氏　　名	RAHOMAN Md Mizanur
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1878 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	Keyword-based Information Retrieval over Enhanced Linked Data
論 文 審 査 委 員	主　査　准教授　市瀬　龍太郎 教授　山田　誠二 教授　武田　英明 准教授　宮尾　祐介 主席研究員　那須川　哲哉 日本アイ・ビー・エム株式会社

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

Linked Data are inference-enable, interlinked network like graph-based data. Usually these data are presented with a machine understandable way. The inclusion of schema information or ontology information of data makes Linked Data machine understandable, while the graph-based structure of data helps to find potential link identification among various data, and construction of data links. Over any dataset, data schema usually describes data about data i.e., meta data. In Linked Data perspective, this schema information is further extended to maintain relationship among the data or incorporate semantics over data which is called as data ontology. So, apart from describing meta data information, Linked Data ontology facilitates inference-enable data structure. Therefore, Linked Data are considered as knowledge with rich semantics.

Currently Linked Data hold vast amount of knowledge, which are also growing rapidly. Success of these contemporary data depends on how effectively they can be used by the users and applications. Like other data, usage of these data primarily relies upon how easily they can be accessed, and how good the data are i.e., the quality of data. A good number of researches have been conducted to investigate these two issues, however, contemporary systems are still not good to tackle them effectively.

The keyword-based query is an easy-to-use information access option for the users because of its familiarity and comfortability. However, traditional keyword-based query over Linked Data is not effective. While an information access over Linked Data system considers a query holistically, a traditional document-based information access system considers a query individually. That is why, an information access over Linked Data system needs to put all keywords of a query, in a semantic order, together so that it can retrieve more exact pitch of information that the query is searching for. While, a traditional document-based information access system retrieves relevant documents among the other documents, considering individual keywords of the query. In such a case, a document-based information access system might not need to consider the entire query, rather a part of the query can also produce required information need. So simple use of keywords will not fit over Linked Data. On the other hand, the information access over Linked Data is also different from the usual graph search because the traditional graph search may not be able to capture the rich semantics of Linked Data, that are presented with schema and ontologies. Therefore, we find that the contemporary systems are not effective. While some researches tried to adapt traditional keyword-based queries, and some tried to adapt subgraph searching, but those researches do not handle Linked Data's structural complexities. Particularly those researches do not give any specific

guide-line that how the query should be built. Rather they proposed ad hoc-based supervised query handling techniques or used language tools to retrieve the information. However, ad hoc techniques are not automatic and require expertise, while tool dependent techniques heavily rely upon performance of the particular tools. Therefore, the contemporary systems still not effective in Linked Data information access. Furthermore, there are also very few systems that can handle specific semantics like ``temporal semantics'' i.e., time and event related queries, however capturing them can leverage Linked Data usability. Since the contemporary systems suffer on handling those issues together, we propose Linked Data information access frameworks that are easy-to-use, effective to handle structural complexities, and facilitated to capture temporal semantics. We analyze structure of Linked Data and propose some defined templates for keywords, and their management to retrieve Linked Data information. While, we capture temporal semantics by text analysis.

On the other hand, usually Linked Data are generated from data sources ranging from manual generated to automatic generated data. In manual data generation, the domain experts of a particular domain e.g., biology domain, medical domain, law domain etc. craft this kind of Linked Data, which heavily depend upon manual activities. Because of manual intervention or expert intervention, this type of Linked Data is generally clean, but less frequent. On the other hand, in automatic Linked Data generation, pattern-based mapping, or rule-based mapping of source data extract Linked Data. This type of Linked Data is more common, but potential to hold less cleaner data than manually generated Linked Data. However, the existing Linked Datasets are not always clean. We understand that both type of Linked Data generation methods can generate unclean data. In manually generated Linked Data, the erroneous data entries could be generated because of human errors. Moreover, data could be generated from multiple sources which sometime differ from one another. On the other hand, in automatically generated Linked Data, erroneous data entries could be extracted because of the wrong contents in source data. However, we do not find many Linked Data quality assessment frameworks that can automatically identify all possible types of errors. Since manual quality assessment over Linked Data is not a feasible choice and data can hold different types of errors, an automatic quality assessment framework is a requirement. We adapt a novel unsupervised nearest-neighbor based outlier detection technique which is automatic, not susceptible to particular types of errors.

The proposed systems are easy to use and effective in their operations. They can support in leveraging Linked Data success.

(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，ユーザが Linked Data を効果的に利用するための課題を解決する研究に取り組んでいる．ここで取り組んだ課題は主に２つある．１つは，Linked Data を容易に検索するための手法に関する研究であり，キーワードに基づく検索手法を新たに提案している．Linked Data においては，通常の文書検索と異なり，キーワードの意味的な要素を勘案しながら検索しなければならないという課題がある．これを解決したというのが，この論文の１つ目の主張である．また，もう１つは，Linked Data におけるデータの質の向上に関する研究であり，データの誤りを検知する手法を新たに提案している．Linked Data においては，誤ったデータが混入することがあり，実用上，大きな問題となっている．データの誤りを検知する手法を新たに作ることで，この問題を解決したというのが，この論文のもう１つの主張である．

本論文は，全８章からなる．第１章「Introduction」では，Linked Data を効果的に利用することに関する本研究の背景，動機について説明している．

第２章「Background」では，この論文に関連する研究について述べている．まず，本論文で取り扱う Linked Data に関して説明し，その後，Linked Data における検索手法，質の評価手法の２つに分けて，関連研究と本研究の関係について，順に説明している．

第３章「BoTRet: The Basic Information Access Framework」では，この研究の基礎となるキーワードを使った Linked Data の検索手法について述べている．最初に，この研究の動機などを説明し，次に，キーワードが２つの場合，３つ以上の場合に分けて，提案手法を説明している．そして，３つの実験を通して，提案した手法の有効性を確認している．

第４章「TLDRet: A Temporal Extension of BoTLRet」では，第３章で述べたフレームワークを拡張し，時間に関する検索を実現する手法について述べている．最初に，この研究の動機などを説明し，次に，検索における時間やイベントに関する考察を行い，新たな検索手法について述べている．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５章「LiCord: A Machine Learning-based Keyword Segmenter」では，自然言語で作られた検索文から，検索に用いるキーワードを作成する手法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，機械学習を応用した言語非依存で検索文からキーワードを抽出する手法について述べている．そして，提案した手法を２つの実験で評価し，その有効性を確認している．

第６章「ALDErrD: An Information Assessment Framework over Linked Data」では，Linked Data における誤りを検知する手法について述べている．最初に，この研究の動機などを説明し，次に，基本的なアイデア，手法の詳細について説明をしている．そして，提案した手法を２つの実験で評価し，その有効性を確認している．

第７章「Discussion」では，博士論文で提案した手法に関して総合的な考察を行っている．

第８章「Conclusion」では，博士論文の結論をまとめている．

上記のように，本博士論文は，Linked Data を効果的に利用するために，キーワードを用いて容易に Linked Data を検索することができる手法を示すと共に，誤りを検知することで Linked Data を質的に向上させることができ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は，近年注目を浴びている知識

(別紙様式 3)
(Separate Form 3)
グラフ利用の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，博士論文の内容は，２本の査読付きジャーナル論文，４本の査読付き国際会議論文として発表されており，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．