Spvryan's International Journal of Engineering Sciences & Technology (SEST)
ISSN : 2394-0905

Framework for Data Extraction using Linked Data in Distributed System

Mrs. Smita S. Sapkal
PG Student, Computer Department
D.Y. Patil College of Engg.
Pune, India
smita.jadhav88@gmail.com

Ms. Soudamini Pawar
Assistant Professor, Computer Department
D.Y. Patil College of Engg.
Pune, India
psoudamini@yahoo.co.in

***Abstract*— Linked data systems are the architectures which are only linking the data source in the huge semantic web system using the ontology. This puts an obligatory situation for the user to visit the provided URL by the linked data systems to get the required data which may be in the form of files. Many times these URL may not be working or they might be perished, so the system wastes the user's time. So to enhance the concept of linked data further ahead and providing the user directly the required data instead of URL we come with a solution which extracts the data from different data servers in the distributed paradigm. The proposed system extracts the tagged images of movie celebrities from data servers of distributed system for the user's query. Query contains only the celebrity name which is actually not enough to fetch the data, so our system uses the information from web pages of the related web service which contains complete celebrity information. User search experience is enhanced using the combination of Fuzzy Logic and Interactive Genetic Algorithm to extract maximum relevant result for user query.**

Keywords*— information retrieval; databases; linked data; web of data; crawler; url; uri; rdf.

I. INTRODUCTION

Linked Data is defined as the standards for interlinking and publishing structured data on the Web. The concept of Linked Data was introduced by Tim Berners-Lee in the note on Linked Data [1] and has been published as the Linked Data principles. These principles are as follows:

1. Use URIs(Universal Resource Identifiers) as names for things.
2. Use HTTP(Hypertext Transfer Protocol) URIs, so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF(Resource Description Framework), SPARQL(Simple Protocol and RDF Query Language)).
4. Include links to other URIs, so that they can discover more things.

The task of linking has different types stated below [6]:

1. Relationship Links**:** Related data from other data sources can be traced with the help of relationship links. Bibliographic data i.e. data about the publications or data about where people live can be pointed out with the use of relationship links.
2. Identity Links: To detect similar real world objects or abstract concepts, identity links are used for the identification of URI aliases mentioned in other data sources. Using these URI aliases, further description about the entities can be retrieved from other data sources. Identity links perform significant social function as they enable the world to express different views on the Web of Data.
3. Vocabulary Links: These links are used to identify or relate similar terms in other vocabularies. Data is pointed to the definition present in the vocabulary using vocabulary links. These links increase the self-descriptive value of data and further enable data across different vocabularies to be integrated for developing Linked Data applications.

The use of identifiers for data and concepts play an important role in Linked Data applications. Resources can be described jointly and linked to the related data created by other individuals or communities. Linked Data openly explores the reusability and recombination of the resources contributed by the expert individuals. Anyone can freely contribute to the existing data resources using the concept of Linked Data. Diverse descriptions can be referred to the same thing by means of identifiers. Libraries can be enriched with the use of rich linkages pointed to complementary data from reliable sources. Thus, value of own data of the libraries can be increased considerably rather than the individual sources.

The prominent feature of the World Wide Web forms the basis of Linked Data: use of URIs to browse the content of the information space. On the World Wide Web, there are numerous websites and web pages available to users and applications; similarly Linked Data makes use of datasets expressed in RDF and users can browse the datasets resolving

the trails of URIs. Thus, the importance of Linked Data for library users is increased with these navigation principles.

Linked Data is merely the extension of the present web and optimization in the form of addition of structured data. RDF in Attributes (RDFa) and microdata technologies is used to express this structured data, which further enables the libraries to enhance its visibility through search engine optimization (SEO).

Information seekers can use the embedded structured data to reuse the library data in services. Thus, citation management has become as simple as cutting and pasting URIs. Library data can be fully integrated into research documents as soon as the task of link creation from web to library resources is accomplished. The interdisciplinary research is facilitated with the use Linked Data and links are created from various knowledge bases, which are domain specific, for the purpose of enrichment of knowledge.

The proposed system enables the user to query the entities from multiple data servers which are linked implicitly to obtain the result under a single roof. Fuzzy Logic[11] and Interactive Genetic Algorithm[12][13] are used to provide enriched results to the user. For implementation, image data set is distributed on multiple data servers over the network and is comprised of related images of Bollywood celebrities. The proposed system allows the user to search celebrity images with respect to categories like actor, director, music director, guest appearance, producer etc. The system returns the result based on search vector. Search vector is created by movie names from the live web site[1] automatically by the system. (1 http://www.gomolo.com/)

The rest of the paper is organized as follows: Section 2 discusses some related work and section 3 presents the design of our approach. The details of the results and some discussions on this approach are presented in section 4 as Results and Discussions. Section 5 elaborates hint of some extension of the approach as future work and conclusion.

