文献情報の多様な要素を考慮したベクトル表現獲得

Acquiring vector representations considering various elements in bibliographic information

米田 拓真 Takuma Yoneda
三輪 誠 Makoto Miwa
佐々木 裕 Yutaka Sasaki
豊田工業大学
Toyota Technological Institute
{sd14084, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 背景と目的

大量の書誌情報から，興味のある文献や著者を見つけることができれば，効率よく知見を広げることができる．これを実現する方法として，書誌情報に含まれる各要素のベクトル表現を獲得することにより，簡単な演算で類似した文献等の検索を可能にすることが考えられる．

従来の手法として，文献をノードとし，それらを引用関係で接続したグラフ構造を作った上で，グラフ構造に対するベクトル表現獲得手法を適用して各文献の表現獲得を行う手法がある，これらの手法には書誌情報に含まれる多様な情報を考慮できないという問題があった．また，あるトピックから文章が作成される過程をモデル化し，著者や文献に含まれる単語間の関係を表現する author-topic model [4] などのトピックモデルも広く研究されてきたが，多様な情報を考慮するモデルを作成することが難しいという問題があった．

書誌情報の表現を得る際，引用関係のみならず，著者や出版年などの多様な情報を考慮できるとより質の高い表現を獲得できると考えられる．

本研究では，各文献に対応するノード同士が引用関係に従って接続されたグラフ構造を用いるのではなく，図 1 のように文献に含まれる著者や出版年といった様々な要素がノードとして文献のノードに接続される，新しいグラフ構造を用いて表現獲得を行うことで今まで扱われてこなかった様々な情報を考慮して，書誌情報における各要素のベクトル表現を獲得する手法を提案する．

実験では，自然言語処理に関する論文の書誌情報を用いて提案手法のモデルを学習し，ある著者に関連する著者及び関連語を予測させ，その結果を author-topic model によるものと比較した．

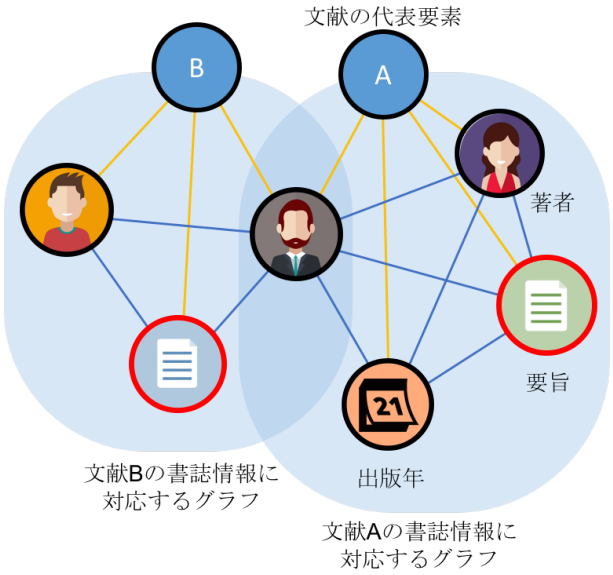

図 1: 提案手法における書誌情報のグラフ構造．この例では，円周が黒いものが非文章カテゴリに，赤いものが文章カテゴリに属する要素である．

2 関連研究

複数のノードがそれらの関係性に応じてリンクで接続された構造であるグラフ構造に対して，グラフに含まれる各ノードやリンクに対して表現を獲得する際には，元のグラフ構造の情報をできるだけ維持することが重要になる．LINE (Large-scale Information Network Embedding) [5] 及び TransE [2] は，グラフ構造における各要素のベクトル表現を得る手法である．グラフ構造において，ある二つのノード *head*, *tail* が関係 *relation* で接続されているとき，これらに対応するベクトル表現をそれぞれ $\boldsymbol{e_h}$, $\boldsymbol{e_t}$, $\boldsymbol{e_r}$ と表す．TransE では，グラフ構造に対し，式 (1) を最小化することで

