D-014

Steiner Tree を利用した Wikipedia における関係の抽出
Extraction of Relationships from Wikipedia Using Steiner Tree

森廣 恭平† 張 信鵬‡ 浅野 泰仁‡ 吉川 正俊‡
Kyohei Morihiro Xinpeng Zhang Yasuhito Asano Masatoshi Yoshikawa

1. はじめに

近年，Wikipedia から有用な知識を抽出する研究が盛んに行われている．Wikipedia の各記事は単一の概念について説明をしている．また，記事同士は相互にハイパーリンクで密接に繋がっている．したがって，記事間のリンク構造は，記事（概念）を節点，リンクを枝とするグラフとして捉えることができる．また，Wikipedia には「国」や「日本の国会議員」などの，概念の集合を扱うためのカテゴリ情報も存在する．

我々は，Wikipedia を用いて，複数の概念間の関係を利用者に提示することを目的とした研究を行っている．特に，本論文では，一つのカテゴリと複数の概念が入力として与えられたときに，そのカテゴリに属し，かつ入力された複数の概念すべてと関係の強い概念を求め，これらの概念間の関係を提示する問題を考える．例えば，入力カテゴリとして「国」，入力概念として「茶」，「仏教」が与えられたならば，茶および仏教と関係の強い国々を求め，これらの国々と茶および仏教との関係を成り立たせているものを提示する．

この問題を解くために，本論文では，top-k グループシュタイナー木を用いた手法を提案する．重み付きグラフの与えられた節点集合に対する最小シュタイナー木とは，その節点集合のすべての節点を含む最小重みの木であり，これを用いてグラフの複数の節点間の関係を求める研究は多く行われている．これを拡張した top-k グループシュタイナー木を用いることで，入力カテゴリに属し，かつ入力概念と関係の強い上位の概念を求め，さらにこれらの概念間の関係を提示することができると考えられる．

これまでにも，Wikipedia のリンク構造を解析することによって，概念間の関係，たとえば与えられた二人の人間の関係や，ある国とある資源の関係などの抽出を行う研究は数多く存在する．これらは主に，ある二つの概念同士の関係や，ある一つの概念と関連の強い概念を抽出するものが多いが，提案手法は与えられた複数の概念と関係の強い概念の発見を目的としている．

2. シュタイナー木問題

節点集合 V，枝集合 E，非負重み関数 $w(E)$ からなる重み付きグラフ $G(V, E)$が与えられたとする．枝 e_i の重みは $w(e_i)$で表す．ターミナル集合 $V' \subseteq V$ が与えられた時，V'の全ての節点を含んだ G の部分木 T を，G, V' に対するシュタイナー木と言う．さらに，その中で重みが最小のものを見つける問題を最小シュタイナー木問題と言う．木 T の重みは，木 Tが含む全ての枝の重みの総和で表す．

最小シュタイナー木問題はグループシュタイナー木問題(GST)に拡張できる．重み付きグラフ $G(V,E)$，n 個のターミナル点集合 $V_1,...,V_n \subseteq V$ が与えられた時，各 V_i の少なくとも一つの節点を含むような部分木が $G, V_1,...,V_n$ に対するグループシュタイナー木であり，そのうち重みが最小のものを求める問題が GST である．

さらに，最小シュタイナー木問題及び GST は「top-k の解を求める問題」へと拡張することができる．Top-k シュタイナー木問題は，重みが小さいものから順に k 個のシュタイナー木を求める問題として定義される．Top-k GST についても同様に，重みの最も小さい k 個のグループシュタイナー木を求める問題として定義される．

Top-k シュタイナー木問題に対しては STAR [1]に代表されるように多くの近似アルゴリズムが研究されている．STAR は関係データベースや RDF などの知識ベースにおいて，意味的情報を利用して高速に top-k シュタイナー木を計算し，項目間の関係を得ている．しかし Wikipedia ではリンクに重みや意味的情報は付与されていないので，STAR をそのまま適用しても良い結果は得られないと考えられる．