II. Related Work

The format of Linked Data optimizes interconnectivity between data. The task of link creation is accomplished with a view to leverage the usability and interoperability of the data. Thus, the data in Linked Data sets becomes more knowledge rich. Identity is the most important factor behind the development of Linked Data. To check the reliability and validity of the data found during search, it is essential to determine the identity of data [2]. Christian Bizer et al.[3] mentions the use of URIs not only to identify Web pages, but also to identify everything (both tangible and intangible). In order to use Linked Data on the Web, a structure was established to define how URIs could be used and how they would be resolved and connected with other URIs.

In Linked Data, sharing of the structured data on the global scale is based on generalized architecture of the World Wide Web [4]. It is essential to know the functioning of classic document web for understanding the Linked Data principles. The Uniform Resource Identifiers (URIs), the Hypertext Transfer Protocol (HTTP) as universal access mechanism and the Hypertext Markup Language (HTML) standards together forms the basis of document web. URI is used to identify the things, HTTP is used to access the things and HTML contains the data wrapped in a structured format [5]. The mentioned standards enable web to outperform different technical architectures. Hyperlink is used to interconnect the data residing on different Web servers and to crawl the Web.

RDF is a significant component for the creation and implementation of Linked Data framework [7]. **OWL** (Web Ontology Language) is used to create ontologies. **RDF datasets** are at the top of Linked Data Stack. For identifying the things in RDF datasets, URIs are used, instead of simply writing it as a string. For a specific domain, numerous vocabularies are available, which can be used interchangeably. Similarly, within Linked Data, datasets can be used interchangeably. Thus, it provides infinite terms and connections between these terms and the datasets can be easily meshed together.

Consider the example of unique URI used in LCSH(Library of Congress Subject Headings) to describe "dogs". The URI used in LCSH might state the exactMatch property and direct a link to unique URI in DBpedia.org to describe dogs. This task requires a lot of time and manual work. Problems of exact definition and granularity can also arise. Use of skos:exactMatch property is very tedious. It strictly requires the two terms to be similar in use and meaning. It becomes difficult when connections are to be made between two different vocabularies. For example, the URI in LCSH for Mark Twain might represent the work as an author while the same URI in DBpedia.org might describe the person. Prior to linking such vocabularies, critical study and philosophical examination is needed.

Tabulator[8] is a user friendly browser for Semantic Web. It enables the user to follow the RDF links and interact with the web of data. The RDF links are followed based on the user's exploration and analysis. DBpedia Mobile[9] elaborates the application of Linked Data using mobile devices. It enables the user to publish content enriched with location information, photos, reviews etc. and further linked to DBpedia resources. Fenfire[10] browser is used to browse Linked Data with a graph view. It provides the user interactive browsing experience and also enables the user to edit RDF graphs.

III. Proposed Method

In this section, we describe the approach of enriching Linked Data system using information captured from related web pages. The 10 main steps described are shown in Fig. 1.

Step 1: In this step, user enters a query as celebrity name in the system to search and retrieve the images from the different data servers over the network.

