文書間の類似度近似と数理最適化を用いた検索結果多様化

大滝啓介　徳久良子　岡田明久　吉田広顕
株式会社豊田中央研究所
{otaki,tokuhisa,okada,h-yoshida}@mosk.tytlabs.co.jp

概要

蓄積されている技術文書から、自分の関心がある文書を適切に探し出す情報検索は重要である。本稿では自然言語処理を用いて「検索クエリと文書」や「文書と文書」の類似度をベクトル埋め込みを通じて計算した上で、文書間の類似度を考慮しつつ、多様な文書を出力する検索結果多様化を扱う。既存のモデルでは文書の親子関係のみを考慮して多様化を行っていたため、表現可能な文書間の関係性に制限があった。本稿ではこの制限に対処するために一般化した数理モデルを提案し、小規模データでの適用例を通じて手法や今後の課題について議論する。

1　はじめに

背景　日々の研究活動では、大量に蓄積される技術文書から自分の関心がある文書を探し出す必要がある。このような情報検索は研究開発にとって欠かせない行為であり、技術文書検索を支えるため様々な技術が検討されてきた [1, 2, 3]。

本稿では関連スコアベースの検索技術に注目する。クエリ q と検索対象の文書 $\{d_1, \ldots, d_n\}$ に対して、クエリ q と文書 d_j との関連性を定量化することで、評価値に従って文書をソートし、検索結果を出力することができる。具体的には、例えば検索クエリ q と文書 d_j、文書 d_i と文書 d_j をベクトルの埋め込みで表現し、これらのベクトル間の類似度を求めることで、関連性を定量化した検索システムを構築することができる [4]。近年では自然言語処理に関する論文を検索するサービスも提供されている[1)]。

課題と本稿の貢献　類似度を関連スコアとして用いて上位 K 個の文書を検索結果として出力すると、出力された文書が相互に類似する傾向がある。一方で、よりユーザ経験豊かな検索システムを実装するためには、検索結果に多様な文書が含まれることが望ましい。このような問題設定は**検索結果多様化** (*search result diversification*) の文脈で研究されてきた [5]。選択した検索結果の多様性は、文書間の類似度などを用いて評価される。これまでの手法 [6] では文書の親子関係のみを考慮することしか出来ない (詳細は 2 章で述べる)。そのため親・子・孫の関係や、一般に木構造で表現されるような文書間の階層関係を表現することができなかった。

我々は数理最適化を用いて、既存研究を拡張した一般の森構造関係を考慮できる数理モデルを提案し、上記の課題を解決する。提案手法によって、文書間の芋づる式の検索や、木構造で表現する複雑な階層関係を想定した文書群の検索をサポートし、より豊かな検索経験が得られると期待される。本稿では特に、新しく提案した数理モデルの動作検証に注力し、小規模のデータに対して適用した結果を通じて手法を議論した結果を報告する。

2　準備

本稿で提案する数理モデルの元になった既存手法と問題設定について説明する。

2.1　ベクトル埋め込みを用いる検索

システムは n 個の文書 $D = \{d_1, d_2, \ldots, d_n\}$ を管理し、ユーザはクエリ q を検索システムに入力する。これらの文書やクエリから $\mathbb{R}^h$ に属するベクトルへの埋め込みが計算可能として、それぞれを $\mathbf{v}_q, \mathbf{v}_{d_i} \in \mathbb{R}^h$ で表す。クエリと文書間の類似度を計算する関数を $r : \mathbb{R}^h \times \mathbb{R}^h \to \mathbb{R}$ と設定し、$r(\mathbf{v}_q, \mathbf{v}_{d_i})$ で関連スコアを表す。また文書間の類似度を計算する関数を $s : \mathbb{R}^h \times \mathbb{R}^h \to \mathbb{R}$ と設定し、文書間の類似度を $s(\mathbf{v}_{d_i}, \mathbf{v}_{d_j})$ で表す。関連スコアと文書間の類似度には、いずれもコサイン類似度を用いる。