ベクトルを獲得する．

$$\sum_{\substack{(h,r,t)\in S \\ (h',r,t')\in S'}} [\gamma + d(\boldsymbol{e_h} + \boldsymbol{e_r}, \boldsymbol{e_t}) - d(\boldsymbol{e_{h'}} + \boldsymbol{e_r}, \boldsymbol{e_{t'}})]_+ \tag{1}$$

$$d(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|^2 \tag{2}$$

ここで S は対象とするグラフ構造内に存在する $(h,\ r,\ t)$ の組み合わせの集合，S' は S に含まれる集合の各要素に対して，h または t のどちらか一方を無作為に入れ替えたものの集合を表している．

一方，LINE では，各ノードに対して目標ベクトル $\boldsymbol{v}$ および文脈ベクトル $\boldsymbol{w}$ の二つのベクトルを学習する．グラフ内のリンクで接続されたノードのペア h, t について，式 (3) で表されるように，ノード h からノード t を予測する確率モデルを考え，式 (4) を小さくするように各ベクトルを学習する．

$$p(t|h) = \frac{\exp(\boldsymbol{\omega_t} \cdot \boldsymbol{\upsilon_h})}{\sum_{t'\in S''} \exp(\boldsymbol{\omega_{t'}} \cdot \boldsymbol{\upsilon_h})} \tag{3}$$

$$-\sum_{(h,t)\in D} \log p(t|h) \tag{4}$$

本手法では，文献の書誌情報を表現する新しいグラフ構造に対してこれらの手法を用いて各要素のベクトル表現を獲得する．

3 提案手法

本研究では，書誌情報に含まれる多様な情報を考慮した表現獲得を行うため，書誌情報を様々な要素を含む新しいグラフ構造に変換してベクトル表現を獲得する手法を提案する．

この章では，まず提案手法により文献情報から作られるグラフ構造について説明した後，各要素の扱いとベクトル表現の獲得手法について述べる．

3.1 タスクの定義

本手法では，データセットは次のような構造を持つことを前提とする．データセットは文献の書誌情報からなり，各文献は複数のカテゴリを持つ．カテゴリは文章カテゴリ (題目や要旨など) と非文章カテゴリ (著者や引用文献など) の二つのグループに分けられる．各カテゴリは必ず一つ以上の要素を持ち，一般に文章カテゴリは多くの要素 (単語など) を持つのに対し非文章カテゴリの要素数は数個である．

3.2 文献情報に基づくグラフの作成

以下に提案手法で用いるグラフ構造の作成方法を述べる．提案手法により作成されるグラフ構造の例を図 1 に示す．まず，ある文献に着目し，その代表ノードを作成して文献に含まれる各要素と接続する．(図中黄色で表されたリンク) 次に，文献内の全要素を互いに接続する．(図中青色で表されたリンク) このとき，他の文献にも現れる要素を含んでいる場合にはその要素に対応するノードは共有されることに注意する．図 1 の場合，中央の著者は文献 A，B どちらの著者でもあるから，対応するノードが二つのグラフで共有されている．これを対象とするデータ内の全文献に対して行って完成するグラフ全体が本手法で扱うグラフ構造となる．なお，このグラフ構造におけるリンクに種類はない．図中には二色のリンクが描かれているが，これは単に図を明瞭にするためのものである．

3.3 各要素のベクトル表現の獲得

本手法では，文献の書誌情報に対応するグラフ構造を作成した後，LINE 及び TransE に基づく手法により各ノードのベクトル表現を獲得する．ノードのベクトル表現の概要を図 2 に示す．

3.3.1 LINE に基づく手法

LINE に基づく手法では，非文章カテゴリに属する各要素に対して目標ベクトル $\boldsymbol{\upsilon}$ 及び文脈ベクトル $\boldsymbol{\omega}$ の二種類のベクトルを対応付ける．これに対して，計算量を削減するために，文章カテゴリの要素に含まれる文章の各語には目標ベクトルのみを対応づけ，それらから要素の代表ベクトルを算出する．この代表ベクトルは目標ベクトルとみなされる．