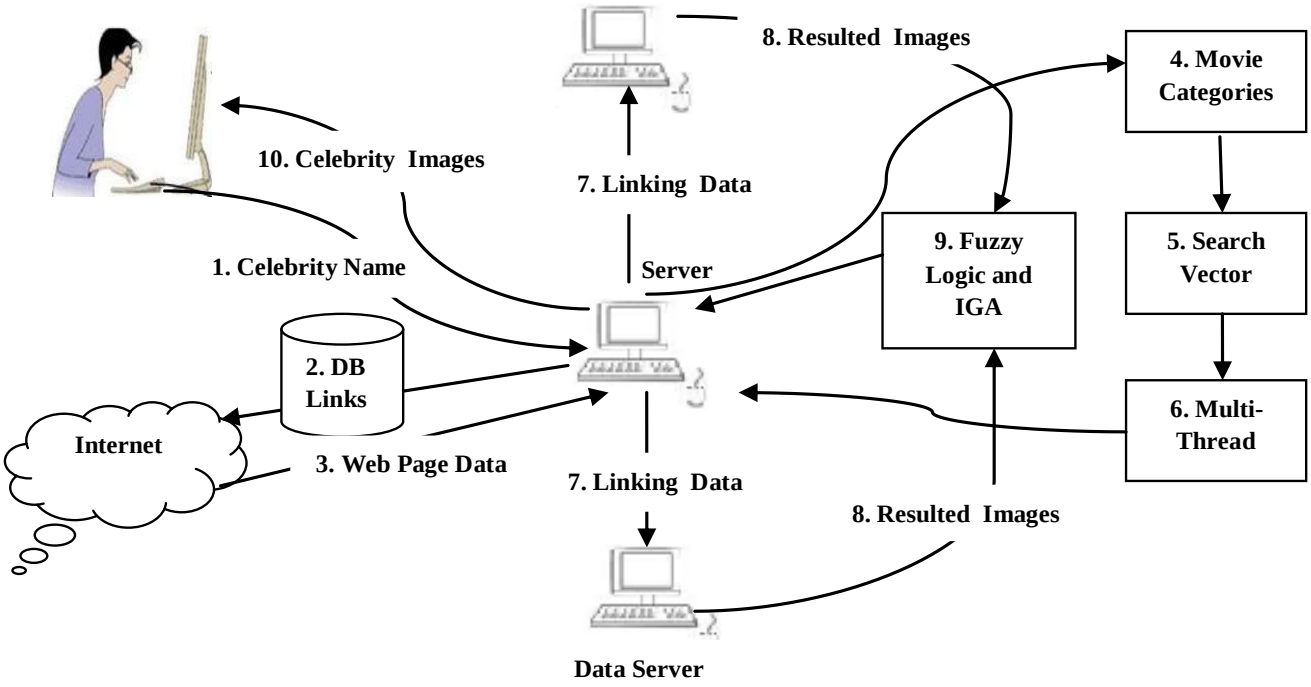

Fig. 1. Overview of our approach

Step 2: Here the celebrity name gets its proper URL of the online site which is stored in the database.

Step 3: This is one of the most crucial phase of experiment, where the system interacts with the live web page of the respective URL. Using a well designed small web crawler, the system fetches data of the web page and then parses all the HTML tags from the web page. Only human readable data is extracted and the advertisement links present on the web page are omitted in this phase.

Step 4: Here the data which has been extracted from the web page in Step 3 consists of only the movie names of the celebrity. These movies are then identified under many categories by the system like actor, music director, narrator, playback singer, presenter, producer, screenplay, story and thanks to.

Step 5: In this step, the movies names which are extracted from the web page are tagged under the category which are identified in the previous step to form a searching vector. This is the master key to link and extract data from data servers.

Step 6: To link the data, system needs to look into several data servers simultaneously. To achieve this, system produces a number of multi-threads based on the number of internal data server over the network.

Step 7: In this step, each multithread is then loaded with the search vector and then finally it gets connected to the data server in the network and thereby finds the images and links them implicitly.

Step 8: After linking, the resultant images from the data servers are transferred to the main server using FTP protocol.

Step 9: This is the main step of proposed system as Fuzzy logic and Genetic algorithms are playing a vital role to provide enriched results to the user.

In order to use a statistical method, it is necessary to represent the movies as vectors of categories. These categories are attributes that attempt to represent the image used for retrieval. We focus on nine categories as stated in step 4 for each search.

Later, fuzzy logic is used to summarize the movie categories. A value from zero to hundred in percentage is obtained for each category. The system provides output based on category features and the available rules in the knowledge base. The resultant value in the output determines the degree of importance of the category in the final summary. The input membership function for each category feature is divided into three membership functions which are composed of insignificant values low (L) and very low (VL), Medium (M) and significant values high (H) and very high (VH). The most important part in this procedure is the definition of fuzzy IF-THEN rules. The desired images are extracted from these rules according to features criteria.

By using IGA (Interactive Genetic Algorithm) which is a branch of evolutionary computation, the fitness is determined by the user's evaluation and not by the predefined mathematical formula. User can interactively determine which members of the population will reproduce, and IGA will automatically generate the next generation of content based on the user's input. In this step, if the user is not satisfied or he/she is willing to retrieve the images based on some different parameters for the respective categories of the images then user can change those values and request to the system to provide new results.

Step 10: In this step, final retrieved images are rearranged according to the data server and movie categories and then finally displayed to the user.