選択した文書の関連スコアを $R(S) := \sum_{d_j \in S} r(\mathbf{v}_q, \mathbf{v}_{d_j})$ で定義する。多様化を考慮しない検索システムは、式 (1) の最適化問題を通じて検索を行う。

1)　ACL2Vec：http://clml.ism.ac.jp/ACL2Vec/

$$\max_{S\subseteq D,|S|=K} R(S). \tag{1}$$

2.2 検索結果多様化

式 (1) を拡張し、検索結果多様化を定式化する。検索結果多様化では、選択した文書集合 S 上の多様性スコアを定義し、関連スコアと多様性スコアの両方を持つ目的関数を用いる。文書集合 S に対する多様性スコアが $\Omega(S)$ の場合、多様性を考慮する検索システムが扱う問題は式 (2) のような多目的最適化問題として定式化される [6, 7]。

$$\max_{S\subseteq D,|S|=K} R(S)+\Omega(S). \tag{2}$$

本稿では多様化手法のうち、文書間の親子関係に着目して S を一度に構築する手法である ILP4ID [6] に着目し、拡張した数理モデルを提案する[2)]。

2.2.1 ILP4ID (ILP for Implicit Diversification)

ILP4ID では文書間に親子関係 $\succ$ を考慮し、親子関係から $\Omega_{\mathrm{ILP4ID}}(S) := \sum_{(d_i,d_j)\in S\times S \text{ if } d_i\succ d_j} s(d_i,d_j)$ を評価した上で、第二項 $\Omega_{\mathrm{ILP4ID}}(\cdot)$ を用いた最適化問題を通じて検索結果を多様化する [6]：

$$\max_{S\subseteq D,|S|=K} \lambda(n-K)R(S)+(1-\lambda)K\Omega_{\mathrm{ILP4ID}}(S). \tag{3}$$

式 (3) では 1 項目の関連スコアと 2 項目の多様性スコアの両項が、全文書数 n と選択する文書数 K とパラメータ $\lambda\in[0,1]$ で重み付けされている。二項目の $\Omega_{\mathrm{ILP4ID}}(\cdot)$ は親子関係に選択した文書間の類似度が相互に高い場合に高くなる項である。そのため、親として選んだ文書集合 S は、式 (3) を最適化した結果として多様化されていることが期待される。なお $\lambda=1$ のとき通常の検索と一致する。

最適化問題としての実装　著者らは整数計画法を用いて目的関数の式 (3) を実装した [6]。二値変数 $x_{ij}\in\{0,1\}, i\in[n]:=\{1,\dots,n\}, j\in[n]$ を用いて、$x_{ii}=1$ によって $d_i\in S$ を、$x_{ij}=1$ によって親子関係 $d_i\succ d_j$ を表す。選択していない文書については、選択した文書に対して親子関係を持ち、制約に $x_{ii}-x_{ji}\geq 0$ を満たす。これは文書 $d_i\in S$ を選択しているときに限り、他の文書 j から i に対して親子関係を認めることを意味する。以上の設定によって、$\Omega_{\mathrm{ILP4ID}}(S)$ は $\Omega_{\mathrm{ILP4ID}}(\mathbf{x}) := \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} x_{ij}\cdot s(d_i,d_j)$ で表すことができる。

2) 本稿では大規模事前学習モデル (Pre-trained Language Models) とコサイン類似度を用いる。対象とする技術文書や教師ラベルの存在によっては、多様化タスクに向けてファインチューニングをしたり、別の類似度を採用してよい。

3 類似度近似に基づく計算手法

本章では 2.2 節で述べた既存手法 ILP4ID を拡張することで、一般的な文書間の構造を表現できる数理最適化問題を提案する。階層関係を親子関係から森構造に拡張することで、芋づる式の親子関係（文書親 $d_1\succ$ 文書子 $d_2\succ$ 文書孫 d_3）などを表現できるようになることが期待できる。一方で Ω_{ILP4ID} で扱った深さ 1 の親子関係 $d_1\succ d_2$ だけではなく、より深い位置の関係（d_1 と d_3 など）最適化問題の中でモデリングする必要がある。

