6M-4

論文発表年数の分散分析を重みづけとして考慮したネットワーク分析による学術俯瞰

大槻　明[†]　　川上　あゆみ[‡]
お茶の水女子大学[†]　お茶の水女子大学[‡]

1. はじめに

学術俯瞰の分野における最近の研究動向は引用ネットワーク分析が主流であり，自動クラスタ化まで実現されているが，クラスタリングにより同定された各領域の特定や主要論文の自動抽出は実現されていない．ゆえに，本研究ではクラスタリングにより同定された各領域の主要論文を自動で特定する手法について研究する．具体的には，引用する側の論文の発表年数の分散を調べることでそれぞれの重要度の計算し，それらの重要度を基に、時間軸を持つ可視化グラフの構築を目指す．

2. 先攻研究

学術俯瞰の分野において，Small[1]は，被引用数は上位 1%の論文からなる共引用ネットワークを分析し，科学分野で成長している領域を追跡する方法を提案した．また，松尾[2]は，図１のとおり，引用ネットワークの構築，最大連結成分の取得，クラスタリング，可視化を行うことで学術論文引用ネットワークを分析した．しかし，クラスタリングにより同定された各領域の特定や主要論文，主要研究者の抽出について自動化はなされておらず，この部分は専門家が手動で分析しているのが現状である．

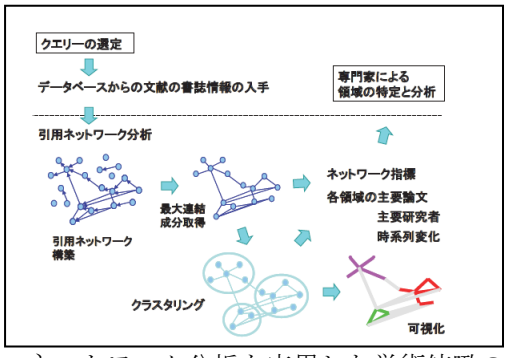

図１　ネットワーク分析を応用した学術俯瞰の手順

3. 提案手法

前節の課題を解決するために，本研究では各領域の主要論文の自動抽出について考える．引用件数が同じでも，「一時期に大量に引用された」場合や、「長期間少しずつ引用されている」場合などが考えられるため、従来の引用分析だけでは、それぞれの重要度を計算する事が難しい．ゆえに，本研究では，上述のそれぞれの「場合」に対し，論文をノード、引用をエッジとする有向グラフと考え，各ノードに発表年数を持たせたうえで，あるノードに入るエッジの元ノードの発表年数の分散を調べることでそれぞれの重要度の計算を試みる．そして、それらの重要度を基に，時間軸を持つ可視化グラフの構築を目指す．以下に全体の流れを記し，次節からその詳細について述べる．

1. 論文 DB からキーワード（クエリ）検索により論文数を絞る
2. 引用論文を中心にリスト化する．
3. 上記 2.のリストから論文発表年数の分散分析を行うことによって、各引用論文に重み付けを行う
4. 上記 3.の重み付けページランクアルゴリズムに適応して各引用論文の重要度を算出する．
5. 重要度（ノード・エッジ）を基に可視化

3.2. 論文 DB からキーワード（クエリ）検索による論文数の絞り込み

本研究では，論文 DB として SCOPUS を採用した．「clustering」というクエリを用いて論文数を絞った結果，87,399 件の論文数に絞り込まれた．

3.3　引用論文を中心にリスト化

前節のリストは，各論文がどの論文を引用したかという並び順になっているが，それを各引用論文がいつ，どのような論文に引用されているのかといった並び順に再リスト化する．

3.4. 論文発表年数の分散分析を行うことによる各引用論文の重み付け

下記 1)～3）により各引用論文の重み付けを行う．

1)　ヒストグラムの最大値を抽出

最も引用数が多い年度を次の関数で抽出し，

MaxYear に格納する．

$MaxYear=max\{y(x)|y(x):=y年に参照された回数\}$ (1)

