International Journal of Communication and Networking System
Volume: 05 Issue: 02 December 2016, Pages No.113-1116
ISSN: 2278-2427

# Semantic Digital Library

Subair KB[1], Nihal Abdulla PT[2], Shibin TV[3], Praveen A [4]Saidalavi Kaladi[5]
[1] Student, Department of CSE, National Institute of Technology, Calicut
[2] Student, Department of CSE, National Institute of Technology, Calicut
[3] Student, Department of CSE, National Institute of Technology, Calicut
[4] Student, Department of CSE, National Institute of Technology, Calicut
[5]HOD, Associate Professor, Department of CSE, National Institute of Technology, Calicut

Abstract — The present search methods are usually keyword based or syntactic in nature. The syntactic search may not always produce relevant results. Unwanted results which contain just the keyword in the document also will be displayed as result. This is why the Semantic web technology was introduced. Here we are implementing the semantic techniques in Digital Library by keeping meta data of all the books in ontology files. This will help the users to find relevant books which contain a specific topic or make a complex query rather than search using a simple keyword.

Keywords : Semantic Web; RDF; Ontology; OWL; SPARQL;

## I. INTRODUCTION

Digital library, most of which are based on keyword search, lacks expressiveness which the user expects. The application of semantic techniques in the field of digital library provides a better way to search for what the user wants with more ease, since the user gets more expressive power as the system can take more complex queries. In semantics, rather than just storing the book name, author name etc., relations are specified. By this project we are aiming to take a step forward in the field of Digital Library.

**SEMANTIC WEB:** Web 2.0 is mainly focused on people, while Semantic Web is focused on machines. For computer it's impossible to perform the tasks required to find, search and analyse its information without human guidance since web pages are designed for human readers particularly. The Semantic Web is a project that aims to change this view by presenting Web page data in such a way that it is understood by computers, enabling machines to do searching, assembling and combining of the Web's information without a human operator [1]. Semantic search, as an application of Semantic Web in the area of information retrieval, has shown significant potential in the function of improving the performance of information retrieval. When Compared with the traditional search engines that focus on the frequency of word appearance, semantic search engines are more likely to try to understand the meanings hidden in retrieved documents and users queries, by means of adding semantic tags into texts, in order to structuralize and conceptualize the objects within documents [2].

**RDF**: RDF (Resource Description Framework), SPARQL and OWL are three basic Semantic Web technologies. Specifically, RDF is basic building block of the Semantic Web. RDF representation is given in Fig 1.

1. A fact is expressed as a triple of the form (Subject, Predicate, and Object).
2. Subjects, predicates, and objects are names for entities.
3. Objects can also be text values, called literal values [3].

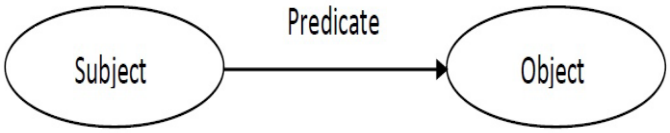


Fig. 1. RDF Representation

**ONTOLOGY**: Ontology is considered as one of the pillars of the Semantic Web. Ontology is a formal, explicit specification of a shared conceptualization [4]. A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are supposed to exist in some area of interest and the relationships that hold among them. A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose [5].

**OWL**: Instead of just presenting information to humans, Web Ontology Language (OWL) is designed for the applications that need to process the content of information. The OWL facilitates greater machine interpretability of web content than that supported by RDF Schema RDF-S), XML and RDF (by providing additional vocabulary along with a formal semantics [6].

## II. LITERATURE SURVEY

The purpose of literature survey is to provide knowledge on concepts and ideas established in the field of semantic web and digital library.

### A. Semantic Digital Library

Semantic Digital Library systems are built upon research on digital libraries, semantic web, social networking and human computer interaction: they integrate knowledge organization systems, delivered by classic digital libraries, with the semantic web and social networking (Web 2.0) technologies. Semantic Web technologies support the interoperability with other services and expressiveness of annotations. The Web 2.0 approach allows users to be engaged in the annotation and knowledge sharing process, making semantic digital libraries more useable [7].

### B. Kosmix Semantic Search

Kosmix provides with a dashboard of search results using trusted branded content creation and distribution platforms to help users see, read, and engage with the search results in a much more organized and context-sensitive way. Their dashboard is different for almost every search done on Kosmix, providing with the most relevant data and information for the terms searched for [8].

