第19回年次大会予稿

引用論文の分散値を重み付けとして考慮したページランクアルゴリズムによる主要論文の抽出

Extraction of key literature based on PageRank algorithm considering variance values of cited literatures as weighting

大槻明[1*], 川上あゆみ[1], 林剛[2], 川村雅義[2]
Akira OTSUKI[1*], Ayumi KAWAKAMI[1], Takeshi HAYASHI[2], Masayoshi KAWAMURA[2]

1 お茶の水女子大学
Ochanomizu University
〒112-8610 東京都文京区大塚2丁目1番1号
E-mail: otsuki.akira@ocha.ac.jp
2 東京大学
The University of Tokyo
〒113-8656 東京都文京区弥生2-11-16東京大学工学部9号館320号室
*連絡先著者 Corresponding Author

学術俯瞰の分野における最近の研究動向は，参考文献の引用分析により実現するサイテーションマップが主流であり、ネットワーク構築やクラスタ化までの自動化はなされているが、各クラスタがどのような集団であるかの意味付けまでの自動化はなされておらず、専門家が手動で分析している現状である．ゆえに、各クラスタの自動解釈を最終的な目的として，本発表では各クラスタの主要論文の自動抽出を目指す。具体的には、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。

Even though Citation Map has been in the mainstream of recent study trend in a field of academic landscape achieving the stage of automated clustering, the reality is that experts manually analyze semantic attachment about what kind of group each cluster is.
Therefore, we try to achieve an automated extraction of key literature of each cluster in the report, by setting automated interpretation of each cluster as the final purpose of the study.

キーワード: 学術俯瞰，引用分析，データベース，ネットワーク分析
Keyword1, Science highangle, Citation analysis, Database, Network analysis

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．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の自動抽出までは実現されてはいない．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，引用される側の論文の発表年数の分散を調べ，その分散値をページランクアルゴリズムに適応することにより各論文の重要度を算出する．また，この重要度を基に、時間軸を持つ可視化グラフの構築を目指す．

2 先行研究

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．

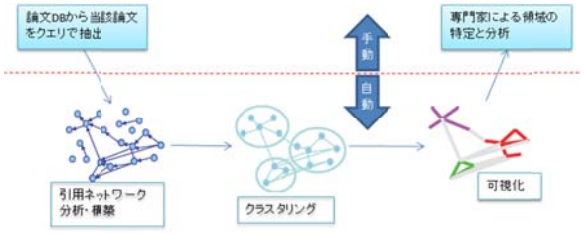

図１　ネットワーク分析を応用した学術俯瞰の手順

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ている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．

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

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．

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

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．

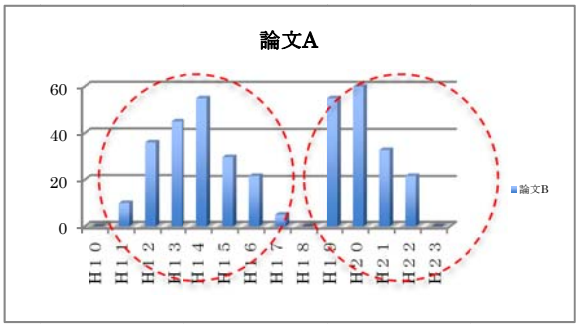

図２ ヒストグラムの形が正規分布から外れているケース

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．

$$Variance = \frac{\Sigma (x-\bar{x})^2}{(n-1)} \quad (3)$$

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Variance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．

(1)各論文は、固有の得点を持っている。
　各引用もまた、固有の得点を持っている。

(2)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をそれぞれ $Variance_1, \cdots, Variance_n$とする。
- Xが他論文を引用している得点をそれぞれ$O_1, \cdots, O_m$とする。

このとき、次が成り立つものとする。

$$Variance_1 + \cdots + Variance_n = P$$

$$O_1 = \cdots = O_m = \frac{p}{m}\left(= \frac{\sum_{i=1}^{n} \mathrm{Variance}_i}{m}\right)$$

すなわち、各論文に「流れ出す」引用の得点の総和と、各論文から「流れ込む」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，各論文から「流れ込む」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,「流れ込む」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「流れ込む」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．なお，本ツールのことをHiAc(Highangle of Academic）と呼ぶ.

