ISS-A-036

注目頂点が指定された大規模グラフの重要部分抽出

柏谷　大吉† 阿部　昇†
†大阪電気通信大学情報通信工学部

1. はじめに

グラフ(graph)とは，頂点(vertex)とそれらを結ぶ辺(edge)と呼ばれる線からなる図形のことである．特に n 個の頂点をもつ単純グラフ G のどの 2 頂点も隣接しているとき，G を完全グラフ(complete graph)といい，K_n と表わす．このとき，K_n の辺数 m は，

$$m = n(n-1)/2s \qquad (1)$$

で求められる[1]．図 1 は K_7 の例である．

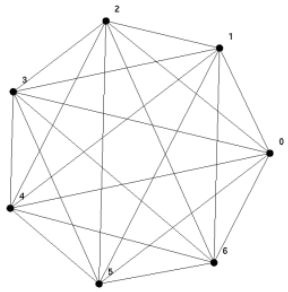

図 1:完全グラフ K_7

例えば，自然言語処理ツールの出力をグラフ化したりするとその出力は完全グラフとなるが，完全グラフは，式(1)から頂点数が増えるにつれ辺数が非常に多くなり，平面上に描画いた際，構造が理解しづらいグラフになることが多い．そこで，本研究では注目すべき 1 つの頂点(以下，注目頂点と呼ぶ)が存在するものとし，その頂点を中心に重要な辺の抽出を行う．

2. グラフ描画に用いるデータ

頂点は x 座標，y 座標ともに-5.0 から+5.0 までの範囲で乱数を発生させることで初期位置を決めた．各頂点は青空文庫の小説 250 冊に対応している．辺の重みは-1.0 から+1.0 の範囲で，各小説間の類似度を入力した．類似度は自然言語処理ツールである doc2vec の出力結果である．

3. 提案手法と描画結果

本研究では2種類の手法を用いて目点を中心に重要部分を抽出し，川西法[2]を用いて描画改善を行い，その後評価を行う．手法 1 の描画例を図 2 に，手法 2 の描画例を図 3 に示す．ここでは，頂点 1 を注目頂点としている．

手法 1 では，注目頂点から類似度の高い順に，頂点次数がある敷居値を超えない範囲で他の頂点を抽出している．そのため手法 1 の描画結果では，注目頂点である頂点 1 からグラフ理論上の距離が離れていることになる頂点 224 まで抽出している．このような注目頂点との関連性が低い頂点が抽出されることは望ましくない．そこで手法 2 として，抽出する頂点に優先順位を設けることにする．初めに注目頂点から距離 1 の頂点の類似度を比較し，類似度の高い順に抽出していく．このとき辺を抽出しすぎることを防ぐ目的で係数αを使用する．αは初期値を注目頂点周りの類似度の平均とし，辺iを抽出するごとに

$$\alpha \mathrel{+}= 1 - S_i \qquad (2)$$

のように増加する．ここでS_iは辺iの類似度である．このαの値を類似度S_iが下回ると距離 1 の辺を抽出するのをやめる．その後αを初期化し，式(2)を用いて距離 2 以上の辺も含めて再度抽出する．距離が近い頂点ほど優先されるようにすることで図 3 のようなグラフが抽出される．

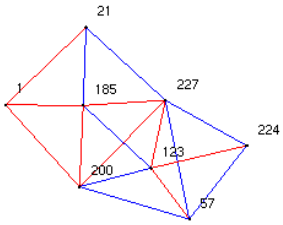

図 2:手法 1 による描画例

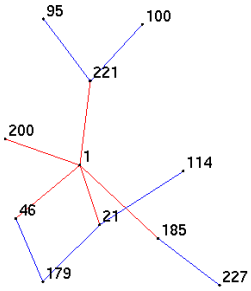

図 3:手法 2 による描画例

4. 評価とまとめ

2 つの手法の評価値を表 1 に示す．手法 2 は手法 1 と比べ，辺の交差数が少なく注目頂点の採用辺数が増えている．よって注目頂点を中心として，より望ましく重要部分を抽出できていると考えられる．辺長平均と分散が増えてしまっているが，これは注目頂点の採用辺数が増えたためであると考えられる．

表 1:各手法の評価

	辺交差数	注目点の採用辺数	辺長平均	辺長分散	頂点と辺の近接接
手法 1	2	3	0.20	0.02	0
手法 2	1	5	0.55	0.22	0

5. 参考文献

[1] 佐藤　公男:原理がわかる工学選書 グラフ理論入門 -C 言語によるプログラムと応用問題-
[2] 川西・増田・山口:“2 種類の理想距離による Eades のグラフ描画法の改良”，電子情報通信学会論文誌，vol.J83-A，No.9 pp.1117-1121，2000.