

BUILDING AND QUERYING SEMANTIC LAYERS FOR WEB ARCHIVES



Vaibhav Kasturia, Pavlos Fafalios, Helge Holzmann and Wolfgang Nejdl

PROBLEM

Despite the increasing number of web archives worldwide, the absence of efficient and meaningful exploration methods still remains a major hurdle in the way of turning web archives into a usable and useful source of information.

For a bit more complex information needs, like exploratory search needs, keywordbased search in web archives leads to ineffective interactions and poor results. Thus, there is the need to go beyond keywordbased search and support more advanced information seeking strategies.

CONTRIBUTIONS

We make the following contributions:

- We introduce a simple but flexible RDF/S data model, called Open Web Archive,
- 2. We detail the process of constructing semantic layers and we present an open source and distributed framework, called ArchiveSpark2Triples, that facilitates their efficient construction.
- 3. We present (and make publicly available) three semantic layers for three different types of web archives: one for a versioned web archive, one for a nonversioned news archive, and one for a social-media archive.
- 4. We present the results of a comparative evaluation using a set of 20 information needs of exploratory nature.

RESULTS

We have introduced a model and a framework for describing and publishing metadata and semantic information about web archives. The constructed semantic layers allow:

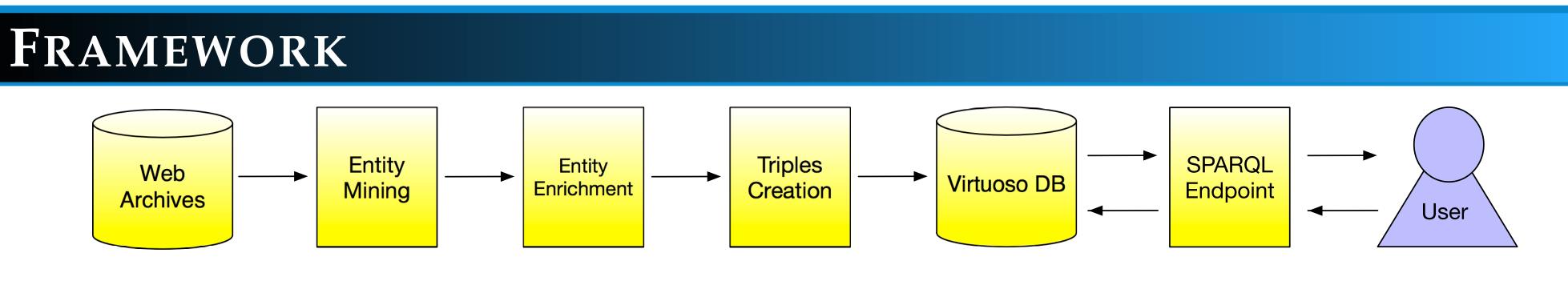
- 1. exploring web archives in a more advanced way based on entities, events and concepts extracted from the archived documents and linked with web resources.
- 2. integrating information coming from multiple knowledge bases and semantic layers.
- 3. inferring new knowledge that is very laborious to derive otherwise.
- 4. coping with common problems when exploring web archives like temporal reference variants and multilinguality.
- 5. making the contents of web archives machine understandable, thereby enabling their direct exploitation by other systems and tools.

The results showed that semantic layers can answer complex information needs that keyword-based search systems fail to satisfy.

DATA MODEL oae:position dc:date rdfs:Literal dc:date rdfs:Literal oae:donfidence rdfs:Literal dc:Event dc:title rdfs:Literal oae:Entity oae*:score* rdfs:Literal skos:Concept dc:format rdfs:Literal oae:detectedAs schema:mentions rdfs:Literal oa:hasBody owa:ArchivedDocument oae:hasMatchedURI owa:VersionedDocument oa:Annotation dc:hasVersioi rdfs:Resource oa:hasTarget oa:hasTarget owa: http://l3s.de/owa/ owa:firstCapture rdfs:Literal oa: http://www.w3.org/ns/oa# rdfs:Resource oae: http://www.ics.forth.gr/isl/oae/core# owa:lastCapture dc:references rdfs:Literal dc: http://purl.org/dc/terms/ owa:ArchivedDocument skos: http://www.w3.org/2004/02/skos/core# owa:numOfCaptures schema: http://schema.org/ owa:VersionedDocument rdfs:Literal rdfs: http://www.w3.org/2000/01/rdf-schema# The Open Web Archive data model dc:date dbo:TennisPlayer 06.01.2012 06:40 _dc:references dc:format rdf:type "text/html" dc:references dbr:Roger_Federer dc:title "An example Page" oae:hasMatchedURI dc:references oae:position schema:mentions dc:references oae:confidence 0.9 oae:Entity rdf:type "Federer" oae:detectedAs http://www.nytimes.com/... owa: http://l3s.de/owa/ schema:mentions rdf:type dc: http://purl.org/dc/terms/ rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# schema: http://schema.org/ dc:Event owa:ArchivedDocument rdf:type oae: http://www.ics.forth.gr/isl/oae/core# Describing an archived article (non-versioned) using the Open Web Archive data model dc:date dbo:TennisPlayer 06.01.2012 06:40 _dc:references dc:format rdf:type "text/html" _dc:references dc:title dbr:Roger_Federer "An example Page" http://...dc:references oae:hasMatchedURI schema:mentions oae:position ••• < dc:references oae:confidence 0.9 oae:Entity rdf:type "Federer" http://www.nytimes.com/... owa: http://l3s.de/owa/ schema:*mentions*

Describing an archived web page (versioned) using the Open Web Archive data model

dc:Event



The process of construction and querying of Semantic Layers

FUTURE WORK

1. Regarding future work and research, user-friendly interfaces should be developed on top of semantic layers for allowing end-users to easily and efficiently explore web archives.

rdf:type

owa:ArchivedDocument

2. Another interesting direction is to study approaches for ranking the results returned by SPARQL queries.

SOURCE CODE

The Semantic Layers are publicly available at: http://l3s.de/ owa/semanticlayers/

dc: http://purl.org/dc/terms/

schema: http://schema.org/

rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#

oae: http://www.ics.forth.gr/isl/oae/core#

The dataset used for evaluation is available at: http://l3s. de/owa/semanticlayers/ SemLayerEval.zip