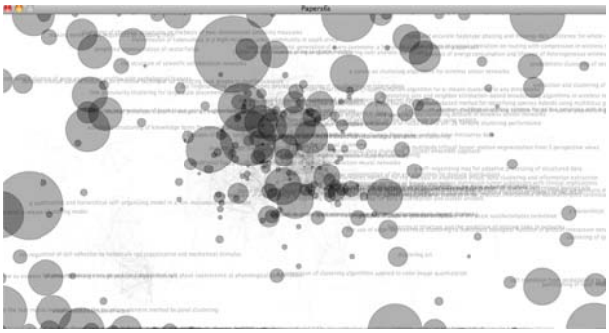

図３　重要度に基づき可視化した例

4 評価実験

4.1 専門家が手動で抽出した主要論文との比較検証

立堀[4]らが2004年に発表した研究動向調査報告は，ＩＢＭの論文データベースを用いて，1999年以降のソフトウエア・アーキテクチャ研究分野の動向を調査し，主要な51論文を手動で抽出したものである．本評価実験では，この専門家が手動で抽出した主要論文をどこまで自動で抽出できるかについて検証する．

立堀らにおける51論文の抽出方法は，GoogleScholarを用いて年あたり引用数を求め，その上位40論文を抽出している．この40論文に加えて，立堀らが特に重要と考えた国際会議に絞ったうえで，ソフトウエア・アーキテクチャに関する11の論文を抽出している．また，立堀らは，定量的な研究動向評価のために，独自の分類方式を採用している，具体的には，ソフトウエア開発プロセスにおいて，ソフトウエア・アーキテクチャの果たす役割を，次の５つの役割のどれに着目しているかによって51論文を分類しており，それを図にしたものが図4である．

・[R]アーキテクチャへの要件(Requirement)
　⇒様々な利害関係者のシステムへの要件をアーキテクチャに反映する．

・[M]アーキテクチャのメタモデル（Metamodel）
⇒アーキテクチャの設計は，メタモデルに基づいてアーキテクトが行う．
・[C] アーキテクチャの用いた利害関係者とのコミュニケーション（Communication）
⇒全ての要件を満たすことができない場合，要件を調整するために，アーキテクチャを用いて利害関係者と交渉する．
・[S] アーキテクチャとシステム間の同期（Synchronization）
⇒抽象的なアーキテクチャは，実際に動くシステムの実装に落とさなければならない．逆に，システムに変更があった場合，それはアーキテクチャにも反映されるべきである．

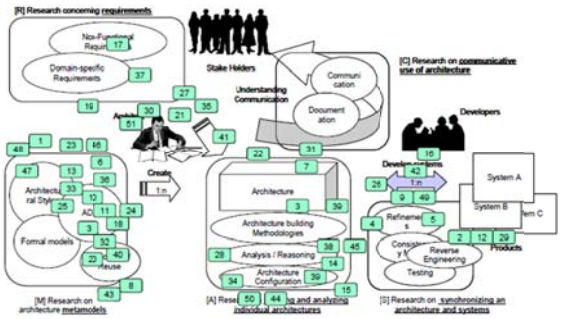

図4　51論文をソフトウエア・アーキテクチャの５つの役割に分類したイメージ

立堀らの報告に対して，HiAcで同様に主要論文を抽出したものが図5である．HiAcにおける論文抽出について述べる．まず，論文DBとして SCOPUSを用いた．そして，クエリとして「Software Architecture」を，また，年代として「1999年～2004年」をそれぞれ設定して論文を抽出した．なお，表1の「SCOPUS」欄のとおり，51論文のうち，○のついた23論文以外は，そもそもSCOPUSには論文が存在しなかったため，対象外としている．また，表1の論文番号を図4及び図5上でも表記している．さらに，図4における５つの役割は，定量的な研究動向評価のために立堀らが独自に設定した分類方式であるため，HiAcの場合（図5）では，手動でクラスタの微調整を行った．

図5から，引用論文の分散値を重み付けとして考慮したページランクアルゴリズム（以下「本アルゴリズム」という）によって可視化した結果，23論文はどれも主要な論文として抽出されていた．特に，論文番号41-51は，立堀らが主要な論文を手動で抽出したものであり，その中でSCOPUSに存在したものは，論文番号44-48であるが，それら全てを主要論文として抽出できていた．

このように，専門家が抽出した論文を本アルゴリズムによって分析することにより，主要な論文として自動で抽出することが可能となる．ただ，[R]のクラスタにおける緑のノードなど，立堀らが抽出していない論文もHiAcでは主要な論文として抽出していたが，これらの論文がどのような意味を持つものかについては，今後の課題として検討していきたい．