3.1 提案手法

一般化した階層関係を考慮するモデルを提案するために、我々は類似度が分解できる仮定を置く。具体的には D 上の類似度行列 $\mathrm{S} := \left(s(d_i,d_j)\right)_{i,j\in[n]}\in\mathbb{R}^{n\times n}$ を低ランク ($l\ll n$) の因子行列 $\mathrm{L}\in\mathbb{R}^{n\times l}, \mathrm{R}\in\mathbb{R}^{n\times l}$ によって $\mathrm{S}\approx\mathrm{LR}^T$ として近似できると仮定し、計算モデルを提案する。

選択した文書を根とした根付き森を表現する手法を説明する。ここで $\langle n\rangle := \{\mathbf{0}\}\cup[n]$ としてダミー文書 $\mathbf{0}$ を用意する。根付き森の上位にダミー文書 $\mathbf{0}$ を仮定し、根付き森の根を $\mathbf{0}$ の子とみなすことで、根付き木を用いて間接的に根付き森を表現する。ILP4ID と同様に、選択する集合は $S=\{d_j\mid x_{\mathbf{0}j}=1\}$ と解釈する。決定変数として $x_{ij}\in\{0,1\}, i\in\langle n\rangle, j\in\langle n\rangle$ を用いて根付き木を表す。ただし $x_{ij}=1$ は $x_i\succ x_j$ を意味する。また木 T は頂点数 $n+1$ の頂点集合と辺集合 $\{\{i,j\}\mid x_{ij}=1\}$ で表現されるとする。スコア計算のために $i\in\langle n\rangle$ について、実変数 $\mathbf{f}_i^{\mathrm{sum}}\in\mathbb{R}_{\geq 0}^{l}$ と $g_i^{\mathrm{eval}}\in\mathbb{R}_{\geq 0}$ を追加で用意する。

以下に実装の詳細を述べる。

木構造の表現　最も基本的な根付き木の表現を実装する。つまり $\{x_{ij}\}$ について以下を採用する。

- $\sum_{i,j} x_{ij}=n$: 木の辺の数に関する制約
- $\sum_j x_{ji}\leq 1\ (\forall i\in\langle n\rangle)$: 木構造の次数に関する制約
- $x_{ii}=0\ (\forall i\in\langle n\rangle)$: 木構造に関する制約（ILP4ID と異なるモデリングによる）
- $\sum_{i\in T',j\in T'\setminus\{i\}} x_{ij}\leq |T'|-1\ (\forall T'\subseteq\langle n\rangle, |T'|\geq 2)$: ここで T' は頂点数 2 以上の全ての部分木に対する制約である。

目的関数の表現　式 (3) の目的関数を、類似度の低ランク近似の仮定に基づいて拡張する。補助変数 $\mathbf{f}_i^{\text{sum}}$ と g_i^{eval} について以下を設定する。

- $\mathbf{f}_i^{\text{sum}} = \mathrm{R}_{i,:} + \sum_j \mathbf{f}_j^{\text{sum}} x_{ij}$ $(\forall i \in \langle n \rangle)$: 子の右側評価値 R_i を伝搬する。
- $g_i^{\text{eval}} = \langle \mathrm{L}_{i,:}, \sum_j \mathbf{f}_j^{\text{sum}} x_{ij} \rangle + \sum_j g_j^{\text{eval}} x_{ij}$: 子から伝搬した右側評価値と多様性スコア、自分自身の左側評価値 $\mathrm{L}_{i,:}$ の内積を用いて、全体の多様性スコアを求める。

