A Framework for Adaptive RDF Graph Replication for Mobile Semantic Web Applications

Bernhard Schandl and Stefan Zander

University of Vienna
Department of Distributed and Multimedia Systems
{bernhard.schandl,stefan.zander}@univie.ac.at

Abstract. An increasing number of applications are based on Semantic Web technologies and the amount of information available on the Web in the form of RDF is continuously growing. The adaption of the Semantic Web for Personal Information Management and the increasing desire for mobility is often accompanied by situations where no network connectivity is available and hence access to remote data is limited. Such situations could be obviated when mobile devices are able to operate on offline data replicas and synchronize changes when connectivity is re-established. In this paper we present our ongoing work in developing a framework allowing for adaptive RDF graph replication and synchronization on mobile devices. We propose to interpose components that analyze various information sources of semantic applications (including ontologies, queries, and expressed user interest) and use them for selecting parts of RDF data bases, which are then made available offline using a proxy SPARQL endpoint on a mobile device. Thus, we provide access to Semantic Web data without the need for permanent network connectivity.

1 Introduction

The original design of the World Wide Web is *document-centric*: digital information resources are published on servers and can be retrieved by using Uniform Resource Locators (URLs). Such documents are mainly HTML pages with embedded media like images, which are connected by hyperlinks. While there exist a large number of static documents (i.e., documents that reside on a server and are delivered to clients as-is), large amounts of data are embedded in the so-called *hidden web*, which consists of virtual documents that are created on request time using data that is stored in other systems, e.g. relational data bases. In most cases, these data are exposed via query forms and are available to clients also in the form of semi-structured HTML documents.

If the consumer of such data is not a human (through the usage of a Web browser) but a machine, it is required to re-extract the raw data from the HTML representation, being optimized for human consumption, which is usually an expensive and error-prone task [6]. It is the goal of the *Semantic Web* [2] to eliminate this source of potential errors by providing the technical infrastructure to

directly publish machine-interpretable information on the Web, thus making it *data-centric*. The Semantic Web builds upon the Web infrastructure [14] and extends it with a meta format for information representation (RDF [13]) and languages that allow publishers to semantically describe their data (e.g., RDF Schema [4] and Web Ontology Language [9]). This technology stack has been complemented by the activities of the Linked Open Data initiative, which demonstrate how to publish and interlink data sets using Semantic Web technologies [3] and hence creating a world-wide distributed database.

Recently, the application of Semantic Web technologies to the problem of Personal Information Management (PIM) has gained lots of interest, most notably in the form of the *Semantic Desktop* [18], which has been investigated in the course of a number of projects (e.g., [12, 15, 19]). With the increasing proliferation of mobile devices like smart phones or netbooks, issues of Personal Information Management are no longer restricted to desktop machines. In mobile scenarios, users frequently face the problem that data is not available because of several reasons: firstly, there may be no physical network connectivity (e.g., because of the lack of mobile network coverage), and secondly, security restrictions may apply (e.g., a VPN connection to the company network cannot be established). In such situations it is desirable to make relevant data available on the mobile device so that applications can operate offline, and to synchronize changes back to the base data set when connectivity is recovered. However, because of the still limited storage and computing power of mobile devices, it is advisable to carefully select the information to replicate; ideally in an automatic, transparent, and adaptive manner.

In this paper we present our ongoing works towards a framework that aims to provide this functionality. Its architecture consists of a number of middleware components that selectively replicate data from an RDF data base to a (mobile) client. This selection is done by considering, on the one hand, automatically derived metrics about the data set and its usage, and, on the other hand, manually defined rules that allow the user to specify subsets of the data to be replicated. On the mobile device, replicated data are wrapped by a SPARQL endpoint to be transparently used by applications.

2 Mobile RDF Replication and Synchronization Architecture

In Figure 1 the typical architecture of RDF-based applications is depicted. Such applications usually consist of two main components:

- A *SPARQL endpoint*, which wraps an RDF dataset and hides its implementation details from a client. The data may actually be stored in a relational database, in the file system, in memory, or it may be accessible via a network protocol. The endpoint implementation accepts SPARQL query strings, executes them on the actual RDF data, and returns the results in the correct target format.