式 (3) から分かるように，*tail* になりうる要素は文脈ベクトル $\boldsymbol{\omega}$ を持たねばならない．本手法では，文章カテゴリの要素は文脈ベクトルを持たないため，*tail* に割り当てることは出来ない．よって，LINE に基づく手法では，グラフ構造内で接続されている一対のノードに対して h, t を割り当てる際に，t が文章カテゴリの要素であるものを除外する．全てのペアを作成後，式 (4) が小さくなるような各ベクトルを獲得する．なお，この手法は米田ら [6] によってまとめられたものである．

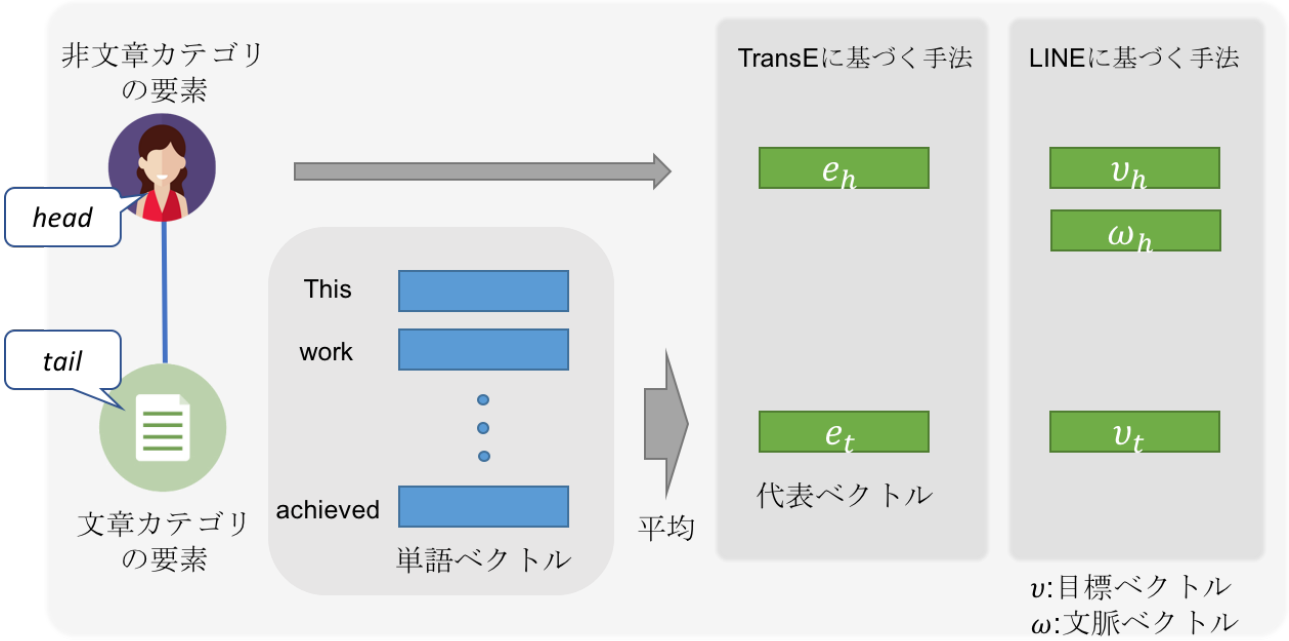

図 2: ノードのベクトル表現

3.3.2 TransE に基づく手法

TransE に基づく手法では，提案手法により作成したグラフ構造において，接続されている一対のノードに対しそれぞれ h, t を割り当て，それらの要素のカテゴリのペアを r とする．なお，ここで LINE に基づく手法に対応して，t に文章カテゴリの要素が対応するものを除外する．非文章カテゴリに属する各要素は各々，対応する一種類のベクトル $\boldsymbol{e}$ を持つ．これに対して，文章カテゴリの要素はそれに含まれる文章の各語がそれぞれ一種類のベクトルを持ち，それらから代表ベクトルを計算したものを文章カテゴリの要素に対応するベクトルとする．グラフ構造に含まれる全てのペアを作成後，式 (1) が小さくなるように各ベクトルを更新する．

表 1: ACL Anthology Reference Corpus の詳細

		要素数		しきい
カテゴリ	種別	元データ	前処理後	頻度
要旨	文章	59,276	10,994	20
著者	非文章	17,260	2,609	5
引用	非文章	10,871	10,871	1
出版年	非文章	16	16	1
文献の代表要素	非文章	19,475	19,475	1

