An Innovative of Semantics-Empowered Sensors, Services and Social Computing on the Ubiquitous Web

E. Fathima

Department of Computer Science,
Bharath University, India

Abstract: Semantic Web architecture is the automated conversion and storage of unstructured data sources in a semantic web database. Semantic Web applications are intended to automatically extract and process the concepts and context in the database in a range of highly flexible tools. The Project presents a method for rapid development of benchmarks for Semantic Web knowledge base systems. At the core, we have a synthetic data generation approach for RDF that is scalable and models the real world data. The data-generation algorithm learns from real domain documents and generates benchmark data based on the extracted properties relevant for benchmarking. This is important because relative performance of systems will vary depending on the structure of the ontology and data used. The proposed system approach helps overcome the problem of having insufficient information on the structure of the ontology and data used for benchmarking and allows us to develop benchmarks for a variety of domains and applications in a very time efficient manner. The proposed system measure and analyze the graph features of Semantic Web (SW) schemas with focus on power law degree distributions and conducted an experiment of synthetic data generation on five Semantic Web knowledge base systems. The Project mainly aims to prove the influence of ontology and data on the capability and performance of the systems and thus the need of using a representative benchmark for the intended application of the systems.

Key words: RDF that is • Project presents • Semantic Web knowledge

INTRODUCTION

A number of prominent ontologies have emerged on the Semantic Web as the de facto choice for describing data in specific domains [1]. Classes and properties in these specifications are identified via common identifiers in the form of URIs, which are consistently used across sources. On the instance level, however, there is still little agreement between sources on the use of common URIs to identify specific entities, such as people, forums, or categories [2]. In fact, since the assignment of URIs to instances is optional within the Resource Description Framework (RDF), many entities are described anonymously without use of a URI. Where agreement on URIs for resources cannot be reached, multiple URIs may exist for the one resource [3]. The problem of object consolidation has received significant attention in the database community under the names of record linkage, instance fusion and duplicate identification [4].

Due to the lack of formal specification for determining equivalences, these approaches are mostly concerned with probabilistic methods [5]. In more formal approaches, properties unique to an entity can be used to determine its identity; examples are personal public services numbers or email addresses which can be used to identify people. On the Semantic Web, these inverse functional properties” are specified in Web Ontology Language (OWL) descriptions. For example, from the DCMI specification one can determine that dc: Publisher (a person’s name) is a property unique to a person; hence, two instances sharing the same value for dc: Publisher properties must be the same real-world entity. There is much more agreement on values that are established in real-world areas or widely used applications such as email addresses or instant messaging usernames than on seemingly artificial URIs as the identifying factor of entities.

Corresponding Author: E. Fathima, Department of Computer Science, Bharath University, India.
Thus, equivalence of identifiers based on inverse functional properties has to be taken into account when merging RDF graphs, since infrequent reuse of instance identifiers across sources leads to problematic data integration: the total knowledge contribution for an entity may be fragmented over multiple instances. One desired outcome of the Semantic Web effort is a massive data graph spanning the Web and agreement on identifiers is crucial to achieving a well-connected graph where associations between entities are recorded correctly. Fusing identifiers is especially important for entity-centric applications relying on RDF data such as search and query engines, interactive browsing tools and data mining systems.

There are a number of challenges related to performing Synthetic data generation on Semantic Web data:

The Web is massive in size, therefore methods operating on considerable fractions of the Web require scalable algorithms. In particular, we require an algorithm that scales object consolidation to the Web. Data gathered from the Web has been created by a large number of people and exhibits a lot of variance in terms of quality and completeness. Therefore, we require our algorithm to be robust in the face of potentially problematic data.

**XML**

**XML Schema:** The DTD and Schema are used to describe the grammar and restriction over data in the XML document. DTD and schema are used to specify the structure of instance documents and the data type of each element/attribute. DTDs used today in the XML originated from the parent SGML specification. Because SGML was designed for a more document-centric model, it did not require the use of complex data typing definitions. The XML Schema specification improves greatly upon the DTD content model by providing rich data typing capabilities for elements and attributes as well as providing OO design principles.

XML Schema was approved as a W3C Recommendation in May, 2001 and is now being widely used for structuring XML documents for e-commerce and Web Services applications.

The two major goals that the W3C XML Schema working group focused on during the design of the XML Schema standard were:

Expressing Object Oriented design principles found in common OO programming languages into the specification.

Providing rich data typing support similar to the data typing functionality available in most relational database systems.