The complete working approach of our model is represented in the algorithm 1 depicted below:

Algorithm 1

// input: Celebrity name C_n

// output : C_{set} , set of celebrity images

function linked_Data(C_n)

1: get C_n URL as C_{url} from DB

2: Fetch WebData as C_{web_data} from C_{url}

Create C_{cat} as Celebrity category **from** C_{web_data}

4. Create C_{movies} as Search Vector **from** C_{web_data}

5. Create Set M_t={ m_{t1},m_{t2}……m_{tn}} for multithread

6. M_t ← DataServerSet(D_{s1},D_{s2}…D_{sn})

7.C_{movies}← DataServerSet(D_{s1},D_{s2}…D_{sn})

8. **for** i ← 1 to n

9. C_{set}← DataServerSet(D_{s1},D_{s2}…D_{sn})

10.C_{set} ← C_{data}

11. (Fuzzy Set)F_{set} → C_{set}

12. C_{set} ←geneticRecall(C_{set})

13.**return** C_{set}

IV. RESULTS AND DISCUSSIONS

To show the effectiveness of the proposed system, some experiments are reported. Selecting a suitable image database is a critical and important step in designing an image retrieval system. Currently, there is no standard image database for this purpose. Also, there is no agreement on the type and the number of images in the database. Since most image retrieval systems are intended for general databases, it is reasonable to include various semantic groups of images in the database.

For experiment, over 2000 images for about 20 different Bollywood celebrities from many different sources were collected. These image dataset were then kept in three different data servers which play their role in Linked data system.

A. Practicability of System Demonstration

An example is stated to test practicability of the proposed system. A user submits a query in the form of a string which is a celebrity name, for which the user needs to search all the possible images by celebrity name or by the celebrity movie name in which celebrity worked in any mentioned categories (Step 4). The system then fetches all the movie names from online web service as mentioned earlier and links the images of the data server implicitly and finally shows the resultant images to the user based on the fuzzy rules set by the system. If the user is not satisfied with the results, then the user can again search the images with other searching parameter using the genetic algorithm.

B. Retrieval Precision/Recall Evaluation

To evaluate the effectiveness of the proposed approach, the system counts on the relevant images that are retrieved for query entered by the user. The retrieval effectiveness can be defined in terms of precision and recall rates. For experimental results, 10 celebrities are considered. Each celebrity has 25 images which are distributed in 3 data servers over the network in LAN.

For more clarity we assign

- A = The number of relevant records retrieved,
- B = The number of relevant records not retrieved, and
- C = The number of irrelevant records retrieved.

So, Precision = (A/ (A+ C))*100

And Recall = (A/ (A+ B))*100

In Fig. 2, it is observed that the tendency of average precision for the retrieved images is 90% which is a better precision result in linking and extracting the data.

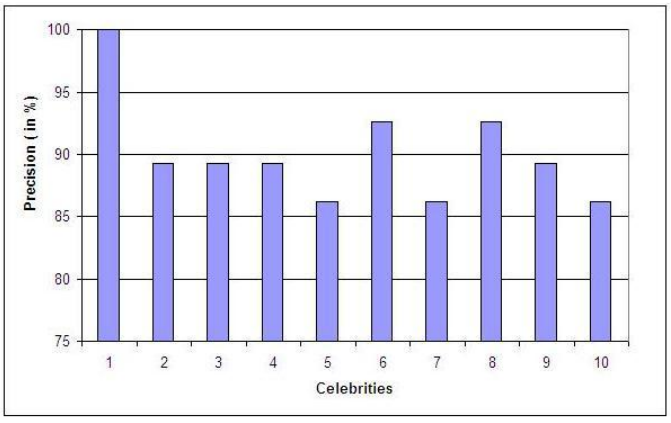

Fig. 2. Retrieval average Precision of the proposed approach

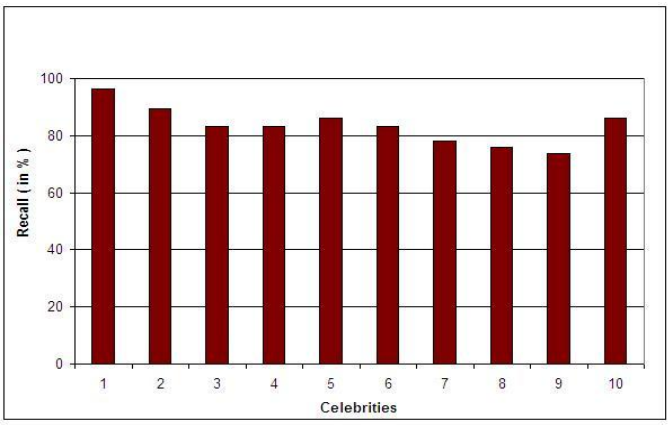

Fig. 3. Retrieval average Recall of the proposed approach

In Fig. 3, it is observed that the tendency of average recall for the retrieved images is about 84% which is actually a better recall result and can achieve 100% within few generations of Genetic Algorithm.

