International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr42018 www.irjet.net p-ISSN: 2395-0072

DATA RETRIEVAL USING MASTER RESOURCE DESCRIPTION FRAMEWORK

Lavanya .R, Manjula .M

Abstract - *The proliferation of heterogeneous Linked Data on the Web poses new challenges to database systems. In particular, the capacity to store, track and query provenance in data is becoming a pivotal feature of modern triplestores. We present methods by extending a native RDF store to efficiently handle the storage, tracking, and querying of provenance in RDF data by using the concept of Master Resource Description Framework. We describe a reliable and understandable specification of the way results are derived from the data and how particular pieces of data are combined to answer a query. This project further aims to give a personalized experience for the users by using a user profile which tracks the user interests and produce results relatively to their interests. Subsequently, we present techniques to tailor queries with provenance data. Finally, we show how using Master Resource Description Framework improves the efficiency of search while retrieving data from databases*.

Keywords— database, linked data,Master RDF, provenance, query, RDF,triplestores

I. INTRODUCTION

As Database Retrieval process is heavy weighted and Time Consuming, it needs to be addressed by RDF-based approach. Every query represents a graph pattern consisting of a set of triple patterns representing the sets of distinguished variables, undistinguished variables, and constants. Hence Retrieval based on query search also meaningless. A solution to a graph pattern q on a graph G is a mapping from the variables in q to vertices in G such that the substitution of variables would yield a sub graph of G. The substitutions of distinguished variables constitute the answers. In fact, It can be interpreted as a homomorphism (i.e., a structure preserving mapping) from the query graph to the data graph.

II. PROBLEM DEFINITION

The existing system fails to effectively handle the massive amounts of files. Also the Database retrieval is a heavy wighted process [11]. This in turn results in slow processing. Cost of querying is also high. These conditions result to other effects like high access latency, not supporting the user run time profiling and also not supporting user preference based queries. Above stated points define the problem for our project. These points are taken forward and solved to the possible extent to make a better retrieval system.

III. RESOURCE DESCRIPTION FRAMEWORK

The Resource Description Framework (RDF) is a general framework for how to describe any Internet resource such as a Web site and its content. An RDF description (such descriptions are often referred to as metadata, or "data about data") can include the authors of the resource, date of creation or updating, the organization of the pages on a site (the sitemap), information that describes content in terms of audience or content rating, key words for search engine data collection, subject categories, and so forth [1]. The Resource Description Framework will make it possible for everyone to share Web site and other descriptions more easily and for software developers to build products that can use the metadata to provide better search engines and directories, to act as intelligent agents, and to give Web users more control of what they're viewing[3]. The RDF is an application of another technology, the Extensible Markup Language (XML), and is being developed under the auspices of the World Wide Consortium (W3C)[1]. A certain amount of metadata is already provided for Web site resources using the Hypertext Markup Language (HTML). For example, when we wrote this page, we added HTML statements containing key words that describe the content of this definition and that are used by search engines for indexing. This RDF is generated by using an API called Jena API which happens to be an open source framework helping in the generation of RDF files [2].

A. Basics of RDF

An Internet resource is defined as any resource with a Uniform Resource Identifier (URI). This includes the Uniform Resource Locators (URL) that identify entire Web sites as well as specific Web pages[1]. As with today's HTML META tags, the RDF description statements, encased as part of an Extensible Markup Language (XML) section, could be included within a Web page (that is, a Hypertext Markup Language - HTML - file) or could be in separate files. RDF is now a formal W3C Recommendation, meaning that it is ready for general use [1]. Currently, a second W3C recommendation, still at the Proposal stage, proposes a system in which the descriptions related to a particular purpose would constitute a class of such like descriptions (using *class* here much as it is used in object-oriented programming data modeling and programming). Such classes could fit into a *schema* or hierarchy of classes, with subclasses of a class able to inherit the descriptions of the entire class [4]. The schema of classes proposal would save having to repeat descriptions since a single reference to the class of which a particular RDF description was a part would suffice [2]. The scheme or description of the collection of classes could itself be written in RDF language.

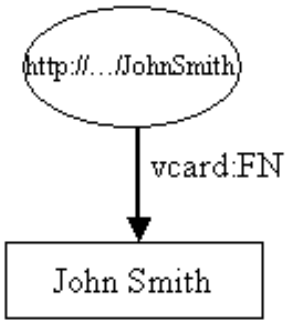

Fig. 1. Diagramatic representation of Resource Description Framework

B. Working of RDF

RDF is a structure oriented approach that exploits the structure patterns exhibited by the underlying data captured using a structure index. For capturing the structure of the underlying data, structure index is used, a concept that has been successfully applied in the area of XML- and semi structured data management [3]. It is basically a graph, where vertices represent groups of data elements that are similar in structure. For constructing this index, consider structure patterns that exhibit certain edge labels containing path.. Further, RDF data partitioning is carried on. To obtain a contiguous storage of data elements that are structurally similar, vertices of the structure index are mapped to tables [12]. The triples with the same property label, triples with subjects that share the same structure are physically grouped. A basic strategy is to match the query against the structure index first to identify groups of data that satisfy the query structure [10]. Then, via standard data-level processing, data in these relevant groups are retrieved and joined. Instead of performing structure- and data-level operations successively and independent from each other like in this basic strategy, an integrated strategy that aims at an optimal combination of these two types of operations is used.