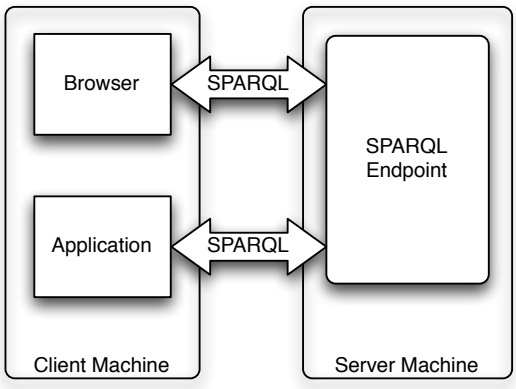

Fig. 1: Typical Architecture of Semantic Web-based Applications

– An *application*, which accesses RDF data by issuing SPARQL queries to the endpoint, and interprets the results[1]. Just as it is the case with applications that build upon relational databases, all details of generating results and processing updates are delegated to the SPARQL engine. The only defined interface between the application and the data set is the SPARQL language and its transport protocol [7].

Naturally, our proposed replication and synchronization mechanisms are beneficial only in situations where these components are distributed over different physical machines and the network link between them is potentially unstable (e.g., when the SPARQL endpoint resides on a company server, while the application is executed on an employee's mobile device).

To introduce a replication and synchronization layer into such a semantic application, it is not necessary to modify any of the existing system components. Instead, we introduce two new components that serve as mediator layer between the client application and the SPARQL endpoint. We denote these components the client-side *replication engine* and the server-side *replication manager*. This extended system architecture is depicted in Figure 2 and described in the following.

Replication Engine The replication engine is instantiated on the client machine and acts as a transparent proxy for applications. The only change to applications is a configuration modification: applications must be re-configured to query the local SPARQL endpoint instead of the original remote endpoint.

[1] We assume that update functionality will be included into SPARQL in the near future; the current effort towards this direction has been subsumed by a corresponding W3C member submission, cf. `http://www.w3.org/Submission/2008/SUBM-SPARQL-Update-20080715/`.

The replication engine is a fully-functional SPARQL endpoint that is able to process queries and return the results to the application. It is configured to establish a connection to the original SPARQL endpoint, as well as to a corresponding replication manager. It has two operation modes, *online* and *offline* mode. In online mode all queries are directly passed to the original (remote) SPARQL endpoint, and results from the endpoint are forwarded to the application where the request originated. In offline mode the replication engine answers queries from its local cache, which holds a subset of the original data set. The virtual endpoint is hence enabled to return at least partial results for application queries, which is a significant improvement compared to situations where no data can be retrieved at all.

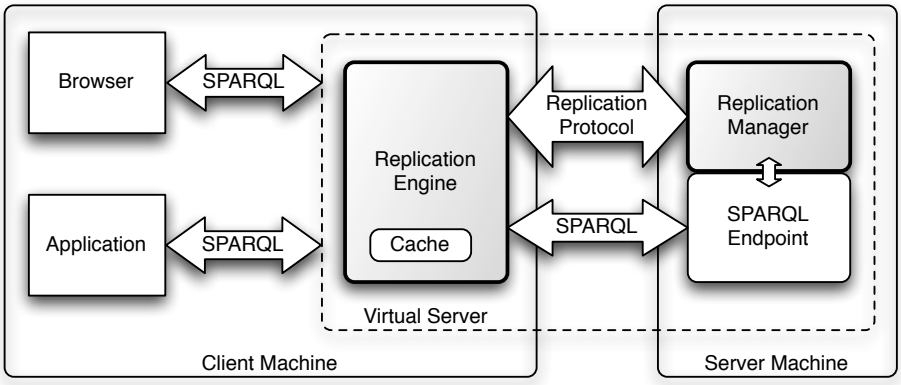

Fig. 2: Proposed Architecture Extension by an Intermediate SPARQL Proxy

Updates are processed in a similar manner: in online mode they are forwarded to both the local cache and the original data base, while in offline mode changes are recorded on the mobile device for subsequent synchronization between the cached copy and the original data set.