以上の補助変数と制約を用いて式 (4) を定義し、検索結果多様化と階層構造の決定を同時に達成する。

$$\begin{aligned} \max_{x, f^{\text{sum}}, g^{\text{eval}}} \quad & \lambda(n-K)R(x) \\ & + (1-\lambda)K\Omega_{\text{ours}}(x, f^{\text{sum}}, g^{\text{eval}}). \end{aligned} \tag{4}$$

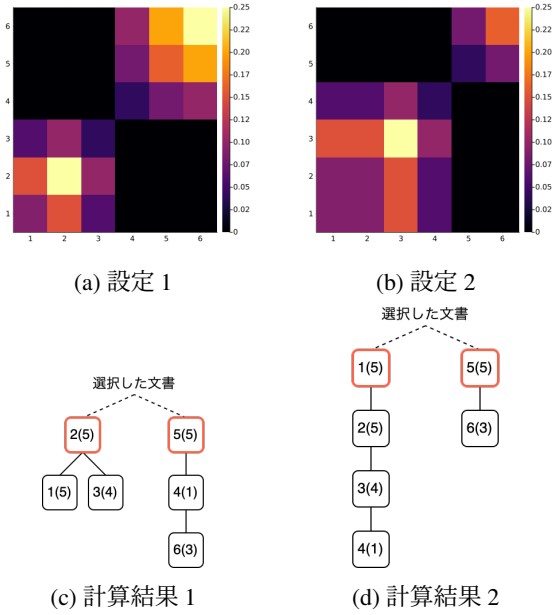
ただし $R(x) := \sum_i x_{\mathbf{0},i} r(q, s_i)$ は関連度スコアを、$\Omega_{\text{ours}}(x, g) := \sum_i x_{\mathbf{0},i} g_i^{\text{eval}}$ は階層構造に基づいて評価した多様化スコアを表す。それぞれの項の重み付けは ILP4ID と同様に設定した。

3.2　原理確認

人工データを用いて、本稿で提案した計算が可能なことを検証する。簡単化のために、$n = 6$ について、$r(\mathbf{v}_q, \mathbf{v}_{d_i})$ がベクトル $[5, 5, 4, 1, 5, 3]$ のように与えられて、$K = 2$ の場合を考える。多様化を行わない場合、スコア 5 である d_1, d_2, d_5 のうち 2 つが選ばれる。(例えばインデックス i の小さい順に d_1, d_2 が検索結果として出力される。

一方で検索結果多様化の文脈では、提案手法を用いることで異なる結果が得られる。図 1 に 2 つの問題設定と最適化問題を解いた結果を示す。類似度行列 S の具体的な例として、クラスタ内の文書数が異なる 2 クラスタの例を想定し、図 1a と図 1b を設定した。計算の結果得られた文書の親子関係を森構造として表現したものを図 1c と図 1d にそれぞれ示す。図中の木ノード中の番号と数値は、文書 d_i について番号 i とスコア $r(\mathbf{v}_q, \mathbf{v}_{d_i})$ に対して $i(r(\mathbf{v}_q, \mathbf{v}_{d_i}))$ と描画した。それぞれの森の根として選ばれた文書番号（図 1c では 2 と 5、図 1d では 1 と 5）は、それぞれ文書 $\{d_2, d_5\}$ と $\{d_1, d_5\}$ が検索結果として得られたことを表す。

結果より、図 1c は図 1a のクラスタ構造を、図 1d は図 1b のクラスタ構造をそれぞれ発見しつつ、階層構造を構築しながら、スコアの高い文書 d_1, d_2, d_5 から $K = 2$ 文書を代表的な検索結果として選択することができた。このとき、森の根として選択した文書は多様化を達成しつつ、各文書 $d_j \in S$ については、それと類似した文書を階層構造（木）として表現できる。そのため、ILP4ID と比較して複雑な構造を森構造として表現しつつ、その根に K 子の文書を選択することが出来た。



