3ZB-3

時系列に基づいた文献参照関係の可視化-文献のクラスタリング-

岡田 拓也† 渡部 秀文‡ 斎藤 隆文†
東京農工大学大学院 生物システム応用科学府†
東京農工大学 工学府・工学部‡

1. はじめに

ある分野で研究を行うにあたり，その分野の文献を調査することは重要である．文献調査の方法としては，参考文献を辿る，検索エンジンを用いる，などの方法が考えられる．しかし，参考文献を辿る方法では，キーとなる一つの文献の参考文献を確認する場合でも 50 前後，そこから新しい文献を辿ると調査する文献は膨大な数となる．検索エンジンを用いた方法では，関連キーワードで検索を行った場合でも数万以上の検索結果が表示される．どちらの方法でも，膨大な量の文献に目を通す必要があり，研究者の求める文献を探し出すことは困難である．

そこで，テキストベースではなく視覚的に文献調査が行える方法として，文献の参照関係の可視化手法が北川らによって提案された[1]．文献をノード，参照関係をリンクで表現し，時系列順にノードを配置することで注目度の高い文献などの発見を可能にした．しかし，北川らの手法では文献数が膨大な場合，リンクの出入が識別できないという問題が発生した．また，同年発表のノードを縦に積み上げており，縦軸には有用性のある情報が付加されていなかった．

本研究では，北川らによる文献参照関係の可視化手法を改善する．グラデーションを利用したリンクで描画することで，膨大な量の参照関係も識別できるようにする．また，参照関係に基づいてクラスタリングを行い，縦軸に対して有用性の高い情報を付加する．これにより，研究者が文献調査に利用可能なシステムを目指す．

2. 文献データ

文献データは ACM Digital Library[2]から取得する．同サイトでは著者，文献名など以外にも参照・被参照関係についても記載されている．本研究では，対象文献の掲載されているページの html ファイルから必要な情報を抽出することで，文献データを生成する．

Visualization of document reference relationship based on time series- Clustering of document-.
Takuya OKADA†, Hidefumi WATANABE‡,Takafumi SAITO†
†Graduate School of Bio-Applications and Systems Engineering, Tokyo University of Agriculture and Technology
‡Faculty/Graduate School of Engineering Tokyo University of Agriculture and Technology.

3. 提案手法

3.1. 可視化手法の概要

可視化は図 1に示すようにして行う．各ノードは文献を表し，擬似カラー(図 2)に従って文献が参照された回数を表現する．ノードは横軸を発表年順，縦軸を階層型クラスタリングによって割り当てられたクラスタとして対応箇所に配置する．一番下のクラスタには，階層型クラスタリングを行った際に，参照関係が乏しく除外したデータなどをまとめる．また，クラスタの左端にはクラスタ間の階層構造を描画する．リンクは文献の参照関係を表し，参照の場合はノードの左側，被参照の場合は右側からリンクを描画する．発表年の新しい文献が古い文献を参照した場合は黒いリンク，同発表年の文献を参照した場合は赤いリンクで描画する．

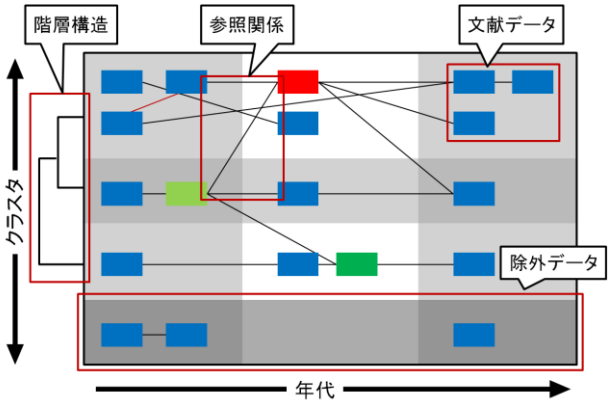
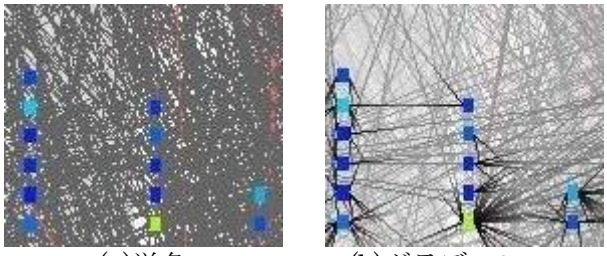


図 1 可視化概要

図 2 擬似カラー



(a)単色 (b)グラデーション

図 3 リンク描画方法の比較

3.2. グラデーションによるリンク描画

可視化対象となる文献データ数が数百以上になるとリンクの出入が識別困難となる．この問題を解決するため，本研究では単色でリンクを描画せず，グラデーションを用いて描画する方法を提案する．ノード付近では濃い色で，ノードから離れると薄い色でリンクを描画することにより，ノードにどの程度のリンクが出入しているかを識別することが可能となる(図 3).