Replication Manager The task of the replication manager is to compute a ranking for the selective replication; i.e., it determines which subset of the data is to be replicated on the client. To accomplish this it needs access to the whole RDF data set, which can in general be achieved through the SPARQL endpoint. In order to achieve better performance, it may however be necessary to integrate these two components more tightly, as SPARQL can not be used to notify the manager about data updates. The degree of such an integration is subject of further research.

Replication Control Protocol The replication manager and the replication engine exchange information about the current status of the original endpoint

and the client's cache via a replication control protocol, which is also used to coordinate the execution of data replication tasks. Possible reasons for initiating a new data replication task include the execution of a SPARQL query or a data update on the client machine. The replication control protocol should ensure a maximum of offline data availability in the engine's cache at any time. This strategy is preferred over manual synchronization on demand because it also holds when the network connection is unexpectedly interrupted. Additionally it enables the client to disconnect at any time, instead of requiring it to start a tedious synchronization procedure before a planned disconnect.

Processing and elaborating on user-related contextual data provided by the replication engine is another important task and serves as the basis for the intelligent RDF subgraph selection according to the user's current activities and intentions. Some selection strategies we are considering in our ongoing works are introduced in the following chapter.

3 Selection of RDF Replica Sets

It is not practicable to replicate entire data sets under the restrictions of mobile devices imposed by technical and user-related context. To provide a tradeoff in such situations, we are investigating algorithms for selective replication of RDF sub-graphs. The goal of these algorithms is to provide a *subjective interest ranking* of RDF triples, where we take into account structural and semantic characteristics of the dataset, as well as user preferences and usage context information. In the following we describe some of our envisioned input parameters in more detail.

1. *Graph Structure and Metrics.* RDF is based on a graph model; therefore, various metrics and analysis algorithms can be applied to it (e.g., degrees of graph nodes). We are currently investigating the applicability of these metrics for deriving conclusions on the relevance of graph elements for offline replication. Such metrics, however, do not take into account the semantics of the RDF model and ontologies [22], which is addressed by the following two information sources, ontology structure and queries.
2. *Ontology Structure and Metrics.* Ontologies are used to express shared conceptualizations between communicating partners. In our work we focus on the Web Ontology Language (OWL) [9], which is one of the standard languages for ontology modelling on the Semantic Web. OWL ontologies consist of three types of elements: classes, individuals, and properties. Their structure as well as the semantics of the relationships between them is expressed using different OWL language constructs, e.g., `subClassOf` or `equivalentProperty`. From the analysis of these expressions we hope to be able to infer information about the importance of instance data that adheres to these ontologies, and to detect redundant data that does not need to be replicated on the client.
3. *Queries.* As described in Section 2, applications usually access RDF data through issuing SPARQL queries. Hence, the structure of these queries as

well as the vocabularies used therein are indicators which data are relevant for an application. To exploit this information we will analyze the syntactic and semantic structure of queries (with the help of ontologies, as described before) and draw conclusions regarding the importance of the data sets that these queries are applied to.

4. *User Context.* Context and context-awareness play a critical role in interactive information systems [8, 10]. Recent research in this area reveals that the prevailing system-centric view of context-awareness should be replaced by a user-centric view [20]. Intelligent and adaptive RDF subgraph selection must therefore elaborate on the user's tasks and information needs on a semantic level to provide appropriate and valuable data. For instance, based on upcoming appointments or events in the user's calendar, the replication engine could infer on the data probably needed. We investigate further approaches on how to utilize user behavior and contextual information to enhance the quality of the data retrieval process.
5. *Explicit User Interest.* The end of the Semantic Web information chain is the human user. In every situation, the user should have the possibility to overrule or supplement automatically replicated datasets. This selection may be carried out on various levels, e.g., using elements from an ontology, using range definitions for attribute values, or even (on the lowest level) the selection of single triples out of the graph. Depending on the user's experience, sophisticated user interfaces are required for this task, especially in cases where the amount of data exceeds certain sizes.

