氏　　名	NGUYEN Tri Phuc
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2163 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　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Semantic Tabular Data Annotation with Knowledge Bases for Data Interoperability
論 文 審 査 委 員	主　査　教授　武田　英明
	教授　相澤　彰子
	准教授　市瀬　龍太郎
	助教　小林　亮太
	准教授　大向　一輝 東京大学大学院　人文社会系研究科

(Form 3)

Summary of Doctoral Thesis

Name in full: NGUYEN Tri Phuc

Title: Semantic Tabular Data Annotation with Knowledge Bases for Data Interoperability

Tabular data is semi-structured data, widely used for many aspects of human life such as reports, notes, books, media, software, and many other places. It is an effective and efficient way to store and represent data for human consumption since its compact representation reflects the logical relation between columns, rows, and cells. In the era of computers and the Internet, tabular data is the most popular structure for relational databases, Web tables, spreadsheets, and Open Data.

Nowadays, with the vision of Open Data, a large number of tabular data have been published on the Web and Open Data Portals. Such tabular data contains valuable information and could be potentially useful in various fields, such as health, food security, climate change, resource management, smart cities and so on. Additionally, our society has become data-driven, where more and more data expected to grow in the near future from large volume, variety, and velocity. As a result, it is promising for establishing transparency, improving the quality of human life, and inspiring business opportunities.

Although these tabular data offer huge potential, these data are difficult to use due to fragmentation, heterogeneous schema, missing or incomplete metadata. Therefore, the usability of tabular data is an open question and should be exploited. There are several works have been made on improving the usability of tabular data such as establishing standard policies for data providers or performing automatic reconstruct semantic meaning for tabular data. The first solution on standard policies takes a lot of time, and effort and difficult to scale, while the second solution is more promising to automation, and scale-up.

This thesis focuses on the automatic reconstruct semantic meaning for tabular data. The methodology is to assign the elements of tabular data into semantic concepts in knowledge bases. As a result, the meaning of tabular data could be interpreted or inferred by knowledge base concepts, therefore, it is easy to use in other downstream applications.

In this thesis, we firstly review the table data annotation for data interoperability including matching tasks, challenges, possible applications. Additionally, we identify potential limitations of tabular data annotation: 1) common text-based approaches are less effective in annotating numerical attributes; 2) entity lookup on one search engine is imperfect on the general and multi-language text. Then, we introduce the novel

solutions to address these limitations of tabular data annotation. We introduce Distribution-based Similarities (DBS), and a deep similarity metric (EmbNum+) for numerical attribute annotation to address the first limitations, and MTab is a general framework for tabular data annotation which addresses the limitations 1 and 2.

The following describes the details of these methods.

First, we present a lightweight solution on semantic annotation called DBS for numerical attributes. Existing approaches rely on the p value of a statistical hypothesis test as a metric to estimate the similarity between numerical attributes, and then assign unknown attribute by the labeled attributes. However, the p value-based metrics strongly depend on the assumptions about the distribution and data domain. In other words, they are unstable for general cases, when such knowledge is undefined. We present effective metrics called Distribution-based Similarities (DBS) to address the limitations of p value-based metrics.

Second, we present an effective and efficient method called EmbNum+ which is an end-to-end system to learn a similarity metric directly from numerical attributes. EmbNum+ was inspired by deep metric learning approaches with which both representations and a similarity metric are learned without making any assumption regarding data; hence, enabling EmbNum+ to be more generalized with a variety of data types and distributions. Evaluations on many datasets of various domains show that EmbNum+ consistently outperformed other approaches in terms of effectiveness and efficiency.

Third, we present a general framework for tabular annotation called MTab. MTab combines the voting algorithm and the probability models to tackle bottleneck problems of tabular data annotation. Additionally, we also adopt more signals from table elements and introduce a novel scoring function to estimate the uncertainly from ranking. This system got the first prize for entity annotations (CEA), type annotations (CTA), and relation annotations (CPA) at the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, the 18th International Semantic Web Conference 2019.

Overall, the objective of this thesis is to improve semantic annotations for tabular data (MTab), in particular, the treatment of numerical attributes is the main focus (DBS, and EmbNum+).