3. 抽出手法の概要と考察

3.1 Top-k グループシュタイナー木を用いた手法

問題の入力は Wikipedia 上の n 個の概念 $a_1, a_2, ..., a_n$ と 1 個のカテゴリ c，そして正の整数のパラメータ z である．このとき，提案手法は，c に属する概念のうち，$a_1, a_2, ..., a_n$ と強い関係を持つ z 個の概念を求め，これらの概念間の関係を提示する．

まず，これらに対応する Wikipedia の記事を節点とするグラフ G を以下のように作成する．MediaWiki を用いて，Wikipedia の全てのデータを取得し，データベースに格納することができる．このデータベースを用いて，$a_1, a_2..., a_n$ の各概念に対応する記事と，カテゴリ c に属する各概念に対応する記事を取得する．さらに，これらの記事からリンクを 2 回以下たどることで到達可能な記事も取得する．取得したすべての記事を節点，各記事間のリンクを枝としたグラフを G とする．各枝の重みは，tfidf を記事内のリンクに適用して概念間の関連度を算出して決定する．

次に，グラフ G とターミナル集合 $V_1=\{a_1\}, V_2=\{a_2\}, ..., V_n=\{a_n\}, V_{n+1}=$｛カテゴリ c に属する各概念の記事｝に対するグループシュタイナー木を重み最小のものから順に求めていく．例えば，「茶」，「仏教」，「カテゴリ：国」を入力とした場合，ターミナル集合は $V_1=$｛茶｝, $V_2=$｛仏教｝, $V_3=$｛「カテゴリ：国」に属する記事の集合｝となる．この例に対する最小グループシュタイナー木を図 1 に示す．実線部が最小グループシュタイナー木の枝であり，最も木の重みが小さくなるように選ばれている．「インド」，「日本」はともに V_3 に属しているが，この最小グループシュタイナー木に含まれている V_3 に属する記事は「日本」であり，これは「茶」と「仏教」に最も強い関係を持つ国は

† 京都大学工学部情報学科
‡ 京都大学大学院情報学研究科

日本であることを表している．さらに，各ターミナルがどのような節点を通って繋がっているかを見ることによって，それぞれの節点間の関係を理解することができる．図 1 では，「禅」は「茶」，「仏教」，「日本」のすべてと繋がっており，三者と深い関係のある項目だということが分かる．

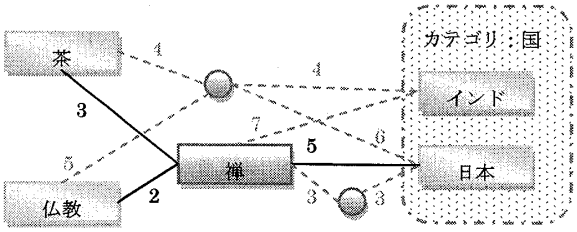

図 1: 最小グループシュタイナー木の例

なお順に求められたグループシュタイナー木のいずれかに現れた V_{n+1} の点の集合を C とし，$|C|=z$ となるまでグループシュタイナー木を求めていくものとする．すなわち，カテゴリ c に属する概念のうち，$a_1, a_2, \ldots, a_n$ と強い関係を持つと判定された概念の集合が C である．ここで求められた top-k グループシュタイナー木の数 k は，z に等しいとは限らない．いくつかのグループシュタイナー木に同じ V_{n+1} の点が重複して現れることがあるからである．

さらに，図 1 の例の「禅」のように，これらの概念間の関係を構成するために重要な節点及び枝を発見するため，求めた top-k グループシュタイナー木の節点と枝に，その重要さを表すスコアを与える．より上位(すなわち重みの小さい)シュタイナー木に現れた節点と枝は重要であると考えられ，また数多くのシュタイナー木に現れる節点と枝も関係が強いと考えられるので，以下のようにスコアを設定する．すなわち，各スコアの初期値は 0 であり，重みが i 番目に小さなグループシュタイナー木に現れた節点と枝のスコアに i の逆数を加算していく．これを i=1 から k まで繰り返す．

