氏　　名	Chang Lin-Hsuan
学位(専攻分野)	博士(統計科学)
学 位 記 番 号	総研大甲第 2153 号
学位授与の日付	2020 年 3 月 24 日
学位授与の要件	複合科学研究科　統計科学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Statistical approaches on the analyses and interpretations of a scientometric database
論 文 審 査 委 員	主　査　准教授　南　和宏
	教授　栗木　哲
	教授　金藤　浩司
	教授　中谷　朋昭 横浜市立大学 学術院国際総合科学群データサイエンス学部
	教授　中野　純司 中央大学　国際経営学部国際経営学科

（様式３）

博士論文の要旨

氏　　名 Chang Lin-Hsuan

論文題目 Statistical approaches on the analyses and interpretations of a scientometric database

Nowadays, the research funds are very limited, and all research institutes and universities are competing against one another for these resources. Government and private funding agencies, on the other hands, make their decisions on distributing these limited resources mainly based on the applicants' research performances. Therefore, it is very important to have a fair method to judge the research performance, so that the resources are distributed in a reasonable and efficient way. As a result, bibiliometics, which is an analysis of research performance based on scientometric database, becomes a hot topic that governments and research institutes advocate and emphasize recently.

Among all bibiliometics, citation plays an important role in evaluating research performance. In the regime of big data, citation information is gathered from large-scale databases and represented in network forms. A citation network analysis is a quantitative method to identify important and impacted literature of a field on the basis of how often a publication is cited in other publications. This analysis has recently become an essential tool to evaluate scientific achievements at different entities, including but not limited to, research articles, individual researchers, scientific journals, international conferences, universities and research institutes, or even countries. Governments and funding agencies make appropriate decisions on the allocation of their resources to these entities according to their research performances. For example, an education department can allocate its educational funds to schools based on quantitative reports of their performances.

The main theme of this thesis is to study how to evaluate the research performance of an entity in a fair way using a large-scale citation network called the Web of Science. In specific, the phenomenon of cross-disciplinary citation is studied first. The brief information of Web of Science is introduced and we analyze the citations within each academic field and citations from other fields toward statistics articles. Some fields tend to use a lot of article as reference and some fields are not. We are also interested in how other fields use statistical article as reference since statistics is a useful tool which can be applied in different areas. We can also found out some overlap between each fields.

Then the article network influence (ANI) method is introduced to evaluate the research performance of an entity, such as research articles, and scientific journals.

Most conventional methods tend to consider only the "directed citation" and ignore the "indirect citation". We think it is not enough to express how influential an article is. Assume we have a small citation network with three articles and two citations, where Article A was cited by Article B and Article B was cited by Article C. If we measure the influence of paper A in this network, conventional methods only take the direct citation into consideration and conclude that Article A has influences only on Article B. However, it is very likely that Article A inspired Article C through Article B. This scenario might happen when one of the above two citations (or both) was interdisciplinary, or Article B was a review-type article. We should consider those scenarios and put all the citations into consideration. The comparison between ANI and other methods are also provided in the thesis. We also developed a journal based network influence (JNI) based on the ANI.

Lastly, we analyze the structure of the article citation network of a particular subject and propose a generative model on how this citation network is evolved. It is natural that different subjects have different characteristics in their citation network. For some theoretical-based subjects, it takes a certain amount of time for readers to digest the theorems in the articles, so their resulting citations will take some years to appear. On the other hand, for some application-based subjects, the results from their articles should be implemented in the shortest time in order to have the best advances, so these articles will receive immediate references after they are published. In order to understand the characteristics of citation networks of different subjects, it is essential to understand their structures and their underlying generative models. Then we might be able to find out the difference between different subjects. We assume that the network is evolved based on the importance of the article. The more important the article is, the more citations it will received. Tapered pareto distribution is used to describe the distribution of importance of the article. We treat the importance of the article as the connection probability of receiving the citation. Instead of studying the whole Web of Science, we particularly study a subset under the topic of statistics.

