Science Graph for characterizing the recent scientific landscape

Takahiro Kawamura[1], Katsutaro Watanabe[1],
Naoya Matsumoto[1], Shusaku Egami[1] and Mari Jibu[1]

Japan Science and Technology Agency

Abstract. Maps of science representing the structure of science can help us understand science and technology development. However, navigating the recent scientific landscape is still challenging, since application of inter-citation and co-citation analysis for ongoing projects and recently published papers has difficulty. Therefore, in order to characterize what is being attempted in the current scientific landscape, this paper proposes a content-based method of locating research projects in a multi-dimensional space using word/paragraph embedding techniques. The proposed method successfully formed a science graph with 78% accuracy from 25,607 project descriptions of the 7th Framework Programme (FP7) from 2006 to 2016.

1 Introduction

Research in scientometrics has developed techniques for analyzing research activities and for measuring their relationships, and then constructed maps of science [1], one of the major topics in scientometrics. Maps of science have been useful tools for understanding the structure of science, their spread, and interconnection of disciplines. However, conventional approaches to understanding research activities focus on what authors tell us about past accomplishments through inter-citation and co-citation analysis of published research papers. Therefore, this paper focuses on what researchers currently want to work on thier research projects. Project descriptions, however, do not have references and can not be analyzed using citation analysis; thus, we propose to analyze them using a content-based method using natural language processing (NLP) techniques. Then, we created a science graph, which is a knowledge graph representing the recent scientific trends, where nodes represent research projects that are linked by certain distances of the content similarity and their semantics.

2 Related Work

Some studies have examined automatic topic classification based on content using lexical approaches such as probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA). One uses LDA to find the five most probable words for a topic, and each document is viewed as a mixture of topics. Thus, this approach can classify documents across different agencies and publishers. However, the relationship between any project and article, such as that involving their distance or semantics, cannot be computed directly.

By contrast, Le and Mikolov [2] proposed a paragraph vector that learns fixed-length feature representations using a two-layered neural network from plain texts, such as sentences, paragraphs, and documents. A paragraph vector is considered another word in a paragraph and is shared across all contexts generated from the same paragraph but not across paragraphs. The paragraph vectors are computed by fixing the word vectors and training the new paragraph vector until convergence. By considering word order, paragraph vectors can also address the weaknesses of bag-of-words models in LDA and pLSA.

3 Measurement of Project Relationships

In this study, we analyzed project descriptions from FP7. Precisely, our experimental data set consisted of the titles and descriptions of 25,607 FP7 projects from 2006 to 2016, including 305,819 sentences in total. All words in the sentences were tokenized and lemmatized before creating the vector space.

Firstly, we constructed paragraph vectors for 25,607 FP7 projects using the current paragraph embedding technique. The hyperparameters were set empirically as follows: 500 dimensions were established for 66,830 words that appeared more than five times; the window size c was 10, and the learning rate and minimum learning rate were 0.025 and 0.0001, respectively, with an adaptive gradient algorithm. The learning model is a distributed memory model with hierarchical softmax. As a result, we found that projects are scattered and *not clustered* by any subject or discipline in the vector space. Most projects are slightly connected to a low number of projects. Thus, it is difficult to grasp trends and compare an ordinary classification system such as SIC codes. Closely observing the vector space reveals some of the reasons for this *unclustered* problem: each word with nearly the same meaning has slightly different word vectors, and shared but unimportant words are considered the commonality of paragraphs. Therefore, for addressing this problem, we introduce an entropy-based method for clustering word vectors before constructing paragraph vectors.

To unify word vectors of almost the same meaning, excluding trivial common words, we generated cluster vectors of word vectors based on the entropy of each concept in a thesaurus. We calculated the information entropy of each concept in the FP7 projects. Next, after creating clusters according to the degree of entropy, we unified all word vectors in the same cluster to a cluster vector and constructed paragraph vectors based on the cluster vectors. The overall flow is shown in Fig. 1.

$$H(C) = -\sum_{i=0}^{n}\Big(\sum_{j=0}^{m} p(S_{ij}|C) \cdot \log_2 \sum_{j=0}^{m} p(S_{ij}|C))\Big) \tag{1}$$

Shannon's entropy in information theory is an estimate of event informativeness. Given that a thesaurus consists of terms T_i, we calculated the entropy of a concept C by considering the appearance frequencies of a hypernym T_0 and its hyponyms $T_1 ... T_n$ as an event probability. The frequencies of synonyms $S_{i0} ... S_{im}$ of term T_i were summarized to a corresponding concept (synonyms S_{ij} include descriptors of terms T_i themselves). In Eq. (1), $p(S_{ij}|C)$ is the probability of a synonym S_{ij} given a concept C and terms T_i. For each concept in the thesaurus,

