Vol. (1) No. (1) March 2018

JK Research Journal in Mathematics and Computer Sciences

DBpedia-An Advancement Towards Content Extraction From Wikipedia

Neha Jain

Government Degree College R.S Pura, Jammu, J&K

Abstract: DBpedia is the research product of the efforts made towards extracting structured content from the data obtainable from the largest web encyclopaedia i.e. Wikipedia. It is a project aiming towards making the information machine readable. DBpedia makes available the content in the form of RDF (Resource Description Framework) triplets which can be easily queried using query languages like SPARQL. The paper highlights this query making procedure. The research work also includes the role of DBpedia in Linked Data initiative.

Keywords: *DBpedia; Wikipedia; SPARQL; Jena; Knowledge Extraction.*

Introduction

DBpedia is the outcome of work undertaken towards extracting structured content form the content which is available on the largest web encyclopedia i.e. Wikipedia. It is a project aiming towards making the information machine readable. DBpedia provides the content in the form of RDF triplets which can be easily queried using query languages like SPARQL. RDF is a flexible data model for representing extracted content and for publishing it on the Web. A number of end points are available which can be used to query DBpedia. In the presented research work, DBpedia data has been queried and processed using Apache Jena toolkit in JAVA. Apache Jena is a free and open source framework for building Semantic Web and Linked Data Applications[1]. The paper highlights this query making procedure. The research work also includes the role of DBpedia in Linked Data initiative.

About DBpedia

The DBpedia initiative has successfully captured a richly loaded knowledge base from its parent site Wikipedia. This knowledge base serves as Linked Data for other initiatives on the Web. Because of its volume, DBpedia is considered to be at the core of the Linked Data Cloud. DBpedia knowledge base to the present date provides 274 million pieces of information pertaining to 2.6 million concepts. As DBpedia is highly diverse and expanded over a wide range of domains, it has a considerable amount of overlap with various other data-sets already available on the World Wide Web. Many data publishers have taken to linking their data sources to DBpedia to enhance interoperability and accessability of data. This has made DBpedia occupy a central position

j.neha10@gmail.com
* Neha Jain

on the Linked Open Data cloud. The BBC is one such organization working in this direction. BBC is working towards using DBpedia as the semantic backbone and controlled vocabulary for its various domains.

Related work

In their work Christian Bizer et.al. described the extraction of the DBpedia knowledge base, the urgent status of interlinking DBpedia with other sources on the World Wide Web and also presented a general idea of various applications that assist the Linked Open Data concept with DBpedia. [2]

S. Auer et al. [3] described the extraction of the DBpedia datasets, and how the resulting information is published on the Web for human- and machine-consumption. They also highlighted some promising applications from the DBpedia community and also presented how website authors can facilitate DBpedia content within their sites. They also presented the current status of interlinking DBpedia with other open datasets on the Web and outlined the method for DBpedia to serve as a nucleus for an emerging Web of open data.

G. Kobilarov et al. [4] described how the BBC is working to integrate data and linking documents across BBC domains by using Semantic Web Technologies and Linked open data and in particular DBpedia.

LehmannJens et al. [5] presented an overview of the DBpedia community project including its architectur, technical implementation, maintenance, internationalization, usage statistics and applications.

S. Hellmann et al. [6]described tools and techniques to use existing techniques to achieve scalability on large knowledge bases available as SPARQL endpoints or Linked Data. The algorithms developed by them are made available in the open source DL-Learner project and can be used in real-life scenarios by Semantic Web applications.

DBpedia Knowledge Extraction Model

Wikipedia is an online encyclopedia which mostly consists of free text to be used by human users. It also contains some information in structured form. Structured text is in the form of infobox templates, categorization information, images, geo-coordinates, hyperlinks etc. The DBpedia Project extracts the updated structured information from Wikipedia and compiles it into a knowledge base.

The DBpedia Ontology is a thin, cross-domain ontology, created by hand-picking information from infoboxes within Wikipedia. The ontology presently consists of 685 classes, arranged in a sub-sumption hierarchy explained by 2,795 different properties. DBpedia presently consists of eleven extractors which process the following types of Wikipedia content:

Labels: The title of the Wikipedia articles can be accessed through this as rdfs: label.

Abstracts: a short abstract which is generally the first paragraph .This can be accessed using the property DBpedia: abstract.

Interlanguage links: These are the links that connect articles about the same topic in different language editions of Wikipedia. These links assign labels and abstracts in varied languages to resources available in DBpedia.