International Journal of Communication and Networking System
Volume: 05 Issue: 02 December 2016, Pages No.113-1116
ISSN: 2278-2427

C. Simplicity

Simplicity (Semantics-sensitive Integrated Matching for Picture Libraries), an image retrieval system, which uses a wavelet-based approach for feature extraction, integrated region matching based upon image segmentation and semantics classification methods. A metric for the overall similarity between images is developed by using a region-matching scheme that takes into account all properties of all the regions in the images [9].

## III. TOOLS

A. Apache Open NLP

The Apache Open NLP library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution. These tasks are usually required to build more advanced text processing services. OpenNLP also included maximum entropy and perceptron based machine learning [10].

B. Sesame

Sesame is used for storage and querying of RDF data. It is is an open source Java framework. The framework is fully extensible and configurable with respect to query languages, RDF file formats, inferences, storage mechanisms and query result formats. Out of the box, Sesame supports SPARQL and SeRQL querying, a memory-based and a disk-based RDF store and RDF Schema inferences. It also supports most popular RDF file formats and query result formats [11].

C. Sparql

RDF query language SPARQL used for databases, able to retrieve and manipulate data stored in Resource Description Framework format [12]. It was made a standard by the Resource Description Framework (RDF) Data Access Working Group (DAWG) of the World Wide Web Consortium, and is identified as one of the key technologies of the semantic web. SPARQL allows for a query to consist of disjunctions, optional patterns, conjunctions and triple patterns [14].

D. SQL

Structured Query Language (SQL) is used to query data from a database. As per ANSI, it is the standard query language for RDBMS (Relational Database Management Systems). SQL statements are used to perform tasks such as insert data on a database or retrieve data from a database or update data on a database. The standard SQL commands such as "Select", "Delete", "Update", "Insert", "Drop" and "Create" can be used to achieve almost everything that one needs to do with a database [16].

E. Wordnet

A lexical database of English WordNet grouped Nouns, verbs, adjectives and adverbs are into synsets (sets of cognitive synonyms), each expressing a unique concept. Synsets are interlinked by means of lexical relations and conceptual semantic. The resulting network of meaningfully related concepts and words can be navigated with the browser [15].

## IV. DESIGN

A. Database Design

A SQL database Test DB is created. This database contains two tables named BOOK DATA and Stored Queries. Book Data in the database. It stores the details of a book like book name, author name, publisher and number of available copies etc. Stord Queries stores the SPARQL queries corresponding to the different user query patterns. This table stores the hash value corresponding to the pattern. It also stores the order in which unknowns in the query occur, the number of variable in the query.

B. Sesame Database

Sesame database is used to store the RDF data. It provides special features for storing and querying the RDF data with SPARQL query. A RDF data entry function is used to feed the sesame database with the RDF files.

C. Module

Sesame database is used to store the RDF data. It provides special features for storing and querying the RDF data with SPARQL query. A RDF data entry function is used to feed the sesame database with the RDF files.Fig. 2 shows the brief design of the model. In first step a user query (in natural language) is entered which is processed and its corresponding SPARQL query is generated. Results from the SPARQL query is used to query in the RDBMS to obtain the final details about the book.

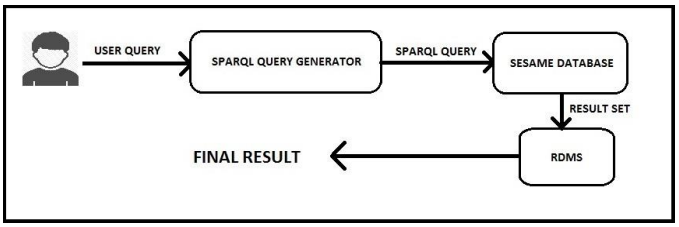


Fig. 2. Model Design

A. Query Generator

The user types the query. This query will be processed. The user query undergoes some processing steps and finally generates a SPARQL query. This includes stop words removal, synonym finding etc. And after matching the processed query with a stored pattern we will get a SPARQL query. This is shown in Fig. 3.

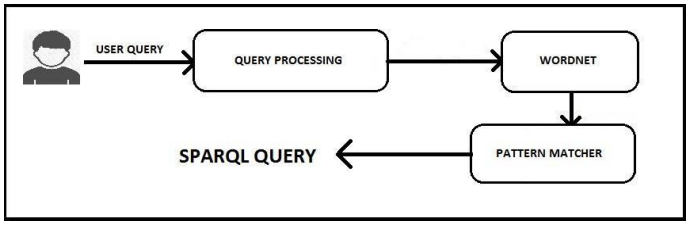


Fig. 3. Query Generator

B. Sesame Database Querying

The SPARQL query obtained from the previous step will be used to query the SESAME database. The result of this step

International Journal of Communication and Networking System
Volume: 05 Issue: 02 December 2016, Pages No.113-1116
ISSN: 2278-2427

contains a book name or author name according to the user query. This is shown in Fig. 4.