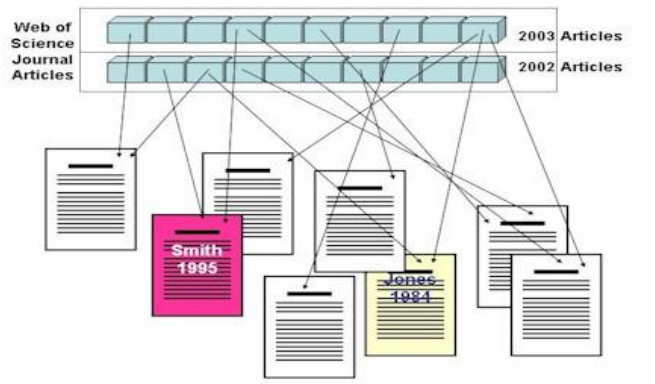

Fig. 2. Working of Resource Description Framework

C. Benefits

- By providing a consistent framework, RDF will encourage the providing of metadata about Internet resources.
- Because RDF will include a standard syntax for describing and querying data, software that exploits metadata will be easier and faster to produce.
- The standard syntax and query capability will allow applications to exchange information more easily.
- Searchers will get more precise results from searching, based on metadata rather than on indexes derived from full text gathering.
- Intelligent software agents will have more precise data to work with.

D. Example

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cd="http://www.recshop.fake/cd#">

<rdf:Description
rdf:about="http://www.recshop.fake/cd/Empire Burlesque">
 <cd:artist>Bob Dylan</cd:artist>
 <cd:country>USA</cd:country>
 <cd:company>Columbia</cd:company>
 <cd:price>10.90</cd:price>
 <cd:year>1985</cd:year>
</rdf:Description>