Images: Images are represented as foaf: depiction property. These are the links pointing to resources at Wikimedia Commons images.

Redirects: The redirection links in Wikipedia are extracted in DBpedia and are used to resolve references between DBpedia resources.

Disambiguation: The disambiguation pages of Wikipedia are used to explain the different meanings of similarly spelled words. The disambiguation links are extracted and represented using predicate DBpedia: diambiguates.

External links: The links to external articles which are present on Wikipedia pages are represented in DBpedia using the property DBpedia: reference. owl: same As property is used to depict RDF links.

Pagelinks: All the links between Wikipedia articles are extracted and represented in DBpedia as

DBpedia: wikilink property.

Homepages: The links to the homepages of entities such as companies and organizations are extracted using this property. It is represented as foaf: homepage.

Categories: Wikipedia articles are arranged in categories, which are represented in DBpedia using the SKOS shared vocabulary.[7] Categories are represented as skos: concepts and the relationships between categories are represented as skos: broader.

Geo-coordinates: DBpedia contains coordinates extracted from info-boxes. Besides that, it also contains geo-coordinated for 1,094,000 geographic locations[8]. World Wide Web Consortium (W3C) Basic Geo Vocabulary is used for expressing the Geo-coordinates.

The model uses Apache Jena Toolkit in JAVA and SPARQL as the query language to gather information from DBpedia datasets. The Jena API which is a free and open source Java framework for building Semantic web and Linked Data applications is used in this work.

In Jena, a data structure called Model encapsulates all state information provided by a collection of RDF triples. The model, in principle, denotes an RDF graph as it contains a collection of RDF nodes attached to each other by labeled relations.

The query language used is a fourth generation query language called SPARQL. It is "data-oriented" language as it only queries and reverts back the information held in the models. The language does not offer any inferencing or reasoning capability.

In the proposed research work, Apache Jena 3.0 framework is used. The API provided by the Jena Framework are imported and used to construct SPARQL queries, call and return results from datasets from DBpedia .

The queries are of the form:

String querystr= "PREFIX dbo:<http://DBpedia.org/ontology/>"+"PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>" +"SELECT ?x"+" WHERE { "+" <http://DBpedia.org/resource/India> rdfs:label ?x . " +" FILTER (lang(?x) = 'en')}";

This query returns the label property value.

String querystr= "PREFIX dbo:<http://DBpedia.org/ontology/>"+"PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>" +"SELECT ?x"+" WHERE { "+" <http://DBpedia.org/resource/India> dbo:abstract ?x . " +" }";

This query returns the abstract property value. Since, there is no language filter, the query returns all the available muti-lingual data.

Conclusion

Jena is a promising development environment as it provides a powerful API to exploit the DBpedia datasets. But it is still not completely robust. It does not provide support to a number of properties like dbo:disambiguates, dbo:wikiPageRedirects, redirection links to images etc. This is not only a limitation with Jena but it is observed that not all localized editions provide fully robust publically available SPARQL endpoints, nor do all localized URI's reference. This is acceptable as the field of linked data is still in the development stages and as more and more local DBpedia research initiatives come up in different parts of the world, the effort to internationalize DBpedia will start to get success.

DBPedia is a promising initiative towards making the World Wide Web intelligent and machine accessible.

References

[1] "https://jena.apache.org/," [Online].

[2] C. Bizer, J. Lehmannb, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann, "DBpedia-A Crystallization Point for the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web,* pp. Volume 7, Issue 3, September 2009, Pages 154–165, September 2009.

[3] S. Auer, C. Bizer, G. Kobilaro, J. L. R. Cyganiak and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," *The Semantic Web,* pp. 722-735, 2007

[4] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer and R. Lee, "Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections," *The Semantic Web: Research and Applications. ESWC 2009. Lecture Notes in Computer Science, vol 5554. Springer, Berlin, Heidelberg,* pp. 723-737, 2009.

[5] LehmannJens, I. Robert, J. Max, J. Anja, K. Dimitris, M. P. N., H. Sebastian, M. Mohamed, K. P. van, A. Sören and B. Christian, "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia," *Journal: Semantic Web, vol. 6, no. 2,* pp. 167-195, 2015

[6] S. Hellmann, J. Lehman and S. Auer, "Learning of OWL class descriptions on very large knowledge bases," *ISWC-PD'08 Proceedings of the 2007 International Conference on Posters and Demonstrations - Volume 401,* pp. Pages 102-103, 2007

[7] "https://www.w3.org/2004/02/skos/," [Online].