XML Schemas provides a means of creating a set of rules that can be used to identify document rules governing the validity of the XML documents that you create. Schemas provide a means of defining the structure, content and semantics of XML documents that can be shared between different types of computers and documents.

**Resource Description Framework - RDF:** RDF - the Resource Description Framework - is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. RDF emphasizes facilities to enable automated processing of Web resources.

RDF metadata can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library by intelligent software agents to facilitate knowledge sharing and exchange in content rating in describing collections of pages that represent a single logical "document" for describing intellectual property rights of Web pages.

RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration and other applications.

**Scope:** The RDF Schema specification is not aimed at theoretical issues, but at solving a small number of immediate problems. Its creators expect that other problems (some of which are illustrated in the examples below) will share similar characteristics and that they also may be able to use the basic classes described in this specification. The RDF Schema specification was directly influenced by consideration of the following problems:

**Platform for Internet Content Selection (PICS):** The RDF Model and Syntax is adequate to represent PICS labels, however it does not provide a general-purpose mapping from PICS rating systems into an RDF representation. One such mapping is described in a separate document.

**Simple Web Metadata:** One obvious application for RDF is in the description of Web pages. This is one of the basic goals of the Dublin Core Metadata Initiative. The Dublin Core Element Set is a set of 15 elements believed to be broadly applicable to describing Web
resources to enable their discovery. The Dublin Core has been a major influence on the development of RDF. An important consideration in the development of the Dublin Core was to not only allow simple descriptions, but also to provide the ability to qualify descriptions in order to provide both domain specific elaboration and descriptive precision.

The RDF Schema Specification provides a machine-understandable system for defining schemas for descriptive vocabularies like the Dublin Core. It allows designers to specify classes of resource types and properties to convey descriptions of those classes, relationships between those properties and classes and constraints on the allowed combinations of classes, properties and values.

**Sitemaps and Concept Navigation:** A sitemap is a hierarchical description of a Web site. A subject taxonomy is a classification system that might be used by content creators or trusted third parties to organize or classify Web resources. The RDF Schema specification provides a mechanism for defining the vocabularies needed for such applications.

Thesauri and library classification schemes are well known examples of hierarchical systems for representing subject taxonomies in terms of the relationships between named concepts. The RDF Schema specification provides sufficient resources for creating RDF models that represent the logical structure of thesauri (and other library classification systems).

**P3P:** The W3C Platform for Privacy Preferences Project has specified a grammar for constructing statements about a site's data collection practices and personal preferences as exercised over those practices, as well as a syntax for exchanging structured data.

Although personal data collection practices have been described in P3P using an application-specific XML tagset, there are benefits to using a general metadata model for this data. The structure of P3P policies can be interpreted as an RDF model. Using a metadata schema to describe the semantics of privacy practice descriptions will permit privacy practice data to be used along with other metadata in a query during resource discovery and will permit a generic software agent to act on privacy metadata using the same techniques as used for other descriptive metadata. Extensions to P3P that describe the specific data elements collected by a site could use RDF Schema to further specify how those data elements are used [6].

**Dublin Core Metadata Initiative:** The Dublin Core Metadata Initiative (DCMI) is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems.

The Dublin Core is a metadata element set intended to facilitate discovery of electronic resources. Originally conceived for author-generated description of Web resources, it has attracted the attention of formal resource description communities such as museums, libraries, government agencies and commercial organizations.

The key characteristics of the Dublin Core are:

- Simplicity
- Semantic interoperability
- International consensus
- Extensibility
- Modularity
- Dublin Core- Elements

**Element Core Metadata Set**

**The Basic Unit of Metadata Is a Statement:**

<table>
<thead>
<tr>
<th>Title</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Identifier</td>
</tr>
<tr>
<td>Description</td>
<td>Source</td>
</tr>
<tr>
<td>Creator</td>
<td>Language</td>
</tr>
<tr>
<td>Publisher</td>
<td>Relation</td>
</tr>
<tr>
<td>Contributor</td>
<td>Coverage</td>
</tr>
<tr>
<td>Date</td>
<td>Rights</td>
</tr>
</tbody>
</table>

**Classes and Properties:** This document specifies the RDF Schema mechanism as a set of RDF resources (including classes and properties) and constraints on their relationships. The abstract RDF Schema core vocabulary can be used to make RDF statements defining and describing application-specific vocabularies such as the Dublin Core Element Set.

**The Type System:** The RDF Schema defined in this specification is a collection of RDF resources that can be used to describe properties of other RDF resources (including properties) which define application-specific RDF vocabularies.

