2M-06

テキスト及び構造に基づく効率的な類似部分木検索手法

溝上拓也 † 天笠俊之 ‡ SavongBou‡

† 筑波大学情報学群情報科学類 ‡ 筑波大学計算科学研究センター

1 はじめに

近年の XML や JSON などの階層データ形式の普及に伴い，データが木構造で表現されることが多くなってきた．こうした木構造データの増加に伴い，木構造データの検索や結合の高速化手法が多く考案されてきた．その中の一つに，類似部分木検索がある．類似部分木検索とは，クエリ木を入力として，データベース木の中からクエリ木と類似した部分木を検索する問題である．類似部分木検索は，木構造データの相互補完，重複除去，コピー検出等，様々な用途に活用できる．類似部分木検索における既存研究では，木構造データの類似度として木の構造に基づく構造類似度やデータレコード (葉ノードのラベル) の情報に基づくテキスト類似度が用いられてきた．構造類似度は，木編集距離に基づく類似度が代表的である．木編集距離に基づく類似部分木検索の先行研究として，TASM[1] や StructureSearch[2], SlimCone[3] などがある．これらの手法では，文献データベースなど似た構造の部分木が多く含まれるデータベース木から検索する場合，候補部分木の葉ノードに含まれるテキスト情報が異なっていてもクエリ木との構造類似度が高くなる傾向があり，真に類似していると言えないものが検索結果に含まれる可能性がある．そのため，検索精度の面で問題が生じる．上記の問題に対処するため，構造類似度とテキスト類似度を併せた総合類似度による類似部分木検索手法が提案されてきた．これまで提案されてきたテキスト類似度として，小柳ら [4] のテキストの単語集合のジャカード係数や小久保ら [5] のジャカード係数と概念階層・同義語階層を組み合わせた類似度指標がある．これらの手法では，葉ノードのテキスト情報の類似性を考慮することで，検索精度の向上を図っている．これらの手法は TASM のような検索時にドキュメント木を全て走査するインデックスフリーな手法であるが，候補部分木に対して先にテキスト類似度を計算し，明らかに類似しえない候補部分木を除外することで，構造類似度のみを用いたインデックスフリーの手法よりも実行時間を短縮している．本研究では，小柳ら，小久保らの手法をベースとして，テキスト類似度にテキストデータから計算した分散表現ベクトルのコサイン類似度を用いた手法を提案する．また，提案手法と小柳・小久保らの既存手法との実行時間・精度の観点から比較する．

Efficient Subtree Similarity Search Based on Structure and Text

Takuya Mizokami†(mizokami@kde.cs.tsukuba.ac.jp), Toshiyuki AMAGASA‡(amagasa@cs.tsukuba.ac.jp) and Savong Bou‡(savong-hashimoto@cs.tsukuba.ac.jp)
†College of Information Sciences, University of Tsukuba
‡Center for Computational Sciences, University of Tsukuba

2 提案手法

問題定義: 本手法では，クエリ木 Q とデータベース木 D，重み係数 $\alpha \in [0,1]$，類似度しきい値 $\theta \in (0,1]$ が与えられたとき，Q との類似度が θ 以上の部分木 T を D から検索する問題を扱う．Q と T の類似度は，Q と T の構造類似度 $Sim_s(Q,T)$ とテキスト類似度 $Sim_t(Q,T)$ を線形結合した総合類似度 $Sim(Q,T)$ によって定義される．ただし，$Sim_t(Q,T)$ には，木の相対的なサイズ差 $d(Q,T)$ によるペナルティが付与されている．

$$Sim(Q,T) = \alpha Sim_s(Q,T) + (1-\alpha)\frac{Sim_t(Q,T)}{1+d(Q,T)} \quad (1)$$

$$Sim_s(Q,T) = 1 - \frac{TED(Q,T)}{|Q|+|T|} \quad (2)$$

$$Sim_t(Q,T) = \frac{1}{2} \times (\cos(\overline{\mathbf{dv}(\mathcal{L}_Q)}, \overline{\mathbf{dv}(\mathcal{L}_T)}) + 1) \quad (3)$$

$$d(Q,T) = \frac{||Q| - |T||}{\min\{|Q|, |T|\}} \quad (4)$$

ここで，$\mathcal{L}_Q$，$\mathcal{L}_T$ はそれぞれ Q と T の葉ノードの集合である．$\overline{\mathbf{dv}(\mathcal{L}_Q)}$，$\overline{\mathbf{dv}(\mathcal{L}_T)}$ はそれぞれ $\mathcal{L}_Q$，$\mathcal{L}_T$ の葉ノードの分散表現ベクトルの平均であり，$\cos(\overline{\mathbf{dv}(\mathcal{L}_Q)}, \overline{\mathbf{dv}(\mathcal{L}_T)})$ は $\overline{\mathbf{dv}(\mathcal{L}_Q)}$ と $\overline{\mathbf{dv}(\mathcal{L}_T)}$ のコサイン類似度である．$TED(Q,T)$ は Q と T の木編集距離である．