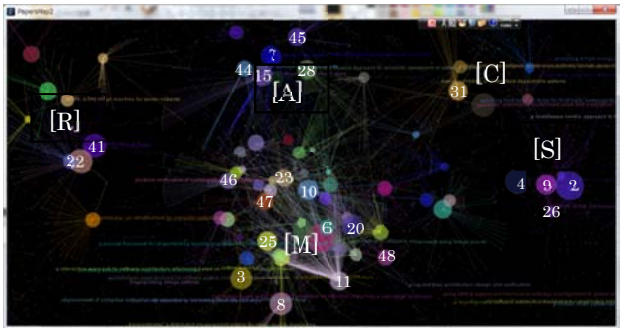

図5　51論文をHiAcで抽出したイメージ

表1　立堀らが抽出した51論文とSCOPUSで抽出した19論文との対比表

No	Conf	Year	Author	Title	SCOPUS
1	WI	20	João Pedro Sousa,	Aura: An Architectural	×

	CSA	02	David Garlan	Framework for User Mobility in Ubiquitous Computing Environments	
2	ICSE	2002	Aldrich, J., Chambers, C., Notkin, D.	ArchJava: Connecting software architecture to implementation	○
3	ICSE	2000	Mehta, Nikunj R., Medvidovic, Nenad, Phadke, Sandeep	Towards a taxonomy of software connectors	○
4	ICSE	2003	Batory, D., Sarvela, J.N., Rauschmayer, A.	Scaling step-wise refinement	○
5	ICSE	1999	Nenad Medvidovic, David S. Rosenblum, Richard N. Taylor	A Language and Environment for Architecture-Based Software	×
6	ICSE	2002	Dashofy, E.M., Van Der Hoek, A., Taylor, R.N.	An infrastructure for the rapid development of XML-based architecture description languages	○
7	ICSE	1999	Kazman, Rick, Barbacci, Mario, Klein, Mark, Carriere, S.Jeromy, Woods, Steven G.	Experience with performing architecture tradeoff analysis	○
8	ICSE	1999	Bowman, Ivan T., Holt, Richard C., Brewster, Neil V.	Linux as a case study: Its extracted software architecture	○
9	ICSE	1999	Keller, Rudolf K., Schauer, Reinhard, Robitaille,	Pattern-based reverse-engineering of design components	○
10	UML	2000	Aler, R., Borrajo, D., Camacho, D., Sierra-Alonso, A.	Reconciling the needs of architectural description with object-modeling notations	○
11	WICSA	1999	Kruchten, P., Selic, B., Kozaczynski, W.	Describing software architecture with UML	○
12	ECOOP	2002	Jonathan Aldrich, Craig Chambers	Architectural Reasoning in ArchJava	×
13	WICSA	2001	Eric M. Dashofy,André van der Hoek,Richard N. Taylor	A Highly-Extensible XML-Based Architecture	×
14	ICSE	1999	Jean-Marc DeBaud,Klaus Schmid	A Systematic Approach to Derive the Scope	×
15	ICSE	1999	Bosch, Jan	Product-line architectures in industry: A case study	○
16	ICSE	2003	Anita Sarma, Zahra Noroozi, and André van der Hoek	Palantír: Raising Awareness among Configuration Management Workspaces	×
17	WICSA	2002	Shang-Wen Cheng, David Garlan, Bradley Schmerl,	Using Architectural Style as a Basis for System Self-repair	×
18	UML	1999	Fiadeiro, J.L., Andrade, L.F.	Interconnecting objects via contracts	×
19	PFE	2001	Jan Bosch, Gert Florijn, Danny Greefhorst, Juha Kuusela,	Variability Issues in Software Product Line	×
20	ICSE	1999	Dashofy, Eric M., Medvidovic, Nenad, Taylor, Richard N.	Using off-the-shelf middleware to implement connectors in distributed software architectures	○
21	SPLC	2000	Martin L. Griss	Implementing Product-Line Features By Composing Component Aspects	×
22	FSE	2001	Uchitel, S., Kramer, J., Magee, J.	Detecting implied scenarios in message sequence chart specifications	○
23	ICSE	2000	Fielding, Roy T., Taylor, Richard N.	Principled design of the modern web architecture	○
24	WICSA	1999	Nenad Medvidovic and David S. Rosenblum	Assessing the Suitability of a Standard Design Method for Modeling Software Architectures	×
25	ICSE	1999	Di Nitto, Elisabetta, Rosenblum, David	Exploiting ADLs to specify architectural styles induced by middleware infrastructures	○
26	OOPSLA	2001	Riehle, D., Fraleigh, S., Bucka-Lassen, D., Omorogbe, N.	The architecture of a UML virtual machine	○
27	WICSA	1999	Mark H. Klein, Rick Kazman, Len Bass, Jeromy Carriere, Mario Barbacci	Attribute-Based Architecture Styles	×
28	WICSA	1999	Jeff Magee,Jeff Kramer,Dimitra Giannakopoulou	Analyzing the behaviour of distributed software architectures: A case study	○
29	ECOOP	2003	Aldrich, J., Sazawal, V., Chambers, C., Notkin, D.	Language Support for Connector Abstractions	×
30	GPCE	2002	Sandeep Neema,Ted Bapty,Jeff Gray,Aniruddha S. Gokhale	Generators for Synthesis of QoS Adaptation in Distributed Real-Time Embedded Systems	×
31	ICSE	2001	Kazman, R., Asundi, J., Klein, M.	Quantifying the costs and benefits of architectural decisions	○
32	WICSA	2001	Bridget Spitznagel, and David Garlan	A Compositional Approach for Constructing Connectors	×
33	ECOOP	2000	Marcus Fontoura , Wolfgang Pree , Bernhard Rumpe	UML-F: A Modeling Language for Object-Oriented Frameworks	×
34	FSE	1999	Pascal Fradet, Daniel Le Métayer and Michaël Périn	Consistency Checking for Multiple View Software Architectures	×
35	RE	2002	Jon G. Hall Michael Jackson Robin C. Laney Bashar Nuseibeh	Relating Software Requirements and Architectures using Problem Frames	×
36	SPLC	2002	Jan Bosch	Maturity and Evolution in Software Product Lines: Approaches, Artefacts and Organization	×
37	RE	1999	John Grundy	Aspect-oriented Requirements Engineering for Component-based Software	×