<rdf:Description
rdf:about="http://www.recshop.fake/cd/Hide your heart">
 <cd:artist>BonnieTyler</cd:artist>
 <cd:country>UK</cd:country>
 <cd:company>CBS Records</cd:company>
 <cd:price>9.90</cd:price>
 <cd:year>1988</cd:year>
</rdf:Description>
.
.
.
</rdf: RDF>
```

IV. MASTER RESOURCE DESCRIPTION FRAMEWORK

Master resource description framework extends the concept of native RDF. It aims to combine multiple RDF files and use a single master RDF for the purpose of retrieving contents. These master RDFs can be a combination of any number of RDFs based on their content size [6]. This project aims to improve the efficiency while searching by using

master RDF concept. This is done by reducing the time taken for the search engine to go and search in each of the RDFs as the content is combined and stored in a Master RDF [7]. The efficiency of which is shown theoretically at the end.

A. Architecture Diagram

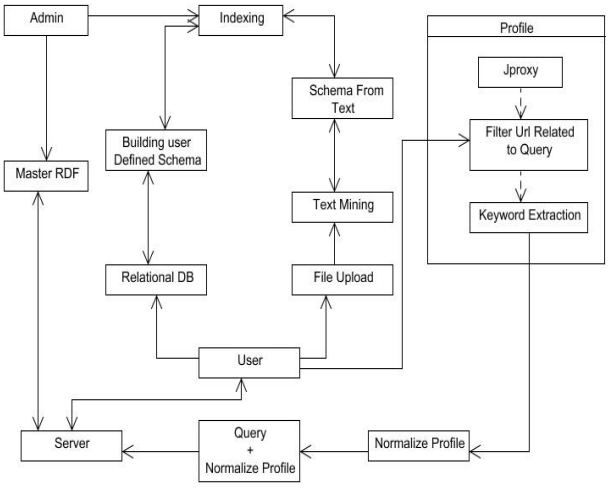

Fig. 3. Architecture diagram

V. USER PROFILE

User profile consists of a user behavior. Each user may have the difference in preferences. First user need to create a profile and add the preference. Whenever the user logins, the proxy will be enabled automatically and runtime profiling will start. Consistent with many previous works in personalized runtime user profiling is failed to achieve. Moreover, our profile is constructed based on the availability of a public accessible taxonomy. User profile is constructed based on the sample taxonomy repository.

VI. RDF FOR USER UPLOADED DATA

The RDF is also generated by mining the text contents uploaded by the users in blogs and the contents of the file are analyzed and the Meta contents are manipulated. The Meta contents are the key for search process so that the file can be rendered on demand [4]. Text mining process analyses the text word by word and also picks up the literal meaning behind the group of words that constitute the sentence [8]. The Words are analyzed in WordNet API so that the related terms can be found for use in the Meta content in generation of RDF.

VII. MASTER RDF GENERATION

Similar data's are grouped together that relate to the same resource. The data level process is subjected to structure level processing by indexing the semantic data elements. Multiple RDFs are grouped and structured together to form a master RDF data that holds all the semantic information's of a Server that support reasoning in any formats of query processing[9]. The different resources are interlinked with high degree of relational factors by the predicates in the triples [4]. The query processing is handled directly in the RDF file by iterating the triples forming a discrete relation with the service query and the URI representing the location of the resource is returned, as this process is handled in web services in real time servers. Hence the structure-oriented approach to RDF data management where data partitioning and query processing make use of structure patterns generated by the RDF [6].

VIII. QUERYING OVER MASTER RDF WITH PROFILING

When a user issues a query on the client, the proxy generates a user profile in runtime in the light of query terms. The output of this step is a generalized user profile satisfying the privacy requirements. Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search [12]. The search results are personalized with the profile and delivered back to the query proxy. Finally, the proxy either presents the raw results to the user, or reranks them with the complete user profile. As the sensitivity values explicitly indicate the user's privacy concerns, the most straightforward privacy preserving method is to remove sub-trees rooted at all sensitive-nodes whose sensitivity values are greater than a threshold[5]. Such method is referred to as forbidding.

IX. PERFORMANCE MEASURE

Assume there are two RDF files with one RDF file containing database content and another being a text file. Let us consider the following:

N=time taken to search RDF file 1

M=time taken to search RDF file 2

S=time taken to navigate from RDF file1 to RDF file2

Now using native RDF file to perform search will be:

Time complexity=O (N+S+M)

Using Master RDF concept for searching will give:

Time complexity=O (N+M)

From the above measures it is evident that the time taken to perform search will reduce by a factor of S which represents the time taken to navigate between RDF files.

X. CONCLUSION

Hence we propose a system for retrieval of data by using the concept of Master Resource Description Framework which acts as an index for all the underlying RDF. This reduces the

time taken to switch between multiple RDF's and also strives to provide personalized experience to the user which helps them in retrieving the exact data that they want. Thus this system brings about a considerable change in the efficiency of retrieval and improves the performance of the data system.

REFERENCES

[1] D. Wood, R. Cyganiak and M. Lanthaler, Eds., "RDF 1.1 Concepts and Abstract Syntax," W3C Recommendation, Feb. 2014. [Online]. Available: http://www.w3.org/TR/rdf11-concepts/

[2] M. Schmachtenberg, C. Bizer, and H. Paulheim, " Adoption of the linked data best practices in different topical domains," in The Semantic Web, P. Mika, et al., Eds. Berlin, Germany: Springer, 2014, pp. 245–260. [Online]. Available: http://dx.doi.org/10.1007/978- 3-319-11964-9_16

[3] G. Grimnes, "BTC2012 stats," 2012. [Online]. Available: http:// gromgull.net/blog/2012/07/some-basic-btc2012-stats/

[4] P. Groth and W. Beek, "Measuring PROV provenance on the Web of data," 2016. [Online]. Available: https://nbviewer.jupyter.org/ github/pgroth/prov-wod-analysis/blob/master/ Measuring PROV Provenance WebofData.ipynb

[5] M. Wylot, J. Pont, M. Wisniewski, and P. Cudr ?e-Mauroux, "dipLODocus[RDF]: Short and long-tail RDF analytics for massive webs of data," in Proc. 10th Int. Conf. Semantic Web, 2011, pp. 778–793. [Online]. Available: http://dl.acm.org/citation.cfm?id=2063016.2063066

[6] M. Wylot and P. Cudr?e-Mauroux, "DiploCloud: Efficient and scalable management of RDF data in the cloud," IEEE Trans. Knowl. Data Eng., vol. 28, no. 3, pp. 659–674, Mar. 2016.

[7] M. Wylot, P. Cudre-Mauroux , and P. Groth, "TripleProv: Efficient processing of lineage queries in a native RDF store," in Proc. 23rd Int. Conf. World Wide Web, 2014, pp. 455–466.

[8] M. Wylot, P. Cudr?e-Mauroux, and P. Groth, "Executing provenance- enabled queries over Web data," in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 1275–1285. [14] M. Wylot, P. Cudr?e-Mauroux, and P. Groth, "A demonstration of TripleProv: Tracking and querying provenance over Web data," Proc. VLDB Endowment, vol. 8, no. 12, pp. 1992–1995, 2015.

[9] L. Moreau, "The foundations for provenance on the Web," Found. Trends Web Sci., vol. 2, no. 2/3, pp. 99–241, Nov. 2010. [Online]. Available: http://eprints.ecs.soton.ac.uk/21691/

[10] J. Cheney, L. Chiticariu, and W.-C. Tan, Provenance in databases: Why, How, and Where, vol. 1, no. 4. Breda, The Netherlands: Now Publishers Inc, 2009.

[11] P. Groth, Y. Gil, J. Cheney, and S. Miles, "Requirements for provenance on theWeb," Int. J. Digit. Curation, vol. 7, no. 1, pp. 39–56, 2012.

[12] P. Cudr?e-Mauroux, et al., "A demonstration of SciDB: A scienceoriented DBMS," Proc. VLDB Endowment, vol. 2, no. 2, pp. 1534– 1537, 2009.