様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　８　月　２２　日現在

機関番号：　８２６２６

研究種目：　若手研究(B)

研究期間：　2015～2016

課題番号：　１５Ｋ１５９９４

研究課題名（和文）Real-time, Best-effort Query Processing of Semantic Web data

研究課題名（英文）Real-time, Best-effort Query Processing of Semantic Web data

研究代表者

Lynden　Steven (Lynden, Steven)

国立研究開発法人産業技術総合研究所・人工知能研究センター・研究員

研究者番号：３０５２８２７９

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：分散RDF問合せ処理のリアルタイム検索において、固定時間制約の下でユーザ評価尺度が最適になるような解を求める新しい技術を開発した。RDFデータを含む文書を横断して固定時間間隔内において検索する上で、鮮度、多様性、または被覆率の3つの評価尺度を使うことが出来る。また、既存のリンクトオープンデータ知識ベースに対する構造化データの自動リンクのための技術を開発した。この技術では、SPARQLエンドポイントの挙動を予測する問題に対して、機械学習を適用することで、対象のエンドポイントに関するメタデータや統計情報などの事前情報を持つことなく、分散問合せ処理プランの最適化を支援することが出来る。

研究成果の概要（英文）：The project achieved the following contributions to Semantic Web query processing technology. Techniques were developed for optimising user criteria during live-exploration based distributed RDF query processing, which have resulted in a novel approach to answering queries within fixed time constraints. The approach enables criteria such as freshness, diversity or coverage to be increased during a fixed-time interval query over Web documents containing RDF data. Techniques were developed for automatic linking of structured data on the Web with existing Linked Open Data knowledge bases, and machine learning approaches were applied to the problem of predicting the behaviour of Semantic Web data providers (SPARQL endpoints) in order to support the optimisation of distributed query plans where no prior information (metadata or statistics) is available about individual endpoints.

研究分野： Computer Science

キーワード： Semantic Web　Linked Open Data　Query Processing

１．研究開始当初の背景

The vision of a Semantic Web is to convert the current World Wide Web, which is dominated by unstructured and semi-structured documents, into a global Web of Data by attaching semantics using common data formats, the most important of which is the Resource Description Framework (RDF). Linked Open Data (LOD) builds on RDF to enable the construction of a global network of interlinked data, with query languages such as SPARQL supporting specific queries as opposed to simple keyword searches. Whereas the vast majority of research into RDF query processing using SPARQL has concentrated on the use of answering queries completely (albeit usually using a limited set of data sources), the research carried out aims to develop more pragmatic techniques which answer queries within fixed time constraints while optimizing user criteria such as freshness, diversity and coverage.

２．研究の目的

The research aims to develop techniques for the optimization of query execution over Linked Open Data using criteria within a 3-dimensional space. This would allow query result attribute prioritization encompassing: (1) freshness (recently added or updated data), (2) finding diverse data (data from different parts of the Web of Linked data), and (3) coverage (the number of results obtained from queries). The technique should provide query results that focus on a user-defined mix of these attributes. The research also aims to support the automatic linking of structured data on the Web with existing Linked Open Data knowledge bases, in order to increase the potential coverage of Linked Data queries. The research also aims to apply machine learning approaches to the problem of predicting the behaviour of Semantic Web data providers (SPARQL endpoints) in order to support the optimisation of distributed query plans where no prior information (metadata or statistics) is available about individual endpoints.

３．研究の方法

In order to support the stated research objectives, algorithms for controlling live-exploration based query processing over distributed RDF on the Web were developed, and experiments performed to analyse performance over the data sources. We developed strategies for live-exploration link prioritization including a novel semantic connectivity-based strategy. The motivation for a semantic connectivity-based strategy is a consequence of the fact that similar resources, connected within RDF graphs, may contain similar properties (such as freshness, diversity, or their relevance in answering a query). Selecting IRIs to dereference that are semantically similar to documents that have improved results with respect to specific criteria (coverage, freshness, diversity) previously during a query's execution process may maximise such criteria during subsequent execution. Live exploration-based query processing approaches, which potentially access the entire Web of Data, often encounter many more data sources than can be investigated (dereferenced and matched with query triple patterns) than it is possible to investigate if the query must be answered within a reasonable time period. The effects of IRI selection strategies were investigated, where user criteria, which should be represented in query results, should be maximised. Sequential selection of IRIs was compared to techniques based on representational (string), and semantic similarity to optimise for the criteria of coverage, diversity and freshness. These results can form the basis of optimisation techniques for use in existing query processors, The proposed techniques of using either string distance-based or semantic connectivity-based techniques have been demonstrated, and given the ease with which they can be implemented can provide worthwhile improvements to query results in a best-effort query answering scenario where user criteria such as coverage, freshness or diversity are specified