Top-k シュタイナー木および top-k グループシュタイナー木の特性として，求められた k 個のシュタイナー木に同じ枝が何回も現れることが挙げられる．このような枝のスコアは極めて高くなる．実際，いくつかの例を検証した結果，このような枝は概念間の関係を構成する上で重要なことが多かった．しかし，他の枝と比べ極めて重みの小さい枝が存在した場合，常にその枝ばかりがシュタイナー木に現れ，結果として同じようなシュタイナー木ばかりが求まってしまい，V_{n+1} の z 個の点が現れるまでに必要なグループシュタイナー木の数 k が非常に大きくなってしまう可能性がある．これを回避するため，グループシュタイナー木を一つ求めるたびに，スコアがある閾値以上となった枝をグラフ G から除去し，次のグループシュタイナー木を求める．

3.2 概念間の関係の提示

3.1 の手法で求めた top-k グループシュタイナー木を，図 2 に示すように一つに重ね合わせて利用者に提示する方法を提案する．k 個のシュタイナー木を個別に提示しても，利用者は複数の概念間の関係の全貌を直感的に理解しづらいと考えられるからである．提案のように，複数のシュタイナー木の結果を重ね合わせて表示することによって，利用者に視覚的に分かりやすく概念間の関係を提示することができると思われる．実際，図 2 の例では，重ね合わせて表示したことによって，「禅」が日本だけではなく，中華人民共和国やインドなど「茶」および「仏教」と関係の強い複数の国々にとって重要な概念であると理解しやすくなっている．

さらに，各節点と枝はスコアに応じて大きさを変えて表示する．これにより，図 2 では，「茶」および「仏教」と「日本」または「中華人民共和国」が強い関係を持っているということをより理解しやすくなっている．また，ターミナル間にある節点でスコアの高いものは，関係を構成する上で重要な節点を表していると考えられる．図 2 では，「茶道」や「禅」が，日本を初めとする国と茶・仏教との関係を構成していることが理解できる．

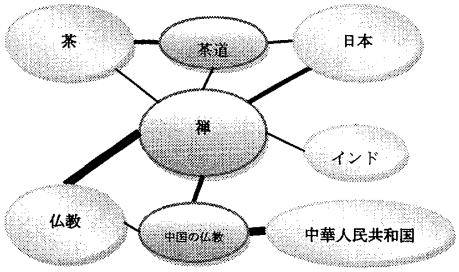

図 2: top-k グループシュタイナー木による関係の提示例

4. 既存研究との比較

Tong と Faloutsos [2]，Koren ら[3]は，グラフ上の複数の節点間の関係を説明するためのサブグラフを，ランダムウォークを用いて抽出する手法を提案している．これらの手法では多くの枝を持つ節点が過小評価されているが，シュタイナー木を用いた提案手法はそのような節点も平等に抽出できるため，重要な枝および節点をより正確に発見できると考えられる．その理論的ないし実験的検証については現在計画中である．

また Zhang ら[4]は減衰流を用いて Wikipedia の二つの概念間の関係の強さを測る手法を提案し，さらにその関係を構成する重要なパスを提示することでその関係を説明している．Zhang らの手法は三つ以上の節点間の関係を扱えないが，我々の手法は三つ以上の節点間の関係を扱うことが可能である．

5. おわりに

本研究では，Wikipedia において，あるカテゴリに属する概念のうち，与えられた複数の概念と関係の強いものを求め，それらの関係を提示するために，top-k グループシュタイナー木を用いた手法を提案した．今後は提案手法の有用性を検証する予定である．

参考文献

[1] G. Kasneci, M. Ramanath, M. Sozio, F. M. Suchanek, G. Weikum, "STAR: Steiner-Tree Approximation in Relationship Graphs", Proc. of 25th ICDE, pp.868-879 (2009).

[2] H. Tong, C. Faloutsos, "Center-Piece Subgraphs: Problem Definition and Fast Solutions", Proc. of 12th ACM SIGKDD, pp.404-413 (2006).

[3] Y. Koren, S. C. North, C. Volinsky, "Measuring and Extracting Proximity in Networks", Proc. of 12th ACM SIGKDD, pp.245-255 (2006).

[4] 張 信鵬, 浅野 泰仁, 吉川 正俊, "減衰流を用いた関係の解析", 第 23 回人工知能全国大会 (2009).