別記様式第３号－１

学 位 論 文 の 要 旨

氏　　名	久永　忠範
学位論文題目	オープンデータの連携に関する研究

ICTの急速な普及によりデータ連携が容易になり，近年ビッグデータやオープンデータの活用が推進され,国や地方公共団体をはじめ多く団体,企業がオープンデータの公開,活用に取り組んでいる．2012年に策定された「電子行政オープンデータ戦略」においては,情報の連携，共有の必要性が明示されている．特に地方公共団体のもつオープンデータの利活用は，地域活性化の大きな要因にもなりえる．

本論文は,地方公共団体のもつオープンデータ間の連携度を述語ベクトル法という手法で測り，データの連携についての研究をまとめたものである．

第１章は，オープンデータの背景と状況の説明を，政府の施策や取組の状況と共に説明した．またデータ連携を行うためには，機械判読しやすいデータ形式でなければならないことと，あまり高度な知識が必要なデータ形式は，地方公共団体職員がそのデータを作ることは容易ではないことを述べた．そのために政府は，地方公共団体の職員向けにデータ作成のための研修やe-learningサイトも開設して，オープンデータの推進を行っていることも述べた．

第２章は，オープンデータに関する関連研究について述べた．政府の取り組むデータ連携のための共通語彙基盤の運用がどのように行われているか説明した.またオープンデータを活用または推進するために,自治体職員の負担減やオープンデータ運用のための費用対効果を考えた受容性向上を目的とした研究や,防災情報を速やかに配信するためのデータ連携の共有についても述べた．そして機械判読によるデータの連携を行うためには,共通のキーとなる語彙が必要であることを述べた．

別記様式第３号－２

第３章は,本研究に必要なオープンデータ間の連携度を測るための述語ベクトル法について述べた.実験に必要なオープンデータの収集方法やそれらのデータの項目名と列データの関係,そして述語ベクトルを生成するための項目判定関数を用いた列間類似度について述べた.またオープンデータ間の連携度を計算する手法やデータの偏りをなくすための重み付けの手法，連携度の計算式，重み付けのない平均，正規分布，減衰関数，閾値を用いた4パターンのうち，どの計算式の連携度の正確性があるのかの考察をおこなった.

第４章は，述語ベクトル法で全国の地方公共団体からランダムに抽出した300のCSVファイルの連携度を列データ間類似度と項目名間類似度による2つのパターンで比較して正確性の実験を行った.列データ間類似度によるオープンデータ間の上位の連携度を考察すると項目名が違っても列データが似通って入れば高い連携度を示していることがわかった．項目名間類似度によるオープンデータ間の上位の連携度を考察すると，同じ自治体の同じような項目名のデータ同士や項目名が違っても一部に同じ単語が含まれていれば連携度が高いが,データの内容自体にあまり連携の可能性が低いことがこの実験でわかった.

第５章は，本研究のまとめと今後の課題について述べた．本研究は，オープンデータの連携度を測り,連携の可能性からお互いのオープンデータを機械判読によって活用することが目的である．その目的達成のために，地方公共団体のオープンデータの連携内容の詳細を調査した．連携を行うための地方公共団体が開示するCSVのデータ形式にはCSV形式として適さないものがあったが,データの内容を使用することで連携度の精度が向上した．また計算時間を短縮するためには，類似度を計算する項目判定関数の精査が必要であり,連携度の高いデータ同士を活用するアプリケーションの作成について述べ総括とした.

別記様式第４号

Summary of Doctoral Dissertation

Title of Doctoral Dissertation:

Study on the cooperation of open data

Name: **Hisanaga Tadanori**

The rapid spread of ICT has facilitated data linkage, and in recent years the use of big data and open data has been promoted, and many organizations and companies including national and local public organizations are working on the disclosure and use of open data. The “Electronic Administration Open Data Strategy” formulated in 2012 clearly shows the necessity of information linkage and sharing. In particular, the utilization of open data owned by local governments can be a major factor in regional revitalization.

Chapter 1 explained the background and situation of open data, together with the state of government measures and efforts. It also states that in order to perform data linkage, the data format must be easy to read by machines, and that data formats that require a high level of knowledge are not easy for local government staff to create the data. It was. To that end, the government also stated that it has promoted open data by opening training and e-learning sites for data creation for local government employees.

Chapter 2 described related research on open data. We explained how the common vocabulary infrastructure for data collaboration that the government is working on is used. In addition, in order to utilize or promote open data, research aimed at improving the acceptability considering the cost-effectiveness of local government staff and open data operation, and data linkage for promptly distributing disaster prevention information He also shared about sharing. He stated that a common key vocabulary is necessary to link data by machine interpretation.

Chapter 3 described the predicate vector method for measuring the degree of cooperation between open data required for this study. The method of collecting the open data necessary for the experiment, the relationship between the item names of these data and the column data, and the similarity between columns using the item decision function to generate the predicate vector were described. Which of the following four patterns uses a method for calculating the degree of linkage between open data, a weighting method for eliminating data bias, a formula for calculating the degree of linkage, an unweighted average, a normal distribution, an attenuation function, and a threshold. We examined whether the calculation formulas are accurate.

Chapter 4 is an experiment of accuracy by comparing the degree of cooperation of 300 CSV files randomly extracted from local governments nationwide by predicate vector method with two patterns of similarity between column data and similarity between item names. Was performed. Considering the higher level of linkage between open data based on the similarity between column data, it was found that even if the item names were different, if the column data entered in a similar manner, a high linkage was shown. Considering the high degree of cooperation between open data based on similarity between item names, the degree of cooperation is the same if the same word is included in some items even if the data of the same item name of the same local government or item names are different. Although it is expensive, it was found in this experiment that the possibility of cooperation with the data content itself is low.

Chapter 5 described the summary of this research and future issues. The purpose of this study is to measure the degree of cooperation of open data, and to utilize each other's open data by machine interpretation from the possibility of cooperation. In order to achieve that goal, we investigated the details of the contents of open data collaboration of local governments. Although some CSV data formats disclosed by local governments for collaboration were not suitable as CSV format, the accuracy of the collaboration was improved by using the data contents. Moreover, in order to shorten the calculation time, it is necessary to carefully examine the item judgment function that calculates the similarity, and the creation of an application that uses data with high cooperation is described and summarized.