**Core Classes:** The following resources are the core classes that are defined as part of the RDF Schema vocabulary. Every RDF model that draws upon the RDF Schema namespace (implicitly) includes these.
**rdfs: Resource:** All things being described by RDF expressions are called resources and are considered to be instances of the class rdfs: Resource. The RDF class rdfs: Resource represents the set called 'Resources' in the formal model for RDF presented in section 5 of the Model and Syntax specification.

**rdf: Property:** rdf: Property represents the subset of RDF resources that are properties, i.e., all the elements of the set introduced as 'Properties' in section 5 of the Model and Syntax specification.

**rdfs: Class:** This corresponds to the generic concept of a Type or Category, similar to the notion of a Class in object-oriented programming languages such as Java. When a schema defines a new class, the resource representing that class must have an rdf: type property whose value is the resource rdfs: Class. RDF classes can be defined to represent almost anything, such as Web pages, people, document types, databases or abstract concepts [7].

**Module Description:**
- RDF Schema Conversion based on DCMI
- Graph Generation and Optimization
- Synthetic Data Generation and access authorization

**RDF Schema Conversion Based on DCMI:** The system can perform automatic Dublin Core-based metadata extracting algorithm from information elements which is described by RDF documents. The Dublin Core Metadata Initiative is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models.

The input of the RDF Schema Conversion system is RDF-based information elements documents, difference functions and extracting functions. The output of the system is metadata which is extracted from the RDF based information elements. The conversion of information elements into RDF schema is performed.

The input of our project is Net Banking Ontology file. This file has contained the data with 3 formats.
**Bank Database**

**Graph Generation and Optimization:** Semantic web schemas are usually represented as directed labeled graphs where nodes are classes or literal types and arcs are properties. These graphs may have self-loops (representing recursive properties) and multiple arcs (when two classes are connected by several properties). In particular, Semantic Web schemas have two different kinds of arcs. The first comprises attributes or relationships among classes, which are called properties. The second comprises subsumption relationships among classes. As the interpretation of these two arc kinds is different, for each semantic web schema we need to define two graphs:

- The property and
- The subsumption graph.

Both graphs have the same set of nodes (i.e., the union of classes and literal types used in the schema) but they comprise different kinds of arcs.

**Property Graph:** The Class Property Graph is a graph that has some property associated with each of the vertices or edges in the graph. As a given graph may have several properties associated with each vertex or edge, a tag is used to identity which property is being accessed. The graph provides a function which returns a property map object [8].

**Subsumption Graph:** The class subsumption graph of a schema is a directed graph \( G_s = (C, Ps) \), where \( C \) is a set of nodes labeled with a class name and \( Ps \) is the transitive and reflexive binary relation that represents subsumption relationships among classes.

**Power Law Distributions:** The project is designed to evaluate whether the property graph exhibits a power-law degree distribution. Examining the distributions involved in the subsumption graph could provide additional hints about the morphology of Semantic Web schemas. Specifically, the project implements power-laws on two different functions of a DRV (Discrete Random Variable) \( X \).

The first, called Complementary Cumulative Density Function (CCDF), i.e., \( P(X = x) \), measures the frequencies of \( X \) values,

- while the second, denoted by VR (Value vs Rank), measures the relationship between the \( i \) - th biggest \( X \) value and its rank (in descending order), \( i \).

- Regarding the property graph, it should be noted that it may contain self-loops (e.g., recursive properties) and multiple arcs (i.e., two classes may be connected by more than one property).

**Synthetic Data Generation:** Knowledge about object attributes and their relationships can be represented through a semantic graph, with vertices representing the attributes and links representing the fact that two attributes are associated. The independent attributes with no incoming links will be represented in thicker border. Values of other attributes are determined by the values of the preceding ones on the graph. For example, the minimum balance depends on the Account type of that person. The idea is to provide as much connectivity in this graph as possible to obtain meaningful relationships. Furthermore, the project will allow for easy addition of new relationships as they are discovered.

Testing: Testing is a means of discovering the quality level of a software system, not a means of assuring software quality. Testing cannot show the absence of defects, it can only show that software defects are present. So when planning is on for software testing, the activity has to be considered as a destructive activity of the software that has been built.

Software Testing is a critical element of software quality assurance and represents the ultimate review of specification, design and coding. The programmer creates a series of test cases that are intended to demolish the software that has been built. In fact, testing is the one step in the software engineering process that could be reviewed as destructive rather than constructive.