(a) 設定 1　(b) 設定 2

(c) 計算結果 1　(d) 計算結果 2

図 1: 原理確認のための設定（上段）と結果（下段）。得られた森構造のうち、木の根が検索結果として選択された文書 $K = 2$ 個に相当する。

3.3　ILP4ID を用いた近似解法

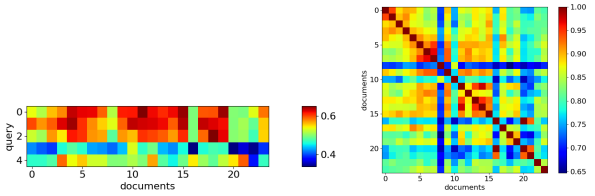
3.1 節で述べた手法によって、全ての部分木構造を表現した最適化問題を議論することができる。一方で、組合せ最適化によって根付き木や根付き森を表現することは、工夫なしでは計算量的に困難なことが多く、現在の定式化と解法では大量の文書を扱うことが難しい。よって本稿の実験では、近似的に木構造を求めることを目的として、深さ 1 の親子関係のみを計算する ILP4ID を繰り返し計算することで、指定の深さまでの階層構造を作成する手法を実装した。

4　実験と考察

提案手法を小規模データに適用し、結果を観察して評価・議論を行う。

4.1　社内文書での適用例

データ　原理検証のために、社内文書のタイトルを d_i と設定し、クエリ q を 5 種類与えて、多様なタイトルの社内文書を階層構造とセットで取得す



(a) 関連スコア　(b) 文書間類似度 **S**

図 2: 実験用のデータ: 2a では y 軸が 5 種類のクエリを、x 軸が 24 本の文書を示す。2b では 24 本の文書間の類似度行列 S を示す。

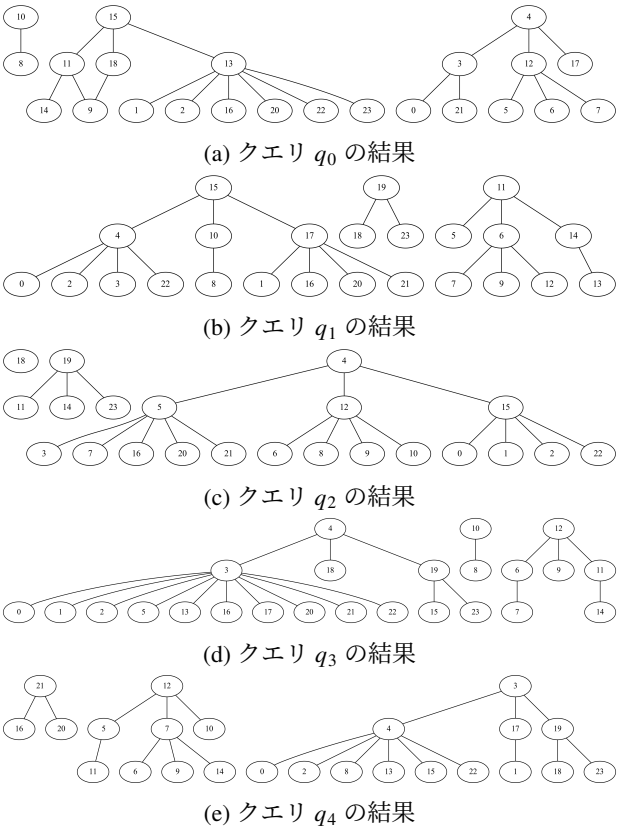
る問題設定を考える。なお予稿では機密情報の扱いから、文書 d_i やクエリ q の情報は公開が容易ではないため、図 2 に類似度の情報のみ可視化したものを掲載する。図 2a はクエリと文書の関連スコアを、図 2b は多様性スコアの定義に用いる文書間の類似度を示す。タイトルやクエリの埋め込みには日本語 ALBERT モデル（`ALINEAR/albert-japanese-v2`）をファインチューニングなしで用いて、ベクトル間の類似度にはコサイン類似度を用いた。

