{n \rightarrow n'} = \frac{C_{n \rightarrow n'}}{\sum_{k=1}^{N} C_{n \rightarrow k}}$$

where $C_{n \rightarrow n'}$ represents the citations from domain $n$ to domain $n'$. Thus, the normalized weight matrix represents interdomain interactions.

We use semantic vectors of domains as input node features derived from the Doc2Vec model \[12\]. The model is first trained on the text of the title and abstract of all patents in our dataset. Using the trained model, for any given patent document, we can obtain its semantic vector using its textual content. Then, the domain semantic vectors are calculated by averaging all the obtained semantic vectors of patents that belong to the corresponding domains. As for the parameters of the GraphSAGE model, we implemented the settings as suggested in the original GraphSAGE paper: non-linearity $\sigma = \text{sigmoid}$ function; aggregator function = Mean; Depth $K = 2$ with neighborhood sample sizes $= 32$. Stellargraph was used for the construction and training of our model. Thus, the derived multimodal domain embeddings can effectively encode both the internal semantic and the external connectivity characteristics of patent domains.

3 Data

We focus on the same patent dataset described in STM \[1\] to train the embedding space. The dataset contains 4,988,929 patents issued by USPTO from 1976 to 2015. Detailed patent information, including semantic information (titles and abstracts) and connective information (forward and backward citations) of patents, can be downloaded from the PatentsView website. There were 54,798,218 citations between patents in the dataset.

STM assigned these patents into a set of 1,757 technology domains via the extended classification overlap method \[13\], statistically estimating the yearly rates of performance improvement for all domains utilizing the centrality measure \[14\]. We applied our multimodal neural embedding method to a citation network of 1,757 domains with their semantic features. As a result, we obtained a 32-dimension vector for each of the 1,757 technological domains. We call the unified vector space the “technology embedding space.” The technology fitness landscape is built on the technology embedding space and technology improvement rates of all domains.

4 Results and Discussions

Fig. 1A shows a two-dimensional (2D) representation of the technology embedding space using the t-SNE method \[15\], where the node color intensity corresponds to the improvement rate of the represented domain. Fig. 1B presents the technology embedding space with domain improvement rates as a contour map. Both scatter and contour maps reveal a single cluster of the fastest domains, or, fastest domains tend to gather in a small cohesive region of the total technology embedding space. In other words, our neural embedding space mapping captures the fastest domains within a cluster, including the following domains as the top five: 719G06F (dynamic information exchange and support systems integrating multiple channels), 709G06F (network management specifically client-server applications), 709G06Q (network messaging system including advertisement), 709H04L (network address and access management), and 726H04L (securing enterprise networks by system architecture). We further developed a web-based interactive visualization for public users to explore the space available at: https://ddi.sutd.edu.sg/technology-fitness-landscape.

Fig. 2 presents the technology fitness landscape map, in which the heights of the domains correspond to their respective improvement rates. The landscape is not very rugged and has a small number of hills. The landscape is characterized by a conspicuous highest hill (the global peak), together with a vast low plain occupying most of the total technology space. In Fig. 2, we annotate several fastest domains in each hill according to their descriptions. We find that the highest hill is highly related to the information, electronics, and electrical technology domains. The slope from the peak of the highest hill to the low-flat plain is steep.

\[2\]https://github.com/stellargraph/
\[3\]https://www.patentsview.org/
\[4\]3D-meshgrid function of MATLAB is used to create (fit) the contour and landscape maps (Figs. 1B and 2).
\[5\]The codes of domains are UPC-IPC pairs, referring to \[1\].
\[6\]We also performed a Non-negative Matrix Factorization topic modeling algorithm \[16\] on the domains of the global peak to examine its theme. Results show that all of the identified topics are about information, electronics, and electrical technologies.
Figure 1: **Technology embedding space with improvement rates of all domains.** A) Each dot denotes a domain, and its color denotes the domain’s performance improvement rate. The darker dots represent faster domains. B) The lighter areas represent faster domains.