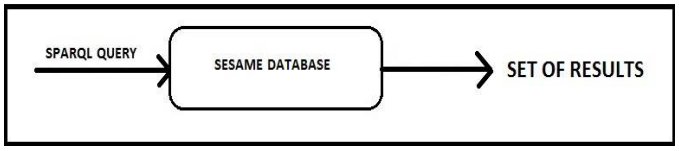


Fig. 4. SPARQL Query Processing

C. RDBMS Querying

The results obtained after querying SESAME database is used here. A SQL query is created using this result and the RDBMS is queried. The final result will be published to the user. This is shown in Fig. 5.

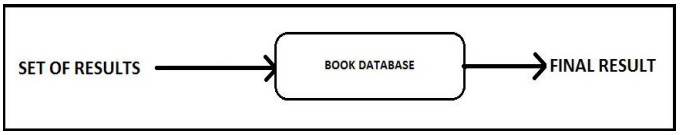


Fig. 5. RDBMS Querying

## V. IMPLEMENTATION

A. Data Collection

RDF's of about 45000 book is collected from Project Gutenberg. For the testing purpose the domain is reduced to 500 books. The data is fed into the database by the administrator. An RDF is added to the repository.

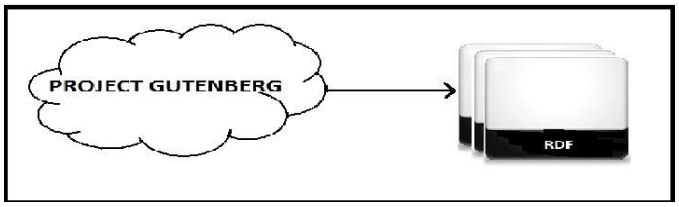


Fig. 6. Data Collection

B. User Query

A graphic user interface is available for the user to enter the search query. The user is assumed to follow certain standards when giving a query like enclosing the author name in single quote and name of the book with double quote. E.g.: Book written by 'Peter Jackson' Book similar to “Sherlock Holmes”

C. Creating SPARQL Query

For every user patterns identified, the SPARQL query corresponding to the pattern is created. This SPARQL query and corresponding patterns are entered in SPARQL query table. A user interface is created for entering this pattern, its corresponding SPARQL and number of variables, when data is entered it finds the hash value and enters the data in to the database.

D. The search process

The user query is further processed. First it is converted to lower case. Then stop words (like the, by, is etc..) are removed. Then the query is modified such that the book name is replaced by tag <bk> author name is replaced by tag <au> and topic with <tp>. Then the query is trimmed so that extra spaces will be removed from it. After this Java WordNet API functions are used to find synonyms of words in the query which is replaced by the best possible substitute based on a sorted ordering of the synset. This enables to understand different queries which has same meaning. Then the hash value of the modified form is found and matched with the hash value found in the SPARQL query table. Thus the corresponding SPARQL query is found. An example is given below to understand the various steps in this process. E.g.: If the user query is: Book written by 'Peter Jackson'

- User Query: Book written by 'Peter Jackson'
- Converted to lowercase and trimmed: book written by 'peter jackson'
- Stop words removal: book written 'peter jackson'
- Substituting <au>for author name: book compose <au>
- Find hash value: -222324243223
- Retrieve SPARQL query: SELECT

After finding the SPARQL query, it is queried in the Sesame database with the replaced tags changed back to its original values. This returns a result set which is further queried on the RDBMS to find the final book details. This results are shown as the output to the user through the user interface.

## VI. RESULT

The result from the proposed system is compared with the results from an online digital library "Universal Digital Library” http://tera-3.ul.cs.cmu.edu/.

Test case 1: Searching for the books by the author name Carroll Lewis. Table shows the results from the two systems. The Universal Digital Library shows the book containing Carrol Lewis in its title. But, the proposed system shows the books written Carrol, Lewis.

| Universal Digital Library | Semantic Digital Library |
|---|---|
| • Lewis Carroll by Florence Becker Lennon<br>• The Collected Verce of Lewis Carroll<br>• The Complete Illustrated Works of Lewis Carroll<br>• The Life and Letters of Lewis Carroll by "Stuart Dodgson Collingwood”<br>• The Complete Works of Lewis Carroll by Alexander Woollcott John Collingwood"<br>• The Life and Letters of Lewis Carroll (Rev C L Dodgson) by "Stuart Dodgson” | • The Hunting of the Snark; An Agony in Eight Fits by Carroll, Lewis<br>• Alice's Adventures in Wonderland by Carrol, Lewis<br>• The Federalist Papers by Carroll, Lewis |

Table.1. Test case 1 results (source : https://goo.gl/Z83WOA)

Test case 2: Second scenario is where we search for a book title, in this case it is Alice’s adventures in wonderland. The results are all the titles matching the input in the Universal Digital Library system. The system proposed by us gives these results as well as other books written by the same author or books in the same genre.