From the analysis of these data it may be possible to derive information that is relevant not only to replication and synchronization, but also for other aspects of the stored data: for instance, the analysis algorithms might reveal that certain parts of a data set are never queried. In this case, it could be advisable to move these parts from the live data store into a long-term archive. On the other hand, analysis of data graphs may evidence that sub-graphs are disconnected, therefore semantic relations between resources are missing. If such a graph is generated from an external data source, this may indicate a potential error in the mapping or in the transformation algorithm.

4 Implementation

As a starting point for a reference implementation we have conducted a survey on existing mobile Semantic Web frameworks. We have analyzed two XML parsers for mobile environments, *NanoXML for J2ME*[2] and *kXML*[3], as well as two mobile RDF frameworks, *Mobile RDF*[4] and *μJena*[5]. Our survey revealed that μJena

[2] NanoXML: `http://sourceforge.net/projects/nanoxml-j2me/`
[3] kXML: `http://kxml.sourceforge.net/`
[4] Mobile RDF: `http://www.hedenus.de/rdf/index.html`
[5] μJena: `http://poseidon.elet.polimi.it/ca/?page_id=59`

is the most advanced framework providing ontology and inferencing support, although its API is currently in prototypical status and only allows for processing RDF data serialized in N-Triples format[6]. However, none of the evaluated frameworks supports queries on RDF data via SPARQL or other query languages. A serialization mechanism between RDF and the internal storage mechanisms used by certain mobile devices for storing data permanently could also not be found. Such mechanisms are however needed since many mobile platforms do not use a file system for storing application data, but provide platform-specific storage systems, such as the Record Management System (RMS) in case of J2ME MIDP[7] applications.

We are currently developing our proposed framework as a Google Android[8] application since the underlying operating system provides substantial advantages compared to other mobile operating system architectures. Android itself is an environment for running Java applications on the *Dalvik Virtual Machine*[9] which is especially optimized for mobile environments. It includes *SQLite*, a lightweight and powerful relational database engine, and makes use of some advanced software design patters such as the Model-View-Controller (MVC) pattern to separate application logic from user interface design and underlying data models. Android provides access to the core system operating functions through standard APIs as well as a complete multitasking environment where each application is executed within its own thread, thus providing the possibility to implement background services, like a synchronization process that is automatically activated when the mobile device has online connectivity to its home network (e.g, by automatically establishing a VPN connection within a public wireless local area network).

As a first step we have implemented an initial prototype consisting of a client application for initiating a request, a minimal replication engine, and the replication manager. The replication manager is able to process a core set of contextual information, such as the number of triples expected by the replication engine, the user's current location, as well as information about the serialization formats the client is able to process. The replication engine takes these values as input parameters and sends them to the replication manager. Based on this information the replication manager selects a subset of the RDF data set and transmits it to the client. A RDF abstraction layer has been introduced in the replication manager so that its implementation is independent from the underlying RDF store. The client locally caches the data and hence makes it available to applications, and changes made to this cache are subsequently forwarded to the replication manager. Currently we are designing a more elaborate framework for RDF persistance on mobile devices. On the replication manager side, we are designing and implementing a ranking pipeline that allows for modular,

[6] N-Triples Syntax for RDF: `http://www.w3.org/TR/rdf-testcases/#ntriples`
[7] Mobile Information Device Profile (MIDP): `http://java.sun.com/products/midp/`
[8] Google Android Platform: `http://code.google.com/android`
[9] Dalvik Virtual Machine: `http://www.dalvikvm.com`

customizable weighting of RDF triples, which is used as the basis for selective replication.