Figure 2: **Technology fitness landscape.** The location of each domain is aligned to the 2D embedding map (Fig. 1A), and the color represents the rate. The heights correspond to the improvement rates of different domains.
Fig. 3 presents the average improvement rates of each 10-quantiles group of domains by their Euclidean distance to the centroid of the N (=1, 5, or 10) domains surrounding the global peak. The results confirm that, in the near field surrounding the global peak, the improvement rates decline rapidly from the hill peak to the low plain of many slow domains.

**Figure 3:** Distance to the global peak (the center of the fastest M domains, M=1, 5, 10) and improvement rates of each group of domains.

To demonstrate different domains’ varying distances to the global peak, we mapped all technological domains into one of the 37 NBER subcategories [17]. These 37 subcategories belong to six categories: (1) computers and communications, (2) electrical and electronic, (3) mechanical, (4) drugs and medical, (5) chemical, and (6) other. Fig. 4 presents the constitution of each category regarding their distances to the global peak (the centroid of the 10 fastest domains). Each item in the matrix represents the number of domains belonging to the corresponding subcategories. The matrix was normalized by column (the sum of each column equals 1). Fig. 4 reveals a clear pattern of technological theme shifts from the global peak to the low plain.

**Figure 4:** The shift of technological themes of domains in NBER subcategories regarding their distances to the global peak.
In Kauffman’s NK fitness landscape, the space comprises genotypes similar or dissimilar to each other to different degrees. The height of an area corresponds to the fitness (or replication rate) of a particular genotype. Each genotype is composed of several nucleotides (DNA sequence), and each position in a DNA sequence can be occupied by four alternative bases: adenine (A), thymine (T), guanine (G), and cytosine (C). Thus, a DNA sequence can be viewed as a 4-dimensional vector. Point mutations in a DNA sequence, which substitute, insert, or delete a single nucleotide, will lead to movements in the landscape. These positive mutations (that increase the fitness of a genotype) can lead to a movement to higher areas of the biological fitness landscape.

By analogy, the technology domains in the total technology fitness landscape are similar to the genotypes. The domain features can be viewed as the nucleotides in the DNA sequence of the technological domain and are now characterized by our 32-dimensional vectors trained on multimodal data. Thus, the mutations of a technological genotype are equivalent to the changes in the values in different dimensions of the embedding vectors, seen as the movements in the technology embedding space. Meanwhile, the improvement rate of a domain can be viewed as its fitness (or replication rate) in the greater technological system that constitutes energy, material, and information processing technologies.

Therefore, innovations in a technological domain can be viewed as positive mutations (that increase the performance of a domain) of its technological genotype in the total technological system (of all energy-information-material technologies), which lead to the movements from lower to higher areas of the technological fitness landscape. Prior studies have pointed out that technological improvement or novelty arises from the recombination or synthesis of existing technologies [18, 6, 19], which, in our cases, can be viewed as such mutations of the technological genotype. For example, the latest innovations in the automotive domain by integrating AI technologies have changed the DNA of automobiles and increased the values of automobiles and their fitness in the total technological system. Using biological evolution terms, one can say that by integrating AI for driving, the automobile’s genotype has been mutated with increased fitness values. In this case, such AI-based automotive innovations and improvements can be viewed as a mutation moving from the low plain to the high mountain in the technological fitness landscape that we created.

For the innovators in presently slow-pacing domains, they may use our fitness landscape map as a guide to mutate their technologies (e.g., innovation) for targeted movements toward the high hill. For the innovators in the fast-pacing domains, they may also use our map to identify slow-paced domains as targets to empower by applying their fast-improving technologies over these. Both ways create value despite different starting points. The success of such mutations might be conditioned by the innovators’ starting positions and neighboring domains in the total space. Making long jumps require learning distant technologies and thus are difficult.

5 Conclusion

This research uses multimodal neural embedding training to derive a unified vector space for the 1,757 domains recently curated and analyzed by Singh et al. in Research Policy [1]. The training encodes both the internal semantic features of patent texts in domains and external interdependent information of patent citations across domains. We map STM’s estimated technological improvement rates for all domains in the technology embedding space to create a technology fitness landscape. The fitness landscape shows that the fastest domains gather closely in the total technology embedding space, while most regions of the space constitute a low plain in terms of improvement rates. A global peak was identified in the landscape with the themes of information, electronics, and electrical technologies. Such a technology fitness landscape provides us with a birds-eye view of the evolution prospects of individual technological domains. It may also potentially inform innovators within specific fields about their embeddedness in the total space and guide them to mutate their technologies toward the direction of interest for innovation.