[8] "http://wiki.DBpedia.org/services-resources/datasets/data-set-39," [Online].

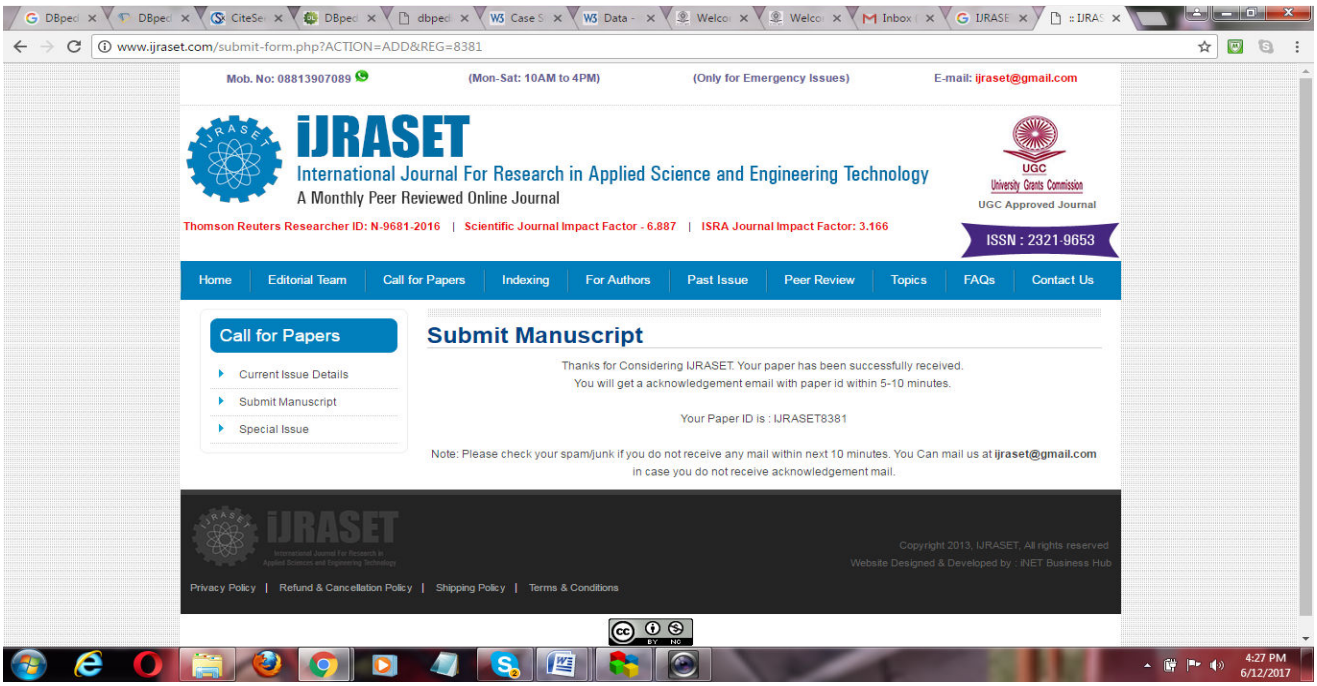
www.ijraset.com/submit-form.php?ACTION=ADD&REG=8381
Mob. No: 08813907089
(Mon-Sat: 10AM to 4PM)
(Only for Emergency Issues)
E-mail: ijraset@gmail.com
iJRASET
International Journal For Research in Applied Science and Engineering Technology
A Monthly Peer Reviewed Online Journal
UGC
University Grants Commission
UGC Approved Journal
Thomson Reuters Researcher ID: N-9681-2016 | Scientific Journal Impact Factor - 6.887 | ISRA Journal Impact Factor: 3.166
ISSN : 2321-9653
Home
Editorial Team
Call for Papers
Indexing
For Authors
Past Issue
Peer Review
Topics
FAQs
Contact Us
Call for Papers
Current Issue Details
Submit Manuscript
Special Issue
Submit Manuscript
Thanks for Considering IJRASET. Your paper has been successfully received.
You will get a acknowledgement email with paper id within 5-10 minutes.
Your Paper ID is : IJRASET8381
Note: Please check your spam/junk if you do not receive any mail within next 10 minutes. You Can mail us at ijraset@gmail.com in case you do not receive acknowledgement mail.
Copyright 2013, IJRASET, All rights reserved
Website Designed & Developed by : iNET Business Hub
Privacy Policy | Refund & Cancellation Policy | Shipping Policy | Terms & Conditions
4:27 PM
6/12/2017

Fig 1