International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Skyline queries over possibilistic RDF data ☆

Amna Abidi [a], Sayda Elmi [a], Mohamed Anis Bach Tobji [a,b,*], Allel HadjAli [c], Boutheina Ben Yaghlane [d]

[a] Université de Tunis, ISG, LARODEC, Tunisia
[b] Univ. Manouba, ESEN, Tunisia
[c] University of Poitiers, LIAS, ISAE-ENSMA, France
[d] University of Carthage, IHEC, LARODEC, Tunisia

ARTICLE INFO

Article history:
Received 15 March 2017
Received in revised form 6 October 2017
Accepted 11 November 2017
Available online xxxx

Keywords:
RDF data
Skyline operator
Semantic Web
Possibility theory

ABSTRACT

Volume and veracity of data on the Web are two main issues in managing information. In this paper, we tackle these two issues, with a particular interest to Resource Description Framework (RDF) data. For veracity management, we rely on a powerful uncertainty theory, namely possibility theory. Therefore, we propose a model for representing and managing possibilistic RDF data. Alongside, to filter the massive amount of RDF data, we use the skyline operator to find out a small set of resources that satisfy predefined user preferences. To this aim, we also propose a skyline operator to extract possibilistic RDF resources that are possibly dominated by no other resources according to Pareto dominance definition. We introduce a dominance operator and a skyline model adopted to the aforementioned kind of data. In addition, we propose an efficient algorithm to compute the skyline with a reasonable performance. Experiments led on the skyline computation showed satisfying results.

1. Introduction

The amount of data on the Web is growing more and more making it hard to find relevant and useful information matching a user query. The reason for creating the semantic Web is to bring structure to the meaningful content of Web pages to facilitate and automate its access. To this end, resources on the Web are marked with labels that describe their structure; these labels are called meta-data. The aim is to make content not only readable by humans, but also by machines, making it possible to analyze and manage huge volumes of Web data. To this purpose, the W3C (World Wide Web Consortium) community introduced a recommendation for semantic annotations; the Resource Description Framework (RDF) [1].

Let us note that large adoption of the semantic Web developed the amount of RDF data on the Web, with datasets increasing in variety and volume. In this paper, we opt for using Preference-based queries that show encouraging results to personalize and filter the massive amount of information residing in today's databases and Information Systems. Among all preference relations, skyline preference relations, defined using Pareto accumulation, have been the most extensively studied [2–5] and extended over Graph Data such as in [6,7]. The skyline operator aims to make multi-objective decisions

☆ This paper is part of the Virtual special issue on Uncertainty Reasoning for the Web, edited by Fernando Bobillo, Kenneth J. Laskey, Trevor Martin, Matthias Nickles.

* Corresponding author at: Université de Tunis, ISG, LARODEC, Tunisia.
E-mail addresses: amna.abidi@ensma.fr (A. Abidi), saida.elmi@ensma.fr (S. Elmi), anis.bach@isg.rnu.tn (M.A. Bach Tobji), allel.hadjali@ensma.fr (A. HadjAli), boutheina_yaghlane@yahoo.fr (B. Ben Yaghlane).

https://doi.org/10.1016/j.ijar.2017.11.005
0888-613X/

over complex data. Given such a multi-criteria preference, the system should be able to identify all potentially interesting data records according to user preferences [8,9].

On the other hand, openness of the Web and diversity of sources on the Internet impacts data veracity. The last two decades have witnessed a profusion of research effort on supporting complex decision making over uncertain RDF data, such as in [10,11]. Our idea is to deal with uncertain RDF data using possibility theory, which is a non-classical theory of uncertainty. It constitutes an alternative to capture different kinds of imperfection, such as imprecision, total ignorance, and partial ignorance that are not representable in probability theory [12].

Thus, we integrated in the structure of RDF data a possibility measure for each subject-property-object triple to reflect the user opinion about the truth of a statement. The possibility measure can be considered as a way to express a source reliability. Our main contributions in this work are:

- Modeling uncertain and imprecise RDF data through possibility theory. We introduce comparison operators between possibility distributions to allow dominance computations in a RDF data set context.
- Extending the skyline operator over possibilistic RDF data. The work of [13] has extended skyline queries over RDF data. To reach the uncertain model, we rethought the dominance operator between two possibilistic RDF points.

The paper is organized as follows: Section 2 includes the background material. In Section 3, we introduce the possibilistic RDF database model. Section 4 explains how the possibilistic-skyline model extends to uncertain RDF data. Then, Section 5 presents algorithms for computing possibilistic RDF skylines, and Section 6 the results and interpretation of experiments. Finally, before concluding in Section 8, we provide a summary of related works in Section 7.

2. Background material

2.1. RDF data model

RDF is a *W3C* framework for representing meta-data and describing the semantics of information in a machine-accessible way [1].

Assume we have a finite set of RDF URI references (U); a finite set of Blank nodes (B); and a finite set of RDF Literals (L). A triple $< Subject, Predicate, Object >$ or $(s, p, o) \in (U \cup B) \times U \times (U \cup B \cup L)$ is called an RDF triple.

RDF describes Web resources (subject) related/characterized (via a predicate or a property) to other resources/literals (object).

Example 1. Below an example of RDF triple data illustrating information about a hotel (Extracted from source $R1$). The hotel with ID h_1 is "Cottages", the street address is "6 Graham Place", located in "Franz Joseph" city, and the country is "New Zealand" (triples 6 and 7). The residence in such hotel costs 5 dollars per night as it is 3 kilometers away from the beach.

1. $(http://hotel.org/h_1, < hasName >, \text{“}Cottages\text{”})$;
2. $(http://hotel.org/h_1, < hasAddress >, \text{“}6\ Graham\ Place\text{”})$;
3. $(http://hotel.org/h_1, < hasPrice >, 5)$;
4. $(http://hotel.org/h_1, < hasDistance >, 3)$;
5. $(http://hotel.org/h_1, < city >, http://example.org/Franz-Joseph)$;
6. $(http://example.org/Franz-Joseph, < hasName >, \text{“}Franz-Joseph\text{”})$;
7. $(http://example.org/Franz-Joseph, < locatedIn >, http://example.org/New-Zealand)$;

A set of RDF statements is a graph for representing meta-data and describing the semantics of information in a machine-accessible way. Therefore, RDF data can be thought in terms of a decentralized directed labelled graph. The edges' labels are the "properties", also called "predicates" or "attributes". The RDF data is stored as a set of $< s, p, o >$ triples, and represented graphically as shown in Fig. 1.

2.2. The skyline operator

The skyline preference operator uses Pareto accumulation. In a set of tuples denoted by S, the skyline consists of the tuples which are dominated by no other tuple [2]. Skyline aims to make intelligent decisions over multi-dimensional data. It consists in finding the most interesting objects according to user-defined criteria (user's preference).

Definition 1. Pareto dominance. Let X and Y be two points in a set of points denoted D with n attributes. A point Y dominates a point X denoted by $Y \succ X$, if $\forall i \in [1, n]\ \ y_i \leq x_i\ \ \Lambda \exists j, y_j < x_j$. The logical dominance concept between two points is modeled as follows:

$$Y \succ X = \bigwedge \Big(\bigwedge_{1 \leq i \leq n} y_i \leq x_i, \bigvee_{1 \leq i \leq n} y_i < x_i \Big)$$