5 Related Work

Although RDF databases are gaining industry strength in terms of performance and memory efficiency, mechanisms for synchronization and offline replication can hardly be found. To the best of our knowledge, many of today's state-of-the-art triple stores, including Jena[10], Sesame[11], and Redland[12], do not include support for (selective) offline replication.

Most of the systems mentioned above can be configured to make use of a relational data base to store RDF data. For this, they employ mapping algorithms in order to represent RDF graphs as relations. One could make use of a RDBMS's replication and synchronization facilities; however, this has two drawbacks: (*1*) it does not consider the special aspects of RDF and semantic graphs, including ontologies, and (*2*) performing selective replication is very hard unless the developer analyzes the exact mapping algorithms for the target system. Usually, those systems don not provide possibilities to elaborate on the meaningfulness and semantics of RDF data sets. Larger-scale database systems like OpenLink Virtuoso [11] and Oracle [1] do not solely focus on RDF but may serve as a data integration point for different sources, including RDF. While these systems often provide support for replication and synchronization, they are not designed to be deployed to mobile devices.

A different approach for selective distribution and replication of RDF data is the Peer-to-Peer (P2P) paradigm, where multiple equal systems exchange data over a network. Such systems are, for instance, Edutella [16] and RDFPeers [5]. These works provide valuable knowledge about efficient distribution and exchange of RDF data, but do not focus on selective replication. Tumarrello et al. [21] describe an algorithm for selective exchange of RDF, based on P2P systems. We aim to extend the results presented by them and apply them to non-P2P environments.

The Open Mobile Alliance (OMA) provides the SyncML framework for data synchronization [17], which allows data of different kinds (including contacts, calendars, and e-mail messages) to be synchronized between devices. The framework also specifies a number of bindings to protocols that are commonly used in the context of mobile devices, as well as limited means to express device context information, e.g., the available memory or the supported databases. Since this framework does not consider a generic data format like RDF, we will analyze potential synergies and links between our approach and the OMA activities.

[10] Jena Semantic Web Framework: `http://jena.sourceforge.net`
[11] Sesame Framework: `http://www.openrdf.org`
[12] Redland RDF Libraries: `http://librdf.org`

6 Conclusions

In this paper, we have outlined our ongoing works towards a framework for selective replication of RDF data sets to mobile devices. The goal of this framework is to provide access to RDF data sets in situations where there is no network connectivity available and hence communication with remote data sources is impossibile. Our proposed architecture extends current Semantic Web applications with intermediate components that handle SPARQL queries transparently, either by forwarding them to the actual data store if connectivity is up, or by answering them from a locally cached partial replica of the data set on the mobile device, if there is no connectivity.

We are currently in the process of specifying in more detail the algorithms and data models that are required to realize such a framework. This includes a model for selective replication of RDF data sets, algorithms for ranking of resources based on their structure and usage, and checkout and update mechanisms that enable mobile devices to stay updated with a base data set. In parallel, we are validating these artifacts by the means of a reference implementation, which is based on the Android mobile platform and a special variant of the popular Jena Semantic Web framework.

Acknowledgements Parts of this work have been funded by FIT-IT grants 812513 and 815133 from Austrian Federal Ministry of Transport, Innovation, and Technology.

References

1. Omar Alonso, Sandeepan Banerjee, and Mark Drake. GIO: A Semantic Web Application Using the Iinformation Grid Framework. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 857–858, New York, NY, USA, 2006. ACM.
2. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American Magazine*, 284(5):34–43, 2001.
3. Chris Bizer, Richard Cyganiak, and Tom Heath. *How to Publish Linked Data on the Web*, 2007. Available at `http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/`, retrieved 02-Dec-2008.
4. Dan Brickley and R.V. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema (W3C Recommendation 10 Februar 2004)*. World Wide Web Consortium, 2004.
5. Min Cai and Martin Frank. RDFPeers: A Scalable Distributed RDF Repository Based on a Structured Peer-to-peer Network. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 650–657, New York, NY, USA, 2004. ACM Press.
6. Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled F. Shaalan. A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411–1428, 2006.
7. Kendall Grant Clark, Lee Feigenbaum, and Elias Torres. *SPARQL Protocol for RDF (W3C Recommendation 15 January 2008)*. World Wide Web Consortium, 2008.