2） 引用期間の特定

年度の古い年度から 1 年度毎に調べ，最初に見つかった MaxYear の 10%以上の引用数の年度を引用がされ始めた開始年度とし StartYear に格納する．そして 10%以下の引用数の年度になった時点で，その年度を *last year* に格納する．そして論文が引用され始めてから引用されなくなった年度までの期間を次の式で求める．

$Period := (LastYear + 1) - StartYear$ (2)

また，図２のように，ヒストグラムの山が複数存在する場合は，この作業を繰り返しそれぞれ *Period0, 1…nに格納する．*

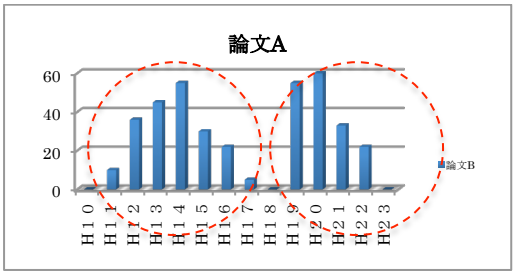

図２ ヒストグラムの形が正規分布から外れているケース

3） ヒストグラムの分散（標準偏差）の算出

当該論文を引用する論文の発表年数の分散（標準偏差）を調べることで当該論文がどのくらいの期間にわたって引用されているかについて調べる．なお，標準偏差の一般的な求め方は次のように表され，求められた標準偏差値は Variance に格納する．

$$Variance = \frac{\sum(x-\bar{x})^2}{n} \quad (3)$$

なお，図２のようにヒストグラムの形が正規分布から明らかに外れているようなケース（山がいくつもある様な場合）は，*Period0, 1…n* の分散（標準偏差）を算出し，それらの平均値を *Variance0, 1…n* に格納する．**そして，** Variance を引用論文の重み付けの値として利用する．

3.5 各引用論文の重要度の算出

PageRankアルゴリズム[3]は，ハイパーリンク構造のような相互参照関係があるときに、 どのページがもっとも「重要」であるかを定量的に算出する手法である．本研究では，このアルゴリズムを利用し，各引用論文の重要度の算出する．なお，この重要度の算出は次のように表される．

ある論文X に対して，
・X の得点を P とする。
・Xが他論文から引用されている得点(重み付け)をそれぞれ$O_1, \cdots, O_n$とする。
このとき、次が成り立つものとする。
$O_1+\cdots+O_n=P$

$$O_1 = \ldots = O_m = \frac{P}{m}\left(= \frac{\sum_{i=1}^{n} I_i}{m}\right)$$

すなわち、各論文に「流れ込む」引用の得点の総和と、各論文から「流れ出す」引用の得点の総和が等しくなるようにして、その総和をその論文の得点と考え，この得点が高いほど、その論文は重要であると考える．そして，この各論文に「流れ出す」引用の得点計算にVarianceの値を適応することにより，各領域における主要論文の特定を目指す．従来のアルゴリズムでは，「流れ出す」引用が複数個あった場合，得点は均等に割り振られていたが，本研究ではVarianceの値が高いものにより多く「流れ出す」と考え計算することで，被引用年次を反映した重要度を計算する．

3.6 重要度を基に可視化

前節で導出した重要度を元に引用ネットワークとして可視化したものが図３である．各ノードには論文名を表示しており，前節の重要度が高いほど，より大きなノードとして表現される．

図３ 3.5節の重要度に基づき可視化した例

4 むすび

本論文では，学術論文引用ネットワーク分析において．クラスタリングにより同定された各領域における主要論文の自動抽出について試みた．具体的には，引用論文の発表年数の分散について分析し，その結果をページランクアルゴリズムに応用することにより各論文の重要度を算出した．そして，その重要度を基に可視化まで行った．今後の課題としては，同じ主要論文の抽出について，専門家が手動で行ったものとの比較検証をすることにより，本研究の有効性について検証していきたい．

参考文献

[1]Small, H. (2006). Tracking and predicting growth areas in science. Scientometrics, 68, 595-610
[2]松尾豊.(2008),'学術俯瞰とウェブからの情報抽出',「イノベーション政策及び政策分析手法に関する国際共同研究」成果報告書No.4,pp43-59
[3]Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, 'The PageRank Citation Ranking: Bringing Order to the Web', 1998