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 NGUYEN Tri Phuc

Title
論文題目 Semantic Tabular Data Annotation with Knowledge Bases for Data Interoperability

出願者はデータ相互運用性を向上することを目的として、表形式データに意味的なアノテーションを行うアルゴリズムの提案とそれを用いたシステム構築に関する研究を行っている。Web 上のデータやオープンデータなど多種多様な表形式のデータがあるが、多様であるが故に検索や統合といったデータとしての活用が難しい。本論文ではこのような表形式のデータに対する意味同定を知識ベースに対するアノテーションによって表現することの研究である。本論文では特に数値データからなる表における意味同定を中心課題として研究を進めている。

1 章では問題の背景と課題を明らかにし、2 章では表形式データの処理について問題の切り分けと関連研究のサーベイを行なっている。Web 上の表形式データは Web の社会での活用やオープンデータの進展によって増加している一方、その多様性が故に処理が難しい。そこでまず表形式データの処理における課題を分別した。大きく構造問題と意味問題に切り分け、前者には表タイプ、ヘッダ、データのタイプ、キー属性などの同定が含まれ、後者にはエンティティ、タイプ、属性の意味同定が含まれる。本論文では、このうち、まず数値データに対する属性の意味同定について取り組み、そのあとこの数値の属性意味同定を含む意味的な同定全体に取り組んでいる。

3 章ではまず数値集合データ間の類似性をどう定義するかについての探求を行い、数値集合の分布への変換による類似性の比較という方法を提案している。この方法を既存のテスト用データセットとオープンデータから生成したテストデータセットで検証し、既存の研究を上回る結果を出すことを示した。なお、後者のデータセットは出願者が本論文のために作ったもので、ヨーロッパ等の政府のオープンデータサイトにある実データから作ったテスト用データセットで、他にないユニークなデータセットである。

4 章では数値集合データの類似性を深層学習アルゴリズムを取り入れた方法で判定し、意味的なアノテーションを生成する一連の方法(EmbNum+)を提案し、既存手法より優れた結果を出している。3 章の方法における類似性の比較を深層学習における表現学習を利用することで不均質な分布のデータで利用可能にしている。ここでは数値集合データの特徴を使い Inverse Transform Sampling により教師事例を増加させることで深層学習を適用可能にしている。また、関連性学習を導入することで、データのスパース性からくる不必要なアノテーションを抑えることを行なっている。この結果、比較的よくデータが整備されている既存の小規模のテストデータセットだけでなく、多様なデータが含まれている大規模なデータセットやノイズが多い実データから生成したオープンデータ・テストデータセットでも既存研究を上回る結果を出している。

5 章では前章のアルゴリズムを一部として文字数値混在の表形式データを知識グラフへのアノテーションするシステムを開発している。このシステムは 2 章で述べた意味同定を全体としてサポートしており、表の項目列のアノテーション（ある項目列がどのようなタイプになるかの同定問題)、エンティティのアノテーション（個別のデータの値が何を指し示しているかという同定問題)、項目列間の関係のアノテーション（表の項目列が何の属性を指し示しているかの同定問題）の 3 つの同定問題を解くことができる。ISWC2019(18th International Semantic Web Conference)の併設の Semantic Challenge の課題として、「表型データのナレッジグラフへのマッチング(Semantic Web Challenge on Tabular Data to Knowledge Graph Matching)」が設定され、上記の 3 つの同定問題が課題となった。出願者のシステムは、すべての課題のすべてのラウンド（計 12 個のコンテスト）において、第 1 位であった。

6 章でまとめと展望を述べている。

以上のように出願者の博士論文は、表形式データの意味同定に関して新たな有効な処理手法を考案しその効果を検証しており、当該分野における研究の発展に貢献するものと認められる。なお、本論文の内容は 1 編の査読付き論文、1 編の査読付き国際会議論文として公表されており、分野における評価もなされている。

以上のような学術的貢献を総合的に判断して、本論文は博士学位を与えるに十分な水準に達していると、審査委員全員一致で認められた。