				Systems	
38	FSE	2001	Nima Kaveh, Wolfgang Emmerich	Deadlock detection in distribution object systems	×
39	FSE	1999	Michel Wermelinger, José Luiz Fiadeiro	Algebraic software architecture reconfiguration	×
40	ICSE	2003	Spitznagel, B., Garlan, D.	A compositional formalization of connector wrappers	○
41	FSE	2004	Uchitel, S., Chatley, R., Kramer, J., Magee, J.	System architecture: The context for scenario-based model synthesis	×
42	FSE	2004	Zhang, X., Young, M., Lasseter, J.H.E.F.	Refining code-design mapping with flow analysis	×
43	ICSE	2004	Hasselbring, W., Reussner, R., Jaekel, H..	The Dublo architecture pattern for smooth migration of business information systems: An experience report	×
44	ICSE	2004	Matinlassi, M.	Comparison of software product line architecture design methods: COPA, FAST, FORM, KobrA and QADA	○
45	ICSE	2004	Caporuscio, M., Inverardi, P., Pelliccione, P.	Compositional verification of middleware-based software architecture descriptions	○
46	ICSE	2004	Grechanik, M., Batory, D., Perry, D.E.	Design of large-scale polylingual systems	○
47	ICSE	2004	François, A.R.J.	A hybrid architectural style for distributed parallel processing of generic data streams	○
48	ICSE	2004	Khare, R., Taylor, R.N	Extending the REpresentational State Transfer (REST) architectural style for decentralized systems	○
49	ICSE	2004	Hong Yan, David Garlan, Bradley Schmerl,	DiscoTect: A System for Discovering Architectures from Running Systems	×
50	ICSE	2004	Bas van der Raadt,Jasper Soetendal,Michiel Perdeck,Hans van Vliet	Polyphony in Architecture	×
51	ICSE	2004	Ian Gorton, Jereme Haack	Architecting in the Face of Uncertainty: An Experience Report	×

5 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，立堀らが2004年に発表したソフトウエア・アーキテクチャの研究分野の動向調査報告と比較検証することにより，専門家が手動で抽出した主要論文をどこまで自動で抽出できるかについて検証した．その結果，対象論文はすべて本アルゴリズムにおいても主要な論文として抽出できていた．つまり，専門家が抽出した論文を本アルゴリズムによって分析することにより，主要な論文として自動で抽出することができた．今後は，本アルゴリズムで自動抽出された論文のさらなる分析を行い，専門家が抽出した論文との比較検証をしていきたいと考える．

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

[4] 立堀道昭，丸山宏，小林真，Daniel Yellin，吉田尚志，川井奈央，：ソフトウエア・アーキテクチャ研究動向の調査報告概要，情報処理学会研究報告，2005