Software developers are by nature constructive people. Testing requires that the developer discard preconceived notions of the correctness of the software just developed and overcome a conflict of interest that occurs when errors are uncovered.
The Objectives of Testing Are: Testing is a process of executing a program with the intent of finding an error. A good test is one that has a high probability of finding an as yet undiscovered error.

A successful test is one that uncovers an as yet undiscovered error. Software testing is an important element of S/W quality assurance and represents the ultimate review of specification design and coding [9].

Unit Testing: Unit testing focuses verification effort on the smallest unit of S/W design. The tests that occur as part of unit testing is illustrated below. The module ‘interface’ is tested to ensure that information properly flows into and out of the program unit under test. The local data structures are examined to ensure that data stored temporarily maintains its integrity during all steps in an algorithms execution.

Boundary conditions are tested to ensure that the module operates properly at boundaries established to limit to restrict processing. All ‘independent paths’ through the control structures are exercised to ensure that all statements in a module have been executed to ensure that all statements in a module have been executed at least once. Finally all error-handling paths are tested.

Unit Testing is conducted for each modules separately with standard input for getting standard output.

RDF Schema Conversion for Unstructured Data: In this Module, standard elearning related data (group of data) are inputs such as studentid, course, documents…, once the get output (related to Input) for this standard Input the analyzed standard output versus input . If there is any changes occur across the module again continue the beginning process of this module otherwise the module should be accepted.

Graph Generation and Optimization: In this module the process of graph generation accepts the rdf schema converted data. The output of the module is property graph and subsumption graph. The output will be compared for expected result by subsets of data. If the output varies from the standard output configuration the process will changed for formatting the output.

In this module the process of optimisation accepts the graph generated and applys powerlaw degree distribution. The output of the module is the values CCDF and VR which will be optained by graph theory for nodes and edges linkage calculation. The output will be compared for expected result by subsets of data. If the output varies from the standard output configuration the process will changed for formatting the output.

Synthetic Data Generation and Access Authorization: In this module the process of synthetic data generation accepts optimized structure optained by applying power law degree distribution. The output of the module is the rdf data file where the schema is synthetically generated based on the optimization vlaues. The output will be compared for expected result by subsets of data. If the output varies from the standard output configuration the process will changed for formatting the output.

Integrating Testing: Integration testing is a systematic technique for constructing the program structure while at the same time conducting tests to uncover errors associated with interfacing. The objective is to take unit-tested modules and build a program structure that has been dictated by design. Top-Down Integration are tested finally.

The modules RDF Schema Conversion for unstructured data, Graph Generation and Optimization, Synthetic Data Generation are tested for integration. The output of the modules is tested for acceptance in the next module for expected results sequentially.

Performance Analysis: The web based graphic user interface for easy viewing of the graph is built and is illustrated in Figure. It allows users to view all graphs (both property and subsumption) or literals only, view graph sorted by different users. In Figure 2.2, detail display is chosen, so queries that were used to produce property graphs are also listed under the queries. One can see that some graphs might be too general. It should be noted that these graphs were produced from only few request over the few data generates. With schema of more data paths, more reliable and interesting graph paths will be found.

To fully test how effectively those degree distribution help to improve the performance of the semantic web, we would deploy the synthetic data generation into the power law degree distributions. At this point, our research effort has not yet been transferred to the access authorization, therefore, we have to run our project more realistically, we extract the rdf file into properties based on users: a particular schema data which is about 30% of all. We then run our power law degree distribution on the data to calculate CCDF and VR.
Two criteria are used to see the effectiveness of the data generated from the schema. The first is the recall; it measures the percent of the testing queries that are matched with the synthetic data. The higher the recall, the more queries are benefited from using synthetic data. The second is the precision, which measures, among queries matched with the synthetic data, the percent of the testing queries that are correctly answered [10].

![Performance Comparison](image)

CONCLUSION

The project described the features, architecture, design and development details for a synthetic data generation process that enables evaluation of consistency of the rdf semantic graph [11-15]. The implementation demonstrates the feasibility of generating synthetic test data using complex semantic graphs with multi-layer and high-dimensional data dependencies. The power-law distribution of degrees implies that the knowledge represented by semantic web ontology can be very inhomogeneous while the majority of concepts use (refer to) a few other concepts. The concepts that have large degrees could be treated as more “essential” knowledge points, as they attract more connections.

REFERENCES

5. HongBin Wang, Da Xin Liu and Wei Sun, 2008. Dublin Core-based Metadata Extracting Algorithm from RDF-based Information Elements Knowledge Discovery and Datamining IEEE.