

Doc2Vec を用いた学会発表概要集の検索手法の検討

A study on classification of meeting abstracts using Doc2Vec

名大院工¹, 理研 AIP², 名大未来研³, 産総研 GaN-OIL⁴

○石川 晃平¹, 沓掛 健太郎², 原田 俊太^{1,3}, 田川 美穂^{1,3}, 宇治原 徹^{1,3,4}

Grad. School of Eng., Nagoya Univ.¹, AIP, RIKEN², IMASS, Nagoya Univ.³, GaN-OIL AIST⁴

°Kohei Ishikawa¹, Kentaro Kutsukake², Shunta Harada^{1,3},

Miho Tagawa^{1,3}, Toru Ujihara^{1,3,4}

E-mail: Ishikawa@unno.material.nagoya-u.ac.jp

【はじめに】応用物理学会をはじめとする学術講演会では、非常に多くの発表が短期間に行われる。前もって発表概要集が配布されるものの、参加者がすべてに目を通すことは不可能である。必然的に、参加者はタイトルを頼りに発表を探す必要に迫られるが、タイトルが内容を適切に表していない場合も多い。結果的に、参加者の多くは本当に求めていた発表を聞き逃してしまう。我々はこの問題を解決するために、機械学習を用いた学会発表概要集の自動分類を検討した。

【実験方法】米国電気化学学会が主催した Meeting of Electrochemical Society の発表概要集(総数: 2371 件)をサンプルとした。本研究では、Doc2Vec を用いることで概要の全文章を 200 次元のベクトルにそれぞれ変換した。発表間の類似度比較にはベクトル間のコサイン類似度を用いた。

【実験結果】Doc2Vec により、発表概要の文章はその意味を含む多次元のベクトルとして表現される。全発表のベクトルを t-SNE 法を用いて 2 次元に表現したプロットを Fig. 1 に示す。プロットの色は各発表の所属セッションを表しており、プロット間の距離は発表間の類似度を表している。ある発表(番号:2158)を例として、全発表を対象にコサイン類似度を用いた検索を行った例を Fig. 2 に示す。発表 2158 は金属 Li 負極に関する研究であり、関連した発表が抽出されていることがわかる。また、電池分野のみならず、Al の電析に関する発表(番号:2352)も抽出されている。このように、本手法は各概要の文書のベクトル化を基礎としているため、セッションを超え、研究手法・内容から本質的な類似度を検索することができる。本手法を用いることで、学会参加者は自身の研究分野からだけでなく、他分野からも類似した発表を発見することが可能になる。

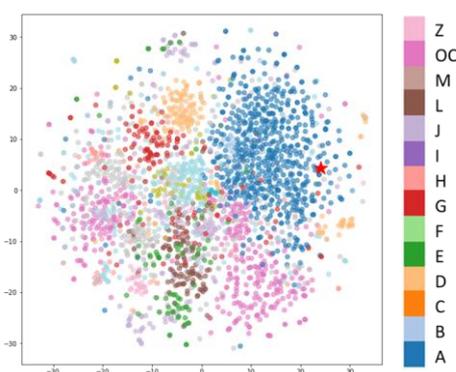


Fig. 1 2D plots of document vectors of ECS meeting abstracts by Doc2Vec.



Fig. 2 Searching ECS meeting abstracts based on cosine similarity of the document's vector.