近似計算の結果得られた階層構造　ILP4ID を繰り返して解く近似計算法を用いて計算を行った。近似解法では各階層で $K = 3$ の文書を選び、深さ 3 まで再帰的に木構造の構築を行った。ILP4ID 自体のパラメータは $\lambda = 0.9$）と設定した。図 1a の y 軸に対応するクエリ q_0 からクエリ q_4 までの 5 つのクエリを用いて検索を行った。最適化問題を近似的に解いて求めた文書の森構造を図 3 に可視化した（クエリ q_0 の結果が図 3a、$q_1, \ldots, q_4$ の順で、q_4 の結果が図 3e に対応する）。各クエリ q_i と文書 $D = \{d_1, \ldots, d_{24}\}$ の関連スコアが図 1a のように異なるために、得られる森もそれぞれ異なっている。

4.2　考察と課題

提案手法を用いてクエリ q に対する検索結果を根付き森として取得することが出来た。式 (4) で定義したように、各部分木の類似度が高くなるように Ω_{ours} が設定されており、構築した木は文書のクラスタを形成する。結果として検索結果多様化が達成される。一方で $R(S)$ については、ILP4ID と同様に選択した文書 S のみ（森の根のみ）で評価し、森構造を利用していないため、評価指標を一般化したり、修正したりする余地がある。

現在の近似解法は、3.1 節で提案した元の最適化問題が計算量に困難であるために手法である。よって元の問題を直接最適化する近似手法や、ニューラルネットワークなどと合わせた手法を構築する余地があると考えられる。



(a) クエリ q_0 の結果

(b) クエリ q_1 の結果

(c) クエリ q_2 の結果

(d) クエリ q_3 の結果

(e) クエリ q_4 の結果

図 3: 得られる高さ 3 までの森構造。森構造のノード i は文書 d_i を表す。

数理最適化は、制約の表現力が高いという特徴がある。文書同士を同じクラスタに属するように制約して森構造を取得したり、ユーザが探索的に文書探索を行う際に、関連した文書を連続的に検索する行為（例えばシリーズ 1、2、3 というように順序がある場合）などに対応した森構造を得ることで、より良い検索体験に向けた手法を検討できると考えられる。このように検索システムを指向した最適化問題へと精緻化することは今後の課題の一つである。

5　まとめ

大量の技術文書から、自分の関心がある文書を適切に探し出すタスクは重要である。本稿では類似度の低ランク近似と数理最適化を用いて、検索結果多様化の新しいモデルを提案して解法を構築し、小規模な実験を用いて議論を行った。今後はオープンデータを用いた大規模な実験や、数理最適化モデルの精緻化・計算手法の高速化、更に検索システムの検証を行う予定である。

参考文献

[1] Christopher D Manning. **Introduction to information retrieval**. Syngress Publishing, 2008.

[2] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. **Modern information retrieval**, Vol. 463. ACM press New York, 1999.

[3] 打田智子, 古澤智裕, 大谷純, 加藤遼, 鈴木翔吾, 河野晋策. 検索システム 実務者のための開発改善ガイドブック. ラムダノート, 2022.

[4] 持橋大地. Researcher2vec: ニューラル線形モデルによる自然言語処理研究者の可視化と推薦. NLP2021, 2021.

[5] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Search result diversification. **Foundations and Trends® in Information Retrieval**, Vol. 9, No. 1, pp. 1–90, 2015.

[6] Hai-Tao Yu, Adam Jatowt, Roi Blanco, Hideo Joho, Joemon Jose, Long Chen, and Fajie Yuan. A concise integer linear programming formulation for implicit search result diversification. In **Proceedings of WSDM2017**, pp. 191–200, 2017.

[7] Guido Zuccon, Leif Azzopardi, Dell Zhang, and Jun Wang. Top-k retrieval using facility location analysis. In **Proceedings of ECIR2012**, pp. 305–316. Springer, 2012.