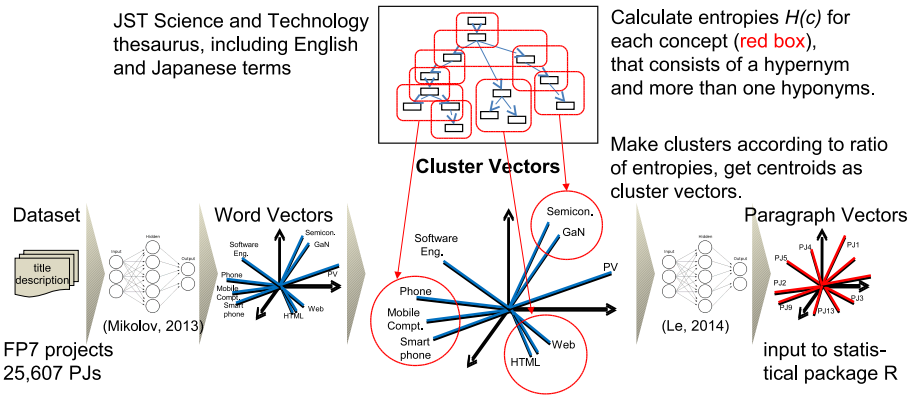

Fig. 1. Construction of paragraph vectors based on cluster vectors.

we calculated the entropy $H(C)$ in the FP7 data set. As the probabilities of events become equal, $H(C)$ increases. If only particular events occur, $H(C)$ is reduced because of low informativeness. Thus, the proposed entropy of a concept increases when a hypernym and hyponyms that construct a concept separately appear with a certain frequency in the data set. Therefore, the degree of entropy indicates the semantic diversity of a concept. Then, assuming that the degree of entropy and the spatial size of a concept in a word vector space are proportional to a certain extent, we split the word vector space into clusters. In fact, our preliminary experiment indicated that the entropy of a concept has high correlation $R = 0.602$ with the maximum Euclidean distance of hyponyms in the concept in a vector space, at least while the entropy is rather high. The vector space is subdivided into clusters proportionally to the ratio of the highest two concept entropies. Each cluster is subdivided until the entropy becomes lower than 0.25 (the top 1.5% of entropies) or the number of elements in a cluster is lower than 10. These parameters were also determined empirically through the experiments. After generating 1,260 cluster clusters from 66,830 word vectors, we considered the centroid of all vectors in a cluster as a cluster vector. Then, we obtained paragraph vectors by calculating the maximum likelihood of L in Eq. (2), which is an extension of the paragraph embedding defined in [2]. $Cl(w)$ means a cluster vector to which a word w belongs, and d_i is a vector for a paragraph i that includes w_t. T is the number of words with a certain usage frequency in the corpus. Using high-entropy concepts in scientific and technological contexts as common points with each paragraph vector (excluding trivial words), paragraph vectors can comprise meaningful groups in the vector space.

$$L = \sum_{t=1}^{T} \log p(Cl(w_t)|Cl(w_{t-c}), ..., Cl(w_{t+c}), d_i) \quad (2)$$

4 Experiments and Evaluation

The science graph for FP7 is publicly accessible at `http://togodb.jst.go.jp/sample-apps/map_FP7/` (click "Drawing Map" button. see the CORDIS website for subject codes). The distributed recursive graph layout (DrL) algorithm,

which produces edge-weighted force directed graphs, was used to visualize the relationships between projects. We computed 328 million cosine similarities for all pairs of the 25,607 projects; however, we kept only those that were above a given threshold (0.35 in this case) as edges due to visualization limitation.

In terms of the *unclustered* problem, we confirmed that the proposed method successfully formed several clusters compared with the baseline method, in comparison with the relationships between the cosine similarities and the number of edges, and the relationship between degree centrality and the number of nodes.

However, since there is no gold standard for evaluating the distance among research projects, we evaluated the accuracy of the similarities based on a sampling method. We randomly extracted 100 pairs of projects with a cosine similarity of > 0.5, to make the distribution similar to the entire distribution. Each pair has two project titles and descriptions, and a cosine value that is divided into three levels: weak ($0.5 \leq \cos. < 0.67$), middle ($0.67 \leq \cos. < 0.84$), and strong ($0.84 \leq \cos.$). Then, three members of our organization, a funding agency in Japan, evaluated the similarity of each pair. The members were provided the prior explanations for the intended use of the graph, some examples of evaluation and the same evaluation data. As a result, we confirmed that 78% of the project similarities (i.e., the distances in the graph) matched majority votes of the members' opinions. Examples misjudged include, e.g., two projects using lots of homonyms with high cos values and two projects which accidentally have some similar sentences with cos values just over the threshold, and so forth. By contrast, the accuracy of the distances in the baseline was 21%. The evaluation results were determined to be in "fair" agreement (Fleiss' Kappa $\kappa = 0.29$).

5 Conclusion and Future Work

Since funding projects do not have references and also recently published articles do not have enough citations yet, we assessed the relationships using a content-based method, instead of citation analysis. At the back end of the graph, bibliographic information is stored as our Linked Data database [3], and they are mainly connected by *similarTo* with similarity values and by *hasConcept* with common concept classes.

As the next step, we will extract new insights from the science graph of research projects, especially in comparison with previous maps of science based on citation analysis of published papers.

References

1. Boyack, K.W., Klavans, R., and Borner, K.: "Mapping the backbone of science," Scientometrics, 64(3), pp.351–74, 2005.
2. Le, Q. and Mikolov, T.: "Distributed Representations of Sentences and Documents," Proc. of ICML 2014, 32, 2014.
3. Kimura, T., Kawamura, T., Watanabe, et al.: "J-GLOBAL knowledge: Japan's Largest Linked Open Data for Science and Technology," Proc. of ISWC 2015, Poster & Demo Track, 2015.