Form8・Separate Sheet
（様式８・別紙）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　Chang Lin-Hsuan

Title
論文題目

Statistical approaches on the analyses and interpretations of a scientometric database

[論文の概要]

提出された論文は全 5 章 87 頁からなり，英語で書かれている．第 1 章は Introduction である．近年，政策当局のみならず研究機関あるいは個々の研究者といったさまざまな立場において，分野内，分野間での業績の客観的な評価の重要性が高まっている．また一方で最近は学術論文データベースが整備され，それを利用することにより，分野内のみならず分野を跨ぐ論文の引用被引用関係を客観的に評価することが可能となっている．本論文は，学術論文データベースを利用して，異分野の論文，学術雑誌を比較することを可能とする指標の開発とそれを用いた実証分析を行うことを目的としている．まず本章では，論文評価のための客観的な指標の重要性が述べられたあと，引用被引用関係は有向グラフで表現されることに着目し，いわゆるグラフ解析において提案されている複数の中心性指標が概観されている．

続く第 2 章では論文全体にかかる予備的な実証分析として全科学分野の論文引用状況が比較される．考察される科学分野は学術論文データベースの一つである Web of Science で与えられている 266 分野である．各分野における平均被引用率と引用論文率が計算され比較されている．これにより各分野の論文引用状況が明らかになった．次に統計学に重きをおいた解析として，各分野からの統計学論文の引用状況を示すために，統計学論文の引用率と平均引用率が計算されている．これらにより各分野における統計学論文の利用状況が明らかとなる．

本論文の主要部である第 3 章では，論文の影響度を測る新しい指標が提案されている．この指標では，直接の引用だけではなく，孫引きなど多段階にわたる間接的な引用が考慮される．各論文における多段階の論文被引用数が，多段階の平均的な引用数の逆数で重み付けされて指標を構成する．その際，重みの非増加性やパスの多重性などが考慮されている．その指標を Article Network Influence (ANI) と名付け，それにより各論文の影響力を測定することが提案されている．また，各学術雑誌に掲載された全論文の ANI のメディアンを Journal Network Influence (JNI) と名付け，それを雑誌比較のための指標として利用することが提案されている．さらにこの指標を用いて統計学分野に注目した実証分析を行っている．30 年間のデータを用いて，各年の 10 年後までの引用被引用ネットワークに対して ANI を計算し，その上位論文を調べることによって，いくつかの知見を得ている．また ANI とページランクおよび FWCI (Field-Weighted Citation Impact) との比較

が行われている．

第 4 章では引用被引用論文ネットワークの確率モデルが検討されている．ここで対象とするデータは 30 年分の統計学論文である．この引用被引用論文ネットワークにはいくつかの特徴がある．まず，論文数が年とともに増加していてほぼＳ字カーブを描いている．また，ある論文の引用数は最初の数年は増加し，その後減少する．そして，各年の論文における引用数の分布は右に裾を引く，いわゆるスケールフリー性を持つ分布に近い．このようなデータの特徴を生かす形で，年ごとの論文数，数年後の引用可能性，各年の論文の影響力のそれぞれにパラメットリックモデルを想定し，それらを統合したモデルを提案した．さらに推定されたパラメータの妥当性，モデルの適合度を，そのモデルからシミュレートしたグラフが，元のデータのグラフと近いかどうかで判断した．有向グラフ間の類似度は直接引用の度数分布間の Kullback-Leibler ダイバージェンスを用いている．対象データに当てはめたモデルの適合度には経年変化がみられ，これは論文の引用構造自体が経年変化することを示唆している．

最後の第 5 章はまとめと今後の研究課題にあてられている．

[論文の評価]

本論文の主な貢献は以下の 2 点である．第一は，第 2 章において全ての分野の論文に対して，予備的ではあるが引用被引用関係の状況を調べあげたことである．このような実証研究は少なく，分野間での競争が重要視される現在，基礎資料としての価値は高い．特に全ての分野における統計学論文の被引用数を調べたことは，統計学分野への貢献として大きいと考えられる．第二は，第 3 章において論文の影響力に対して新しい指標 ANI を提案したことである．本指標は直観的で意味が分かりやすいこと，直接引用数による従来の指標の自然な一般化である点が評価できる．さらに統計学分野に関する ANI を用いた実証研究として，一定期間において影響力の強い論文を特定し，それを基に研究の変遷に関して知見を得たことも評価できる．第 4 章の引用被引用グラフに関する確率モデルの提案も新しい試みである．

また予備審査において指摘されていた 3 つの事項（利用したデータベースである Web of Science の限界に関する記述，第 2 章のまとめの追加，本研究の限界と今後の研究に関する議論）も適切に対応されている．

以上の議論を踏まえ，審査委員会は，本論文は学位の授与に値すると全員一致で判断した．なお，第 2 章の内容は「統計数理」に採択済み，第 3 章の内容は IEEE Access に掲載済みである．