International Journal of Communication and Networking System
Volume: 05 Issue: 02 December 2016, Pages No.113-1116
ISSN: 2278-2427

| Universal Digital Library | Semantic Digital Library |
|---|---|
| • Alice's Adventures In Wonderland by Carroll Lewis | • The Hunting Of The Snark: An Agony In Eight Fits Carrol, Lewis<br>• Alice's Adventures in Wonderland, Aesop's Fables Carroll, Lewis |

Table.2. Test case 2 results (Source: https://goo.gl/Lxo5hc )

Test case3: Complex query is given as input here. The query "book written by the author of "Alice's adventures". While our system processed this query using SPARQL and gives output the Universal Digital Library displayed No results.

| Universal Digital Library | Semantic Digital Library |
|---|---|
| No results | • The hunting of the shark; an agony in eight fits carroll, lewis<br>• Alice's adventures in wonderland carroll, lewis |

Table.3. Test case 3 results (Source : https://goo.gl/FqXmmM)

Semantic web technology is the future of current web. It offers much expressive power for the user and provides interoperability between human and machine. The application of Semantic web technology is not just restricted to Internet but it can be applied in Digital Library also. With this in mind we have designed a model for Semantic Digital Library. Where it gives the user more diverse choices of queries and provide better results than keyword search. After deploying this model of semantic digital library we have been able to increase the ease of search. Any user who wishes to search for a topic can get relevant results with one search rather than searching for different books and checking for the presence of topic of interest. As this can be seen from the results shown in the result section.

***Acknowledgment***
We are highly indebted to Mr. SAIDALAVI KALADY, Associate Professor, Head of Department of Computer Science & Engineering, NIT Calicut for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project design. His motivation has always left us spellbound. We would also like to express our sincere thanks to Mr. JAYARAJ P B (Assistant Professor, Department of Computer Science & Engineering, NIT Calicut) for co-ordinating the project. We would like to express our gratitude towards our parents & friends for their kind co-operation and encouragement which helped us in different stages of completion.

***References***
[1] Imran Alam Shoeb Ahad Siddiqui Nida Khan Deptt Of CS Deptt Of CSE Deptt Of CSE Jamia Hamdard, Delhi Integral University, Lucknow Integral University, Lucknow. The Semantic Web Converting the Current Web Services. International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 3, March 2013
[2] System and Technology Advancements in Distance Learning. Vive (k) Kumar (Athabasca University, Canada) and Fuhua Lin (Athabasca University, Canada).
[3] Amrita Bhandari Department of Computer Science and Engineering, Thapar University, Patiala, Punjab, India, Shalini Batra' Department of Computer Science and Engineering, Thapar University, Patiala, Punjab, India, SEMANTIC RETRIEVAL FOR HOMONYMS USING RDF AND SPARQL. Journal of Global Research in Computer Science
[4] http://semanticweb.org/wiki/Ontology.html
[5] http://www.obitko.com/tutorials/ontologies-semantic-web/specification-of-conceptualization.html
[6] https://www.w3.org/TR/owl-features/
[7] Projes Roy , Dipti Arora Social Semantic Digital Library: The Future Article (PDF Available) in DESIDOC Journal of Library & Information Technology 31(4) · July 2011
[8] http://www.tippingpointlabs.com/branded-content-distribution-kosmix-contextual-search/
[9] J.Z. Wang, Jia Li, G. Wiederhold,Dept. of Comput. Sci. & Eng., Pennsylvania State Univ., University Park, PA, USA SIMPLIcity: semantics-sensitive integrated matching for picture libraries. IEEE Transactions on Pattern Analysis and Machine Intelligence ( Volume: 23, Issue: 9, Sep 2001 )
[10] https://opennlp.apache.org/documentation/manual/opennlp.html
[11] http://www.academia.edu/22262350/An_analysis_of_Linked_Data_Tools_for_supporting_Architectural_Knowledge
[12] Web Application Architecture, Principles, Protocols and Practices: Computer science, Computer science, Cram101 Textbook Reviews
[13] Web Application Architecture, Principles, Protocols and Practices: Computer science, Computer science, Cram101 Textbook Reviews
[14] https://en.wikipedia.org/wiki/SPARQL
[15] https://wordnet.princeton.edu/
[16] http://www.sqlcourse.com/intro.html