C. Comparison with Other Methods

In order to show the superiority of proposed approach, it is compared with system like those in [13]. 90% of precision and about 84% of recall (recall can be achieved 100% by further generation of genetic algorithm) are achieved by the system which not only links the data in the file system of the data servers in distributed network but also it can extract and provide the resulted data to the user under a single roof.

There are many systems, which link the data that have been already defined in RDF, href tag of the html file or in any ontology form, but very few systems are developed like the proposed system which is capable of performing both jobs of linking and extracting data from the data servers.

V. CONCLUSION AND FUTURE WORK

This paper has presented a naïve user-oriented framework in interactive Link Data system, in contrast to conventional approaches which are only linking the data defined in ontology or with the hyper reference tag of html file. The searching of images from data servers in the network is performed successfully by using implicit linked data techniques. User query is implicitly optimized with the online available information and exponentially enlarges the searching parameters for images with the celebrity movie names under all available respective categories. Finally, the system enriches the user results by fuzzy logic which is well aided with the interactive genetic algorithm which improves the user searching experience generation by generation. The experiments are performed with around 2000 images distributed over 3 data servers in the network. The results show the best average precision and recall to summaries produced by the fuzzy method.

The proposed approach is extended by using a combination of Fuzzy Logic and Genetic Algorithm methods and extracts the best possible image result for the user query.

The proposed idea is useful to develop a system for users, who needs to search a document from an online library system of a college. The query can connect to the different college library servers and obtain the result from implicitly linked servers with a direct downloading option.

The future enhancement of this system would be to determine whether Linked Data could be used to direct users to their websites through search engine results and also to download their required data implicitly along with linking. Such systems could be developed for museums, libraries etc.

REFERENCES

[1] T. Berners-Lee. (2006, Jul. 27) *Linked Data - Design Issues* [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html.

[2] H. Glaser, H. Halpin, "The Linked Data strategy for global identity," *IEEE Internet Computing*, vol.16, no.2, 2012, pp.68-71.

[3] C. Bizer, T. Heath, and T. Berners-Lee., "Linked Data - The story so far", Int. J. Semantic Web Inf. Syst., vol. 5, issue 3, pp. 1–22, 2009.

[4] I. Jacobs and N. Walsh. (2004, Dec. 15) *Architecture of the World Wide Web* [Online]. Available: http://www.w3.org/TR/webarch/.

[5] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a global data space", 1st ed., *Synthesis Lectures on the Semantic Web: Theory and Technology*, San Rafael: Morgan & Claypool, 2011, pp. 1-136.

[6] S. Auer, L. Bühmann, C. Dirschl, O. Erling et al. (2010, September 1) *LOD2-Creating Knowledge out of Interlinked Data*, Collaborative Project. [Online]. Available: http://static.lod2.eu/Deliverables/deliverable-1.2.pdf.

[7] J. Mixter, "Linked Data in VRA Core 4.0: Converting VRA XML records into RDF/XML," M.S. thesis, Kent State Univ., 2013.

[8] T. Berners-Lee, Y. Chen, L. Chilton et al. "Tabulator: Exploring and analyzing linked data on the semantic web", presented at the 5th International Semantic Web Conference, Athens, Georgia, 2006.

[9] C. Becker, C. Bizer, "DBpedia mobile - A location-aware semantic web client," presented at the 7th International Semantic Web Conference, Karlsruhe, Germany, 2008.

[10] T. Hastrup, R. Cyganiak, and U. Bojars, "Browsing Linked Data with Fenfire," presented at the 1st Linked Data on the Web workshop, Beijing, China, 2008.

[11] A short fuzzy logic tutorial (2010, Apr. 8) [Online]. Available: www.cs.bilkent.edu.tr/~bulbul/depth/fuzzy.pdf.

[12] Interactive evolutionary computation (2013, Oct. 30) [Online]. Available: http://en.wikipedia.org/wiki/Interactive_evolutionary_computation.

[13] D. Beasley, D. Bull, R. Martin, "An overview of Genetic Algorithms: Part I, Fundamentals," *University Computin*, vol. 15, no. 2, pp. 58-69, 1993.

[14] S. Dietze, H. Qing Yu, D. Giordano, E. Kaldoudi, N. Dovrolis, and Davide Taibi, "Linked education: Interlinking educational resources and the Web of Data,". in *The 27th ACM Symposium On Applied Computing, Special Track on Semantic Web and Applications*, Trento, Italy, March 2012, pp. 366-371.