8. Joëlle Coutaz, James L. Crowley, Simon Dobson, and David Garlan. Context is Key. *Commun. ACM*, 48(3):49–53, 2005.
9. Mike Dean and Guus Schreiber. *OWL Web Ontology Language Reference (W3C Recommendation 10 February 2004)*. World Wide Web Consortium, February 2004. Available at `http://www.w3.org/TR/owl-ref/`.
10. Paul Dourish. What we talk about when we talk about context. *Personal Ubiquitous Comput.*, 8(1):19–30, 2004.
11. Orri Erling and Ivan Mikhailov. RDF Support in the Virtuoso DBMS. In Sören Auer, Christian Bizer, Claudia Müller, and Anna V. Zhdanova, editors, *CSSW*, volume 113 of *LNI*, pages 59–68. GI, 2007.
12. Tudor Groza, Siegfried Handschuh, Knud Moeller, Gunnar Grimnes, Leo Sauermann, Enrico Minack, Cedric Mesnage, Mehdi Jazayeri, Gerald Reif, and Rosa Gudjonsdottir. The NEPOMUK Project - On the Way to the Social Semantic Desktop. In Tassilo Pellegrini and Sebastian Schaffert, editors, *Proceedings of I-Semantics' 07*, pages pp. 201–211. JUCS, 2007.
13. Patrick Hayes. *RDF Semantics (W3C Recommendation 10 February 2004)*. World Wide Web Consortium, 2004.
14. Ian Jacobs and Norman Walsh. *Architecture of the World Wide Web, Volume One (W3C Recommendation 15 December 2004)*. World Wide Web Consortium, 2005. Available at `http://www.w3.org/TR/webarch/`.
15. David R. Karger. Haystack: Per-User Information Environments Based on Semistructured Data. In Victor Kaptelinin and Mary Czerwinski, editors, *Beyond the Desktop Metaphor*, pages 49–100. Massachusetts Institute of Technology, 2007.
16. Wolfgang Nejdl, Boris Wolf, Changtao Qu, Stefan Decker, Michael Sintek, Ambjörn Naeve, Mikael Nilsson, Matthias Palmér, and Tore Risch. EDUTELLA: A P2P Networking Infrastructure Based on RDF. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 604–615, New York, NY, USA, 2002. ACM Press.
17. Open Mobile Alliance. *OMA Data Synchronization V1.2.1*, 2007. Available at `http://www.openmobilealliance.org/Technical/release_program/ds_v12.aspx`.
18. Leo Sauermann, Ansgar Bernardi, and Andreas Dengel. Overview and Outlook on the Semantic Desktop. In Stefan Decker, Jack Park, Dennis Quan, and Leo Sauermann, editors, *Proceedings of the 1st Semantic Desktop Workshop*, volume 175, Galway, Ireland, November 2005. CEUR Workshop Proceedings.
19. Bernhard Schandl. SemDAV: A File Exchange Protocol for the Semantic Desktop. In *Proceedings of the Semantic Desktop and Social Semantic Collaboration Workshop*, volume 202, Athens, GA, USA, November 2006. CEUR Workshop Proceedings.
20. Hong-Siang Teo. An Activity-driven Model for Context-awareness in Mobile Computing. In *MobileHCI '08: Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pages 545–546, New York, NY, USA, 2008. ACM.
21. Giovanni Tummarello, Christian Morbidoni, Joackin Petersson, Paolo Puliti, and Francesco Piazza. RDFGrowth, a P2P Annotation Exchange Algorithm for Scalable Semantic Web Applications. In Ilya Zaihrayeu and Matteo Bonifacio, editors, *P2PKM*, volume 108 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2004.
22. Denny Vrandecic and York Sure. How to Design Better Ontology Metrics. In *Proceedings of the 4th European Semantic Web Conference (ESWC2007)*, 2007.