氏　　名　Joomi JUN

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2413 号

学位授与の日付　2023 年 3 月 24 日

学位授与の要件　複合科学研究科　情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　An ethnic classification using machine learning and surname data and its applications to social science

論 文 審 査 委 員　主　査　水野　貴之
情報学専攻　准教授
武田　英明
情報学専攻　教授
岡田　仁志
情報学専攻　准教授
宮尾　祐介
東京大学　大学院情報理工学系研究科　教授
大西　立顕
立教大学　大学院人工知能科学研究科　教授

(Form 3)

Summary of Doctoral Thesis

Name in full Joomi JUN

Title An ethnic classification using machine learning and surname data and its applications to social science

People leave their birthplace and settle in another land for political, religious, or economic reasons. Thus, people from different origins and political and cultural backgrounds mix, making society more diverse. To understand society, one needs to first understand the difference and characteristics of members of society. In this study, I focus on ethnicity to understand social phenomena. Ethnicity combines physical, cultural, linguistic, and other characteristics. Its definition and level of subdivision differed by country. Therefore, quantitative analysis was difficult. To address this limitation, this study uses surnames to identify individual's ethnicity and applies research on social phenomena.

I develop a classifier that could verify the origin of people using business people's surname from ORBIS dataset and Recurrent Neural Networks. Surname-origin classifier identifies individuals' origin from his/her surname. This classifier can identify 77 origins. A test using Olympic data yielded an average accuracy of 66.8%. In addition, compared to the human judgment result obtained through MTurk, origin was classified similar level of local people.

Using surname-origin classifier, I analyze the ethnic diversity in sociology and the ethnic homophily in economics. For investigating the ethnic diversity in sociology, I first analyzed the ethnic diversity of each country with calculating the entropy of the origin composition. Through this, it is possible to compare ethnic diversity between countries with an overwhelmingly high proportion of a dominant ethnic group, such as Japan, and countries like Canada, which do not.

Next, I focus on the social groups in the United States and showed how the racial composition differed by group. Based on the racial proportions of US demographics, it is confirmed that Asians occupy a relatively high proportion in the economic and physician groups, and non-Hispanic whites occupy a higher proportion in public figures than statistics.

Finally, I analyze the spatial distribution of origin in African continent and Europe countries. The racial composition of all subjects of analysis is complex, and they are in an environment where it is difficult to analyze them statistically. In this study, origins are classified by surname, and regions with similar origin distributions are grouped by Jensen-Shannon Divergence. In the case of the African continent, it was divided into four groups. This yielded similar results to subdividing the African continent by

language and religion.

For analyzing the ethnic homophily in economics, I focus on the international trade. First, I analyzed the frequency of transactions between ethnic groups focusing on companies in the United States. Considering the majority ethnic group of each company and using the conditional probabilities to visualize the frequency of transactions between ethnic groups, I found strong ethnic homophily, especially in the Asian and Middle East ethnic groups.

Next, using the gravity model to measure the statistical significance of ethnic factors in international trade, I analyzed the global trade using the WTFC dataset, to find a positive ethnic factor effect, although this differed by country. Although there was no significant ethnic factor effect in the English (GB) and European (DE) groups, positive effects could be observed in the Chinese (CN) and Korean (KR) groups. I also found significant ethnic factor effects even after removing the language barriers. A preference could be found between identical ethnic groups not in domestic, but in international trade.

Finally, I introduce case study which interethnic trade is activated when an obstacle occurs in international trade. For this, I used the US trading data obtained from FactSet and the PIERS bill of lading data obtained from IHS, and analyzed how ethnic networks in international trade are activated around specific events such as the U.S.–China trade war and the Arab Spring. This study assumed that ethnic linkage can be used to overcome the existing difficulties in international trade. I could thus conclude that exports from China to the United States decreased slightly compared to the average during the 2018 US–China trade war. However, when a Chinese executive was employed in the company of the consignee during the same period, the company's imports from China increased compared to the average. Thus, ethnic networks can be activated to overcome inter-country trade disturbances. During the Arab Spring period, Arabian's interethnic trade was also observed to be active.

This study is meaningful in that it examines the social science problems of ethnic composition using massive data and machine learning techniques adopting the informatics approach. The study is crucial in that it addresses the ethnic diversity and homophily issues that affect social phenomena.

This study can be extended to other fields, for example, to resolve ethnic conflicts and health care issues by ethnic characteristics. It can also be applied to cryptocurrency market, which is expected to be active in interethnic transactions because language dependence becomes stronger to obtain information. I leave these tasks to future works.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　Joomi JUN

Title
論文題目　　An ethnic classification using machine learning and surname data and its applications to social science

本学位論文は、「An ethnic classification using machine learning and surname data and its applications to social science」と題し、全８章から構成されている。
第１章「Introduction」では、社会現象に影響を与える民族（Ethnicity）や出自（Origin）の多様性と分断を、名字と国籍（Nationality）のビッグデータと機械学習の技術により、実証的に分析するという新たな研究手法を導入することの重要性が述べられている。
第２章「Related work」では、健康科学や生物医学、社会科学、統計学、情報学分野における名字を用いた民族に関する関連研究について述べられている。
第３章「Dataset」では、複数の英字での名字と国籍のビッグデータ、商取引のビッグデータについて、それらの内容と本研究における使用目的について述べられている。
第４章「Surname-Origin Classifier」では、全世界３億人の企業関係者の名字と国籍とで学習された Recurrent Neural Network（RNN）が名字から予測する国籍は、人間が名字から判断する出自と強く相関することが示され、この RNN は名字-出自分類器として利用が可能であることが述べられている。
第５章「Application to Sociology」では、第４章で構築された名字-出自分類器を用いて、名字に由来する国家や社会集団、地域の多様性を実証的に調査する社会学への応用が述べられている。アフリカ大陸の各地域を結ぶ出自ネットワークが実証的に可視化されている。
第６章「Application to Economics」では、名字-出自分類器の分類結果を用いて出自をクラスタリングしたものが、使用言語の類似度で定義された民族分類と近いことを利用して、名字-民族分類器を構築し、この分類器を用いて商取引に与える民族性の影響を実証的に調査する経済学への応用が述べられている。貿易では同一言語圏でさえ取引量が取引相手との民族の類似性に依存する一方で、国内の企業間取引では、そのような特徴が統計的に有意ではないことが示されている。
第７章「Ethical risks to consider」では、民族分類の倫理的側面と、民族分類研究を進める上で研究者が注意しなければいけない留意点について述べられている。
第８章「Conclusion」では、これまでの章のまとめと民族間取引の活発化が期待される仮想通貨市場への応用という将来の展望が述べられている。
公開発表会では博士論文の章立てに従って発表が行われ、その後に行われた論文審査会及び口述試験では、審査員からの質疑に対して適切に回答がなされた。質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究が、見えざる民族のネットワークを情報学の手法を応用して実証的に見せる新しい視点を研究分野に導入したことが評価された。

以上を要するに本学位論文は、名字-出自分類器の作成方法とこの分類器を社会科学に応用する具体的な研究事例を示したものであり、研究分野の発展に貢献しているという点で学術的価値が大きい。また、本学位論文の成果は、学術雑誌論文２件、ショートペーパー査読付き国際会議論文１件として発表され、社会的な評価も得ている。以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。