Co-word Network Embedding: A Way to Detect Emerging Technology

Introduction

Emerging technologies are often perceived as having large undeveloped economic potential and being capable of promoting the society status quo. Although emerging technologies are a focus for policy makers, researchers and companies, the way to identify it have yet to achieve consensus[1]. With technological development, practical applications, or both largely unrealized, how to capture new terminology that may represent emerging technology, and what does the terminology mean? This paper aims to provide a useful way to detect emergent terms and its relevant concepts via text analysis to address the above question.

Many attempts have been tried in identifying emerging technologies, most of them are based on publication and patent analysis. The text and structural information of publication and patent make the analysis suitable to conduct methods like bibliometrics, social network analysis, Natural Language Processing (NLP), etc. Co-word network is usually utilized to analyze relations among different keywords through building a matrix or network, which can be visualized and reveal the hotspot and trend in certain technological area. Topic model can cluster documents and get multiple topics, which is widely used in detecting emerging clustering or items.

Radical novelty and *relatively fast growth* are put forth by Rotolo et al. as 2 out of 5 major attributes of emerging technology[2]. To measure attribute of technology's emergence, quantitative analysis is required, and how implicit information can contribute to indicator such as growth should be focused. While a single method has threshold exploiting implicit information and carrying on quantitative analysis, we develop a new method that combines co-word network and neural network model to learn from the co-word network and get dense representation of items for further indicator analysis. This method behaves well in emerging items mining and getting items highly correlated with the emerging ones.

Data and methods

The dataset of this study is retrieved from WoS in *Measuring Tech Emergence* 2018-2019 contest. WoS xml records include fields frequently used for analysis, such as title, abstract, author, organization, keyword, etc. In order to be terminology concentrated, our method focuses on titles, abstracts, author keywords and keywords plus. In data cleansing, we argue that multi-word terms are more significative and abandon single-word terms. Terms under a frequency of 5 in titles and abstracts are also abandoned since it creates volatility during indicator analysis. Furthermore, we merge terms into its primary form.

We then propose our new algorithm to identify the emerging technology. The process mainly includes the following steps:

1 Terms extraction. We segment Abstracts into sentences and combine them with Titles as a sequence of Documents (D). Keywords-Author (KA) and Keywords-Plus (KP) are used as original candidate terms. Topmine[3] is used to extract additional keywords (KE) from the Documents. Then extended candidate terms (eCT) are generated according to Formula (1) and filter them based on cleansing principle mentioned above.

$$eCT = (KA \cup KP \cup KE) \ \ occur \ \ in \ \ D \quad (1)$$

2 Building co-word network. A co-word network can reveal the technical relevance between terms. The extended candidate terms are used to build the co-word network which can be described as graph[4]. The graph consists of nodes (N), edges (E) and weights (W), where N means terms, E means co-word relation between two terms if they occur in one document, W means times that two terms occur together. The following Figure 1 shows steps to build co-word graph.

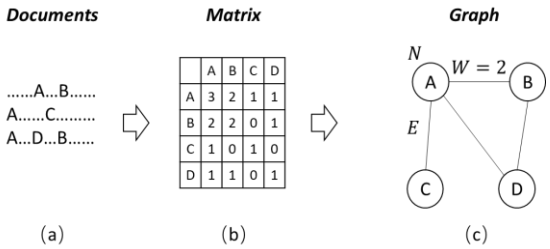

Figure 1. Steps to build co-word graph:
(a) Titles and sentences from abstracts
(b) Matrix contains co-word frequency (weights) between terms (nodes).
(c) Co-word graph consists of nodes, edges and weights.

3 Vector representation. To carry on quantitative analysis, one of the prevalent ways is vectorization. We conduct representation training by using Node2vec[5], so as to represent term nodes and their implicit relation in a vector space. This includes a step of generating sequences of term nodes, which uses random walk algorithm in the network. The walk path is decided by the weights between nodes and two parameters (return parameter p and in-out parameter q). These sequences affect much the representation outcome. Therefore, we experiment groups of parameters in random walk to balance the bias between community feature and structural feature of term nodes. Finally we set p to 1 and q to 0.5. The outcome of this step is dense vector representation of candidate emergent terms.

4 Emergence indicator design. To measure technological emergence, we use Euclidean distance ($EDis$) to calculate Notability ($Nota$) of each terminology by their representation ($P_i = (x_1, x_2, \dots, x_m)$). The Notability is defined as formula (2), which is relevant to the local density of term vectors in the vector space. Emergent terms will then be selected and ranked according to the growth rate of their Notability.

$$Nota(P_i) = \frac{1}{n}\sum_{j=1, j\neq i}^{n} \frac{1}{EDis(P_i, P_j)} \qquad (2)$$

5 Relevant concept. With term nodes represented in a vector space, a cosine distance (*CosD*) matrix among nodes can be computed as formula (3). According to the matrix, smaller cosine distance stands for higher similarity between terms. We then rank the candidate terms according to their cosine distance to emergent terms and pick up to 10 terms in the top. In light of central emergent terms are already picked out, candidate terms close to the emergent terms can be considered as sub-technology or relevant concepts.

$$CosD(P_i, P_j) = \frac{P_i \cdot P_j}{|P_i| \cdot |P_j|} \qquad (3)$$

Results

As a result, we obtain 10 emergent terms from the given dataset, which may represent emerging technologies, with their relevant technologies and concepts as detailed description respectively. In this case, we provide a useful way that can help policy-makers and entrepreneurs figure out which technologies are more important and their relevant items (such as other technologies, equipment, methods and so on), so that further research or decision could proceed. The result is submitted and will have further discussion in the future.

References

[1] Alexander J , Chase J , Newman N , et al. Emergence as a conceptual framework for understanding scientific and technological progress[C]// Technology Management for Emerging Technologies. IEEE, 2012.

[2] What is an emerging technology?[J]. Research Policy, 2015, 44(10):1827-1843.

[3] El-Kishky, A., Song, Y., Wang, C., R. Voss, C., et al. Scalable Topical Phrase Mining from Text Corpora. Proceedings of the VLDB Endowment. 2014, 8: {}.

[4] Jin, C., Bai, Q., Text Clustering Algorithm Based on the Graph Structures of Semantic Word Co-occurrence. 2016, p. 497-502.

[5] Grover, A., Leskovec, J. node2vec: Scalable Feature Learning for Networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2016. 855-864.