3.3. クラスタリング

縦軸の配置位置を決定するため，文献の参照関係に基づいて階層型クラスタリングを行う．また，生成された木構造から参照関係の乏しいデータを除外し小規模クラスタの出現を抑える．

3.3.1. 参照関係に基づいた類似度の算出

階層型クラスタリングの距離関数として利用する類似度は参照関係に基づいて算出する．類似度は表 1の項目に従って加点し，値が高いほど 2 文献間は類似しているとする．同表(a)は対象文献が参考文献に直接記載されている場合に，(b)は参考文献を辿ることで間接的に参照されている場合に加点する．(c)は 2 文献が同じ文献を参照していた場合に加点する．図 4に示すのは類似度の算出例であり，赤いノードとその他のノード間の類似度を表している．例えば，赤いノードと C は直接の参照関係を持つとともに同じ文献 B を参照しているため 5 点となる．

3.3.2. 木構造の加工

階層型クラスタリングの欠点として，外れ値の扱いが困難である点が挙げられる．本研究においては，0，またはそれに近い類似度しか持たないデータが外れ値に相当する．研究分野の中には非常に小規模のものも存在するほか，収集した文献データの中に分野が異なっているデータが混入している，という可能性も考えられる．これらのデータは小さな節として現れ，これらを残した状態でクラスタの割り当てを行うと，小規模なクラスタが多数出現する．そこで，本研究では木構造の中から任意の大きさ以下の節を除外する．これにより，比較的規模の大きなクラスタを割り当てることを可能にする．木構造から除外したデータは，可視化結果では一番下に一つのクラスタとしてまとめて描画する．

可視化結果を図 5に示す．使用した文献データは，SIGGRAPH で 1990 年から 2000 年までに発表された 542 報の文献である．このデータ群により生成された木構造に対して，15 以下の大きさの節を深さ 20 まで除外する．加工した木構造に対して，深さ 5 までで分割し，それぞれをクラスタとして割り当てたものが同図となる．

表 1 類似度の算出方法

(a)直接的な参照関係	4 点
(b)間接的な参照関係	経路毎に(4-経由した文献数)点
(c)共通の参照関係	(共通の参考文献数)点

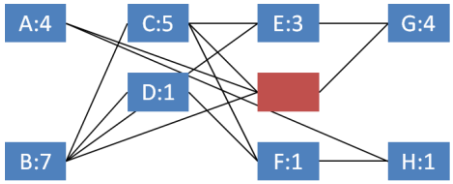


図 4 類似度算出の例

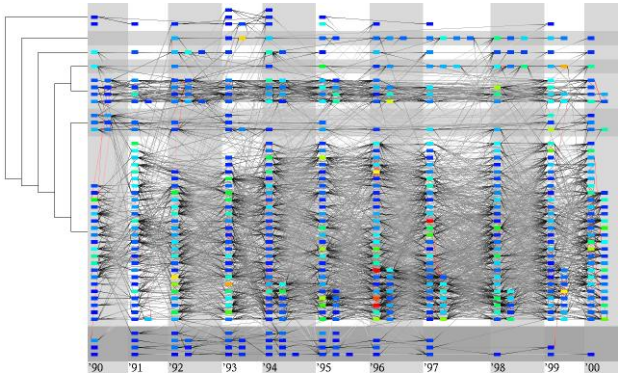


図 5 可視化結果

4. おわりに

本研究では，時系列に基づいた文献参照関係の可視化手法の改善手法を提案した．

リンクをグラデーションで描画することで，どこにリンクが出入しているか識別可能となった．また，参照関係に基づいたクラスタリングにより類似性の高い文献を近い領域に配置し，縦軸に有用な情報を付加することができた．

今後の課題として，木構造の加工方法の改善が挙げられる．除外データの中にも類似性を持つデータが存在するため，それらを適切なクラスタに配置する必要がある．また，加工用パラメータの設定も現在は感覚的に行なっているため，容易に利用できる GUI を用意する必要がある．また，リンクの改善案として，Edge Bandling[3]を用いて複数のリンクを一つに束ねる手法を検討している．これにより，リンクの大まかな流れを把握することが可能になる．

参考文献

[1] 北川晴香, 宮村浩子, 古谷雅理, 斎藤隆文, "文献における参照関係の可視化," 第 70 回情報処理学会全国大会論文誌, 2ZE-5, 2008.
[2] ACM Digital Library, http://portal.acm.org/dl.cfm
[3] A. Lex, M. Streit, C. Partl K. Kashofer, D. Schmalsting, “Comparative Analysis of Multidimensional Quantitative Data,” IEEE Trans. on Visualization and Computer Graphics, Vol. 16, N0. 6, pp.1027-1035, 2010.