This study has several limitations. First, the visualization and analysis in this study rely on STM’s domain classifications and estimated improvement rates. Researchers focusing on the different granularities of technological domains can leverage our method to retrain the model and generate a new landscape based on their dataset. Second, although our neural embedding method learns both intrinsic and connective features from domains and can be more holistic than unimodal spaces, other types of knowledge of domains exist, such as affiliation information and visual images. A desirable multimodal space should consider an even broader range of features. Third, the 3D landscape presented in this research was built on the 2D locations of domains obtained via the t-SNE algorithm. Notably, such a dimension-reduction process may involve information loss to some degree. The technological embedding space is high-dimensional, but it is difficult to visualize and understand. Except for the visualization, all the quantitative analysis works in this research are based on a high-dimensional landscape. We provide open access to all the data, codes, and results of this research. (The Github link will be provided here once the paper is accepted.)
Acknowledgements

The authors acknowledge the funding support for this work received from the SUTD-MIT International Design Center and SUTD Data-Driven Innovation Laboratory (DDI, https://ddi.sutd.edu.sg/), National Natural Science Foundation of China (52035007, 51975360), Special Program for Innovation Method of the Ministry of Science and Technology, China (2018IM020100), National Social Science Foundation of China (17ZDA020). Any ideas, results and conclusions contained in this work are those of the authors, and do not reflect the views of the sponsors.

References

[1] Anuraag Singh, Giorgio Triulzi, and Christopher L Magee. Technological improvement rate predictions for all technologies: Use of patent data and an extended domain description. Research Policy, 50(9):104294, 2021.
[2] Stuart A Kauffman. The origins of order: Self-organization and selection in evolution. Oxford University Press, USA, 1993.
[3] Daniel A Levinthal. Adaptation on rugged landscapes. Management science, 43(7):934–950, 1997.
[4] Koen Frenken. A complexity approach to innovation networks. The case of the aircraft industry (1909–1997). Research Policy, 29(2):257–272, 2000.
[5] Martin Ganco. NK model as a representation of innovative search. Research Policy, 46(10):1783–1800, 2017.
[6] Lee Fleming and Olav Sorensen. Technology as a complex adaptive system: evidence from patent data. Research policy, 30(7):1019–1039, 2001.
[7] Daron Acemoglu, Ufuk Akcigit, and William R. Kerr. Innovation network. Proceedings of the National Academy of Sciences of the United States of America, 113(41):11483–11488, 2016. ISSN 10916490.
[8] Lee Fleming and Olav Sorensen. Science as a map in technological search. Strategic management journal, 25(8-9):909–928, 2004.
[9] Jianxi Luo. Total Technology Space Map as a Digital Platform. Proceedings of the 52nd Hawaii International Conference on System Sciences, pages 6331–6338, 2019.
[10] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. Transactions of the association for computational linguistics, 3:211–225, 2015.
[11] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 1025–1035, 2017.
[12] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In International conference on machine learning, pages 1188–1196. PMLR, 2014.
[13] Christopher L Benson and Christopher L Magee. Technology structural implications from the extension of a patent search method. Scientometrics, 102(3):1965–1985, 2015.
[14] Giorgio Triulzi, Jeff Alstott, and Christopher L Magee. Estimating technology performance improvement rates by mining patent data. Technological Forecasting and Social Change, 2020.
[15] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. Journal of machine learning research, 9(11):2579–2605, 2008.
[16] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791, 1999.
[17] Bronwyn H Hall, Adam B Jaffe, and Manuel Trajtenberg. The NBER patent citation data file: Lessons, insights and methodological tools, 2001.
[18] Lee Fleming. Recombinant uncertainty in technological search. Management science, 47(1):117–132, 2001.
[19] Yuejun He and Jianxi Luo. The novelty ‘sweet spot’of invention. Design Science, 3, 2017.