提案手法: $Sim(Q,T)$ は，Q と T のノード数の差が大きくなるほど低くなる傾向があり，Q のノード数 $|Q|$ は一定であるから，候補部分木 T のサイズ $|T|$ の範囲 $[\mathrm{minT}, \mathrm{maxT}]$ が事前に求められる．これにより，T のサイズが $[\mathrm{minT}, \mathrm{maxT}]$ の範囲に入らない部分木に対する検索を省略することができる．また，計算量の重い構造類似度を計算する前にテキスト類似度を計算し，明らかに類似しえない候補部分木を除外するこ

とで，構造類似度の計算回数を減らす．

3 評価実験

データセット データセットとして，医学を中心とする生命科学の文献情報を収集したオンラインデータベースである MEDLINE[1] を使用した．MEDLINE では，文献の情報を著者やタイトルなどのデータフィールドで管理しているが，そのうちの一つに，文献の内容を表す MeSH(Medical Subject Headings) [2] と呼ばれる専門用語辞書がある．MeSH は，Descriptor(定義語)，Qualifier(副題語)，Supplementary Concept Record(SCR) などから構成されており，Descriptor は階層構造を持つ．よって実験ではこれを概念階層として利用した．同義語規則には MeSH の Entry Term を利用した．データベース木とクエリ木は，MEDLINE データセットを加工して作成した．MEDLINE の冒頭 30000 文献に対して，ランダムに選んだ 1 つの文献をクエリ木とした．データベース木については，クエリを除いた 29999 件の文献に，クエリ木の各テキストノードのラベルを 30% の確率でランダムに変更した文献を 1000 件追加し，計 30999 件の文献をデータベース木とした．ラベルを変更する際には，同義語規則・概念階層から類似度が 0.8 以上のテキストに変更するようにした．データベース木のサイズは 190MB，部分木数は 9280877 個である．一方，クエリ木のサイズは 5.07kB，頂点数は 206 である．

テキストノードの分散表現ベクトルは，NLP ライブラリである Spark-NLP[3] を用い，MEDLINE データセットで訓練された Bert モデル[4] を用いて計算した．分散表現ベクトルの 786 次元の 32 ビット浮動小数点数である．

表 1 に実験結果を示す．実験では，類似度しきい値 θ を最大値 1.0 から、再現率 (Recall) が 1.0 になるまで下げていき，そのときの適合率 (Precision) と実行時間を測定した．構造類似度のみの手法では重み係数を 1 とし、それ以外では 0.5 とした．実験結果から，提案手法は既存手法に比べて精度・実行時間ともに劣っていることがわかる．

表 1: 実験結果

手法	実行時間 (s)	Precision	Recall
構造類似度のみ	46.5	0.12	1.00
小柳ら [4]	14.4	0.72	1.00
小久保ら [5]	58.6	0.12	1.00
提案手法	178.6	0.01	1.00

4 まとめ

本研究では，テキスト及び構造に基づく効率的な類似部分木検索手法を提案した．現状の提案手法では，実行時間・精度ともに既存手法に劣っている．今後は，テキスト類似度の評価指標として，小久保らの手法のように各ラベルの表現ベクトルのコサイン類似度のマッチングによる類似度を検討する．

謝辞

この研究の成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の「ポスト 5G 情報通信システム基盤強化研究開発事業」（JPNP20017）の委託事業，JST CREST（JPMJCR22M2），科学研究費補助金（JP22H03694, JP23K16888）に支援によるものである．

参考文献

[1] Augsten, N., Barbosa, D., Böhlen, M. and Palpanas, T.: TASM: Top-k Approximate Subtree Matching, in *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pp. 353–364 (2010).

[2] Cohen, S.: Indexing for Subtree Similarity-Search Using Edit Distance, in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, p. 49–60, New York, NY, USA (2013), Association for Computing Machinery.

[3] Kocher, D. and Augsten, N.: A Scalable Index for Top-k Subtree Similarity Queries, in *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, p. 1624–1641, New York, NY, USA (2019), Association for Computing Machinery.

[4] 小柳涼介, 天笠俊之, 北川博之：大規模 XML データにおける効率的な重複データ検出, 第 76 回全国大会講演論文集, Vol. 2014, No. 1, pp. 607–608 (2014).

[5] 柚真小久保, 俊之天笠, 博之北川：複数の類似度を考慮した木構造データに対する類似部分木検索, 第 83 回全国大会講演論文集, Vol. 2021, No. 1, pp. 349–350 (2021).

[1] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

[2] https://www.nlm.nih.gov/databases/download/mesh.html

[3] https://sparknlp.org/

[4] https://sparknlp.org/2021/08/31/sent_bert_pubmed_en.html