We also investigated the hypothesis that analysing URLs may be a promising technique when determining, from a large set of possible URLs, which ones to investigate to retrieve relevant structured (semantic) data. Two complementary approaches were developed for eliminating or reducing the reliance of applications on centralised knowledge bases and third-party search services for retrieving structured data. Both approaches utilise machine learning techniques to analyse

URL content alone, exploiting the emergence of conventions such as Semantic URLs, to aid the discovery of structured data. Our study, limited to data obtained from DBpedia resources representing creative works such as movies and books, and associated Google search results, provides a proof-of-concept of the proposed approach.

We also developed an approach for learning to predict response times of SPARQL endpoints, an important problem during the execution of federated queries over multiple SPARQL endpoints. Key findings of our experimental evaluation include the fact that it is possible to achieve a non-negligible degree of accuracy (compared to just taking the mean value of observations) using machine learning methods. Furthermore, we have demonstrated that it is possible to achieved a non-negligible degree of prediction accuracy when only a small set of queries (900) are available as a training dataset, which may be a useful technique to employ in scenarios where it is necessary to model the behaviour of the endpoint from only a small number of prior interactions. In general, it can be seen that parameter optimisation (e.g. using cross-validation) also needs to be performed in an endpoint specific manner as it can be seen that the optimal value of K (for k-nearest neighbor regression), or number of random forest predictors for random forest regression, varies considerably between different endpoints. Our approach is applicable in scenarios where clients have no prior information about the behaviour of SPARQL endpoints and must learn from trial and error. Using estimates of the response times for given queries, client applications can make suggestions of query modifications and issue warnings about execution times, for example: (1) Clients can choose to modify a query that may take too long to execute or not give a required level of completeness. (2) Clients can choose to abort queries if they are not likely to obtain a required completeness or not executable within a given time frame.

Furthermore, the work is expected to be of benefit during distributed query processing over multiple SPARQL endpoints, where a query planner must make optimisation decisions in order to effectively execute queries.

４．研究成果

The first part of the research undertaken which studied the optimization of freshness, diversity and coverage in Linked Data Queries resulted in the enhancement of distributed Linked Data query processing techniques described in a paper published at the Web Intelligence Mining and Semantics (WIMS) Conference 2016. An important factor to consider when assessing the significance of the results of this research is that experiments are based on IRI selection strategies that are optimised at runtime during the execution of individual queries. For query processors with cache components and/or the ability to utilise query history data during query execution, the benefits can potentially be multiplied and can significantly improve performance.

The further research undertaken on automated linking of structured Web data (published at the WIMS 2017 conference) and predicting the behavior of SPARQL endpoints both applied machine learning approaches to contribute novel methods supporting querying over Linked Open Data, which can be directly applied within existing query processors to enhance query results.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　0 件）

〔学会発表〕（計　2 件）

Steven Lynden, Makoto Yui, Akiyoshi Matono, Akihito Nakamura, Hirotaka Ogawa, Isao Kojima: Optimising Coverage, Freshness and Diversity in Live Exploration-based Linked Data Queries. Web Intelligence Mining and Semantics (WIMS) Conference 2016, June 2016, Nimes, France.

Steven Lynden. Analysis of Semantic URLs to support automated linking of Structured Data on the Web. Web Intelligence Mining and Semantics (WIMS) Conference 2017, June 2017, Amantea, Italy.

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
LYNDEN Steven（リンデン　スティーブン）
「国立研究開発法人産業技術総合研究所・人工知能研究センター・研究員」
研究者番号： 30528279

(2)研究分担者
(　　　)

研究者番号：

(3)連携研究者
(　　　)

研究者番号：

(4)研究協力者
(　　　)