4 実験

4.1 データセット

実験は，ACL Anthology Reference Corpus version 20160301 [1] を用いて行った．このデータセットは自然言語処理に関する 19,475 論文からなり，各文献には書誌情報として著者・題目・要旨・出版年・引用の情報が含まれている．前処理として，各文献の題目と要旨を結合して一つの文章とし，全単語の小文字化および数字のみからなる語の削除を行った． 続けて word2phrase tool[1] を用いて 3 単語までの語を繋げ，フレーズ化した．最後に，出現頻度の低い要素によって獲得する表現が悪影響を受けないよう，カテゴリ毎に頻度の閾値を設け，頻度が低い要素を削除した．各カテゴリに設定した閾値と前処理前後における要素数を表 1 に示す．

4.2 評価

提案手法を用いて書誌情報からグラフ構造を作成し，それに対して LINE に基づく手法及び TransE に基づく手法を用いて各要素のベクトル表現を獲得した．なお，どちらの手法においても，計算量を削減するため文章カテゴリの要素の代表ベクトルを要素に含まれる

[1] https://github.com/tmikolov/word2vec

	Author-topic model		LINE に基づく手法		TransE に基づく手法	
著者	関連語	関連著者	関連語	関連著者	関連語	関連著者
Philipp Koehn	machine translation	Hieu Hoang	alignment	Chris Dyer	muc-4	Philip Williams
	hmeant	Alexandra Birch	translation	Qun Liu	heading	Hieu Hoang
	human translators	Eva Hasler	align	Hermann Ney	rithms	Eva Hasler
Ryan McDonald	dependency parsing	Keith Hall	parse	Michael Collins	heading	Hiroshi Ichikawa
	extrinsic	Slav Petrov	sentense	Joakim Nivre	user friendly	Yuji Matsumoto
	hearing	David Talbot	parser	Jens Nilson	satellite	Slav Petrov

表 2: LINE に基づく手法とトピックモデルにより出力された関連要素の比較

文章の各語に対応するベクトルの平均とした．これらを用いて文献の書誌情報に含まれる各要素のベクトルを獲得することで，要素のカテゴリにかかわらず，様々な要素間の類似度を計算することができる．

4.3 結果と考察

提案手法により，文献の書誌情報に含まれる多様な要素間で可能になった類似度計算の一例として，何人かの著者に関する関連語及び関連著者を出力し，author-topic model によるものと比較した結果を表 2 に示す．各著者に関連すると考えられる単語が関連語として予測されており，author-topic model によるものと比較しても概ね妥当なベクトル表現が獲得できていると考えられる．

5 まとめと今後の課題

本論文では，従来手法では難しかった，書誌情報に含まれる多様な情報を活用した各要素のベクトル表現獲得を実現するため，各文献の書誌情報を多種類の要素を含む新たなグラフ構造に変換し，LINE 及び TransE に基づく表現獲得手法により，各要素のベクトルを獲得する手法を提案した．実験では，提案手法による異なるカテゴリ間の類似度計算の結果を author-topic model によるものと比較し，各要素を表現する妥当なベクトル表現が得られていることを示した．今後は，文章カテゴリの代表ベクトルに，平均ベクトルではなく単語の順序を考慮したベクトルを採用したり，文章要素に含まれる各単語に対する注意機構 [3] を導入し，各単語の重み付き平均を取ることで表現を獲得することなどを考えている．

参考文献

[1] Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan R. Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*, 2008.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 2787–2795. Curran Associates, Inc., 2013.

[3] Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W Black, Isabel Trancoso, and Chu-Cheng Lin. Not all contexts are created equal: Better word representations with variable attention. In *EMNLP*, pp. 1367–1372, 2015.

[4] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *UAI*, pp. 487–494, 2004.

[5] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: large-scale information network embedding. In *WWW*, pp. 1067–1077, 2015.

[6] Takuma Yoneda, Koki Mori, Makoto Miwa, and Yutaka Sasaki. Bib2vec: Embedding-based search system for bibliographic information. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 112–115, Valencia, Spain, April 2017. Association for Computational Linguistics.