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

令和　元 年　 ６ 月 １３ 日現在

機関番号：８２６２６

研究種目：若手研究(B)

研究期間：2017～2018

課題番号：１７Ｋ１２７８６

研究課題名（和文）Framework for context-sensitive fact extraction over web data.

研究課題名（英文）Framework for context-sensitive fact extraction over web data.

研究代表者

ＬＥＢＬＡＹ　ＪＵＬＩＥＮ（LEBLAY, Julien）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究員

研究者番号：７０７５７３７７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では、Web上の知識を表現し公開するための優れたモデルの一つである知識グラフから文脈（主に時間情報を含む）を抽出して探索するためのツールと技術の開発を行い、将来の応用のために我々はデータジャーナリズムにおける最新技術を広く調査した。
具体的には、不確実あるいは不完全である可能性がある中で、オントロジデータの文脈について推論するための言語を定義した我々の先行研究に基づいてプロトタイプWebアプリケーションを実装した。
更にこれと並行して、知識グラフにおける事実の経時的妥当性を動的に推論するための機械学習の手法を調査、提案し、評価を行った。

研究成果の学術的意義や社会的意義
We published one survey paper and one demonstration poster and two tutorials in international conferences, and one paper in an international workshop. We are extending this work with neural network-based models, and plan to explore non-temporal contexts in the future.

研究成果の概要（英文）：In this project, we developed tools and techniques to extract context (mostly temporal) from knowledge graphs, one of the prominent model for representing and publishing data on the web, and survey real world applications. This led to an extensive survey and tutorial focusing on applications to data journalism.
We implementing a prototype application based on some earlier work defining a language to reasoning about the context of ontological data in the presence of uncertainty and incompleteness. In parallel, we investigated machine learning approaches to automatically infer the validity over time of the facts in a knowledge graph.

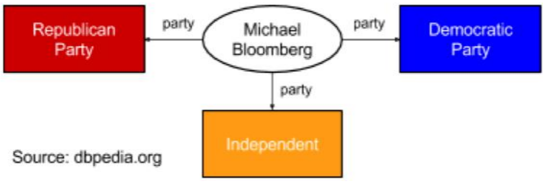
研究分野：情報科学

キーワード：文脈依存推論　機械学習　時間知識グラフ

１．研究開始当初の背景

Knowledge graphs (KG) are graphs in which nodes represent real world concepts (e.g. Nation) or objects (e.g. Japan) and labels directed edges represent relationships among those entities. In other words, a *fact* is each pair of nodes linked with an edge. KGs have become increasingly popular among large Internet services (Google, Facebook, Wikipedia, etc.) to represent, store and exchange web data, in part because they can accommodate a wide of variety data, from schema-less to schema-rich.

DBpedia, a knowledge graph extracted from Wikipedia is a widely used example of such knowledge graphs. Such extracted knowledge graphs however suffer from a lack of context which can lead to inaccuracies or mistakes. This is exemplified in the figure, which shows how Michael Bloomberg is described in DBpedia as being a republican, a democrat and an independent. While all these facts have been true at some points in time, we lack the information about when each of them held makes it hard to use that knowledge for any practically purpose.



Source: dbpedia.org

２．研究の目的

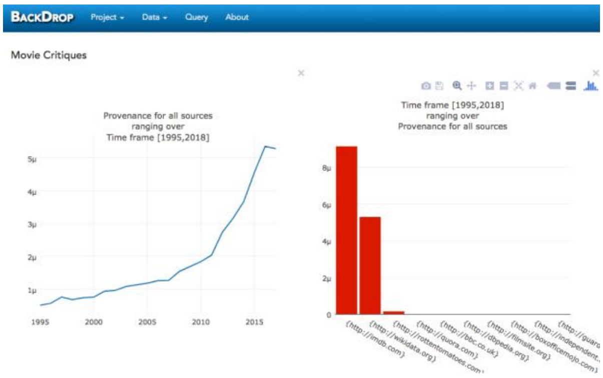
The goal of this research was to investigate (i) how to improve the state of the art in knowledge graph extraction by enhanced the context in which the extracted fact holds, (ii) how such context enrichment could be used in practice. As exemplified above, there are many cases where context is missing. However, there may also be cases where enough contexts are present in the data, but their interaction make the interpretation of the data tricky. Yet, tools for spotting such cases can be valuable in scenarios like computer-assisted journalism, for instance by selecting important periods of time or ruling out unreliable information sources.

