氏　　名	Khai Hoang NGUYEN
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1888 号
学位授与の日付	平成28年9月28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	Instance Matching for Large Scale Heterogeneous Repositories
論 文 審 査 委 員	主　査　准教授　市瀬 龍太郎 教授　武田 英明 教授　佐藤 健 教授　相澤 彰子 国立情報学研究所 准教授　福田 直樹 静岡大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

The recent years have witnessed the fastest period of the development of digital information. Individuals and organizations are publishing data at the highest acceleration. Accompanying with the immense amount of data are many challenges. Among them, instance matching, which identifies different instances of the same entity (aka. coreferences) in various data sources, has been considered as a critical problem. The reason is that the independently created instances are usually incomplete, because of the inconsistency nature of data publication (e.g., purpose, tool, user, and scale). Instance matching helps not only to collect the multiple aspects of entities but also to improve the consistency and non-redundancy of the data.

The dissertation summarizes our contributions to several issues of instance matching. First, we focus on scalability, which is very important for deploying large matching tasks. We develop a time and memory efficient framework named ScSLINT. ScSLINT is a specification-based framework. It generates coreferences on the basis of given instructions, such as matching properties, similarity metrics, and filtering strategy. ScSLINT promotes the matching task to at least 10 times faster compared to state-of-the-art frameworks. ScSLINT is also the unique framework successfully tested on quadrillion scale dataset using a memory-limited machine. Then, based on the architecture of ScSLINT, further systems and algorithms have been introduced.

We propose systems and algorithms for two scenarios of instance matching: supervised and non-supervised. These scenarios are different in the presence of training data. For supervised matching, we propose a specification-based system and a feature to enhance classification-based systems.

For non-supervised instance matching, we propose ASL. ASL is a schema-independent instance matching system for linked data. Because of the inconsistency in the schemas of different repositories, it is important to develop a general system that can work on any repository with any schema. ASL finds the equivalent properties and constructs the matching specification automatically. Experiments on 246 datasets with different schemas and domains show that ASL obtains high accuracy and significantly improves the quality of discovered coreferences against the previous systems.

For supervised instance matching, we propose ScLink system and R2M ranking feature. ScLink is a system for the specification-based matching. The most important part of a specification-based system is the construction of the specification. Existing specification learning algorithms are either ineffective or inefficient. Furthermore, there is space for improvement of scalability as previous systems have not optimized the candidate generation step. ScLink is the combination of two novel algorithms cLearn and minBlock. cLearn finds the optimal matching specifications by detecting high-quality equivalent properties and optimizing the similarity metrics. minBlock enhances the blocking step in order to generate less the candidates but retain as many correct candidates as possible. This is very important because it affects both the scalability and accuracy. This algorithm restricts the matching

task into a compact subset instead of the huge pairwise alignments between the input repositories. In addition, ScLink employs a novel string similarity metric namely mBM25, which aims at the better disambiguation against the existing metrics. We evaluate ScLink using 15 standard matching tasks on relational databases and linked data. The experiment results show that cLearn significantly increases the F1 score compared to the existing specification learning algorithms. Meanwhile, minBlock discards up to 95% of unnecessary candidates and therefore considerably contributes to the reduction of processing time. mBM25 also shows its usefulness on real datasets.

While the specification-based instance matching is good at scalability, the classification-based approach has the advantage of generalization based on the solid theory of machine learning. Therefore, we also approach the problem of instance matching in classification-based fashion. We find that the common limitation of current classification-based matching systems is the ignorance of ranking mechanism. The ranking is an important factor in instance matching because it contributes to the disambiguation, especially for large and ambiguous data. We propose the ranking feature R2M for the classification-based matching systems. R2M significantly improves the quality of the trained classifiers and advances them to significantly better performance. Compared to other systems, a classifier with R2M also outperforms ScLink as well as existing classification-based matching systems.

We also compare the performance of the proposed systems and the classifier with R2M ranking feature. We show that the usage of our systems and algorithms depends on the matching task and should be considered under the trade-off between accuracy and scalability.

(別紙様式 3)
(Separate Form 3)

博士論文の審査結果の要旨
Summary of the results of the doctoral thesis screening

博士論文では，様々なデータにおいて，同じものを指し示すインスタンスを同定するインスタンス・マッチング問題の研究に取り組んでいる．この問題に関して取り組んだ課題は３つある．１つ目は，巨大なデータに対応するためのスケーラビリティに関する課題である．２つ目は，異なるデータスキーマを持つ場合のデータの不均一性に関する課題である．３つ目は，同じ値でも同じものを指すとは限らないなどの問題を引き起こす，データの曖昧性に関する課題である．これらの課題に関して，教師付き，教師なしの２つの問題設定で取り組み，これを解決したというのがこの論文の主張である．

本論文は，全７章からなる．第１章「Introduction」では，インスタンス・マッチング問題に対して，研究の背景を述べ，本論文の貢献について説明している．

第２章「Background and related work」では，本論文に関する研究背景と関連研究について述べている．最初に，この論文で取り組む課題，インスタンス・マッチング問題とそれに関連する概念について説明し，その後，いくつかのカテゴリに分けて関連研究を説明している．

第３章「ScSLINT: Framework for large scale instance matching」では，スケーラビリティの課題に対応するためのフレームワークについて述べている．最初に，この研究の動機を説明し，次に，全体のワークフローを説明した後に，細かい提案手法の説明をしている．そして，実験的に提案手法の評価を行い，その有効性を確認している．

第４章「ASL: Schema-independent specification-based instance matching」では，教師無しの問題設定に対して，スケーラビリティ，不均一性の課題に対応するための手法について述べている．まず，この研究の動機，問題設定について述べ，次に，ある基準で取り出したプロパティのマッチングを取ることで，不均一性の課題を解決する手法などを説明している．そして，３つの実験を通して，提案手法の有効性を確認している．

第５章「ScLink: Supervised specification-based instance matching」では，教師付きの問題設定に対して，スケーラビリティ，不均一性，曖昧性の課題に対応するための手法について述べている．まず，この研究の動機，問題設定について述べ，次に，学習を使うことで，マッチングの判定に使う類似度などを自動的に設定し，曖昧性の課題を解決する手法などを説明している．そして，５つの実験を通して，提案手法の有効性を確認している．

第６章「R2M: Ranking features for classification-based instance matching」では，教師付きの問題設定に対し，不均一性，曖昧性の課題に対応するより精度の高い手法について述べている．まず，研究の動機，問題設定について述べ，次にこの手法に関連する概念について述べている．そして，類似度の順位を利用することで性能を向上させる提案手法について説明し，実験的に提案手法の有効性を確認している．

第７章「Conclusion」では，博士論文で提案した手法に関してそれぞれの関係を考慮しながら，総合的な考察をし，展望を述べると共に，結論をまとめている．

上記のように，本博士論文は，様々なデータにおいて，同じものを指し示すインスタンスを同定するインスタンス・マッチング問題における３つの課題を解決する手法を示した点で，この研究分野の発展に貢献するものである．また，この研究で示した考え方により，インスタンス・マッチングを精度高く実現することが可能となり，多様なデータを統合して運用するための基盤技術開発という観点からも意義があると認められる．さらに，博士

(別紙様式 3)
(Separate Form 3)

論文の内容は，１本の査読付きジャーナル論文，１本の査読付き国際会議論文，２本の査読付き国際会議ポスター・デモ論文などで発表されており，社会からも評価されている．

以上より，本論文は博士論文として，十分な水準であると審査委員全員一致で認められた．