３．研究の方法

This research was split into two main branches, closely following the above goals, namely the investigation of context understanding on Knowledge Graph, and applications of this context-based extraction with an emphasis on computational journalism. Journalists increasingly relying on data and understanding and narrow the contexts in which statements are a key part of their work. To tackle the goal (ii) above, we run an extensive review of the area, which was published as a survey in a special track at the WWW conferences, from which we also derived two tutorials presented respectively in two majors conferences (WWW, VLDB). For goal (i), we approach the problem from two distinct angles: exploring how contexts affect the veracity of claims using techniques borrowed from database theory, couple with symbolic and probabilistic reasoning. We also investigate the problem of context extraction from Knowledge Graphs using Machine Learning, namely building models to infer the most time (or period thereof) some given fact could be deemed valid.

４．研究成果

①　Backdrop



Query answering of ontological data can be regarded as a way of produced new knowledge from a background KG. When such a KG has context information (such as time or provenance), it is important to track how each answer to a query holds with respects to the context in which it was computed. Query answering of ontological data can be regarded as a way of produced new knowledge from a background KG. When such a KG has contextual information (such as time or provenance), it is important to track how each answer to a query held with respects to the context in which it was computed. The goal of model described in (Leblay, 2017), which we implemented and demonstrated in the BackDrop prototype. The system can store data sets made of ontological facts and axioms. In this model, axioms can we endowed with a weight reflecting their likelihood in the real worlds. In addition, each axiom or fact can be endowed with any combination of contexts. Typically describing when did the axiom hold and which sources does it come from. For instance, this can model laws that hold in some countries but not others. When computing answers to queries on such data, BackDrop computes a "veracity score" for every possible context and for each answer and displays it in a user-friendly manner, as depicted on Figure 2. The code for this software was publicly released under open source license.

② Fact checking Survey

In the pursuit of applications for context-sensitive knowledge graphs, we investigated the state of the art in computational fact checking. This emerging field tackles the problems of verifying claims automatically or semi-automatically against reference data. One of the difficulties with this issue that the context in which a claim is made can have a critical influence on the intended meaning, and consequently its verification process. We run an expansive survey of the state of the art in computation fact checking which we detailed in international workshop. In addition, this work was presented as a 3-hour tutorial in an international conference on data management.

③ Deriving temporal scope for knowledge graphs

As mentioned before, there are well-established techniques to automatically extract knowledge from web data, but the extracted knowledge typically lacks crucial context information. Rather than devising new technique to extract such context along with the along, we adopted the alternative approach of leveraging on the work already done by such extracted and rather, extract context from their output. To simplify the problem, we focused one type of context extraction, namely time, i.e. in other word assigning validity time to otherwise non-temporal facts. For this first extended existing technique for knowledge graph extraction, e.g. TransE (Borders, 2013) and RESCAL (Nickel, 2011), to support time as an additional dimension. These first approaches being infructuous, we turned to factorization machines (Rendle, 2010). We did extensive feature engineering and found that non-temporal information (fact for which item validity is irrelevant) can still be used as features to learn about temporal facts. This is in part because such features contain temporal clues (such as date) in natural language. This work is ongoing, as we have started investigating the use of Neural Networks for this takes. Among other advantages, this allows us to support multiple via transfer-learning. Indeed, our current approach learns time validities for increasingly fine time granularity (e.g. centuries, decades, years, months, days, etc.) While still specific to temporal contexts, the approach is generic enough to be extended to other hierarchical contexts, such as space, in future works.

References

Julien Leblay. "A Declarative Approach to Data-Driven Fact Checking." *AAAI* 2017: 147-153

Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." *Advances in neural information processing systems*. 2013.

Nickel, Maximilian, et al. "A Three-Way Model for Collective Learning on Multi-Relational Data." *ICML*. Vol. 11. 2011.

Rendle, Steffen. "Factorization machines." *2010 IEEE International Conference on Data Mining*. IEEE, 2010.

５．主な発表論文等

〔雑誌論文〕（計　1　件）

① Sylvie Cazalens, Julien Leblay, Ioana Manolescu, Philippe Lamarre, Xavier Tannier, Computational fact-checking: a content management perspective. PVLDB 11(12): 2110-2113 (2018)

〔学会発表〕（計　4　件）

① Julien Leblay, Weiling Chen, Steven J. Lynden, Exploring the Veracity of Online Claims with BackDrop. CIKM 2017, 2491-2494

② Julien Leblay, Melisachew Wudage Chekol, Deriving Validity Time in Knowledge Graph. WWW (Companion Volume) 2018, 1771-1776

③ Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, Xavier Tannier, A Content Management Perspective on Fact-Checking. WWW (Companion Volume) 2018, 565-574

④ Julien Leblay, Ioana Manolescu, Xavier Tannier, Computational fact-checking: problems, state-of-the-art, and perspectives. WWW (Tutorial Track) 2018

〔図書〕（計　０　件）

〔産業財産権〕
○出願状況（計　０　件）

○取得状況（計　０　件）

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：
ローマ字氏名：

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。