Journal of International
Academic Research for Multidisciplinary

A Global Society for Multidisciplinary Research

www.jiarm.com

Editorial Board

DESIGN OF SVRDBML SYSTEM FOR CLASSIFICATION OF RDF CONTENTS

ASAAD SABAH HADI*
ABBAS M. AL-BAKRY**

*Dept. of Software, Information Technology College, Babylon University, Iraq
**Dept. of Software, Information Technology College, Babylon University, Iraq

ABSTRACT

The Semantic web has change the way for the most researcher to solve the problems of the classical web. The user search for a specific information all the time but most search engines fail to meet the user requirements. The search engine filter the pages from searching unnecessary ones, but the challenge is to answer intelligent queries received from users based on information available in web pages . Most search engine try to access the web pages through Web APIs. The Web API do not assign globally unique identifiers to data items, then it is not possible to set hyperlinks between data items provided by different APIs. Web APIs therefore slice the Web into separate data silos. The Linked Data Principles solve the fragmentation problem. The RDF is the database for the semantic web and the SPARQL is the language for informing the information inside the RDF. The user must know the contents of the RDF before using the SPARQL. In this paper, we design and implement a system that treat the user querying process as a classification problem and we implement this system using BBC nature wildlife website which it is a real database from the Linked Opened Data. The system gives us an excellent result for most classifiers.

KEYWORDS: Semantic Web, Linked Data, Linked Open Data (LOD), SPARQL, RDF, RDFS, Matrix, Classification, and Machine Learning.

INTRODUCTION

The information retrieval from the World Wide Web is a challenging tasks. The researchers in the information retrieval are concern with the content portion of the hyperlinks that connects various document s[1]. Information is the main source of intelligent [16]. Just by typing any keyword on any search engine will provide you with millions of information, but the amount of relevant information are very low and the user must search manually on all these result, therefore this type of search is not user friendly [16]. The machine misunderstanding of the classical web contents is one of the main problems, because the information does not have any meaning. In other word, to process a WebPages intelligently, machine must understand the contents of these WebPages [2]. Tim Berners-Lee has innovate

the term Semantic Web, which solve some of the problems of the classical web. The classical web does not implement applications against all the web data as a separate silos [3,4].

The data is the main part of the web and knowing the connection way of these data is the dreams of many researchers that is done by Tim Berners Lee who find the principles of the Linked Opened Data(LOD). The main idea of his work is publish the data in a general graph that make these data more accessible to the users and machines [5]. Instead of webpage, the primary data model of the Linked data is RDF (Resource Description Framework) which is a W3C (World Wide Web Consortium) standard for representing any information. The formula of the RDF is: Subject – Predicate –Object. To retrieve the information from RDF we need to use the SPARQL (SPARQL Protocol And RDF Query Language)[6,18].

The LOD extend the web by publishing various data sets and put a link between these data and the resulting data is termed the Linking Open Data cloud. The DBpedia is the major linking hub of the LOD cloud which can help us by linking it with our unstructured data [7,19].

There are three main approach for querying linked data, the first approach employing the information retrieval based on keyword search, the second approach focus on natural language queries and the third approach is structure SPARQL queries over distributed database. In this paper we suggest and implement the fourth type that aim to convert the RDF into matrix vectors and treat the querying manner as a classification problems which is better than SPARQL which need a good experience from the user before making any query.

OVERVIEW

In [20], Olivier Corby, et. al focus on IR(Information Retrieval) on the semantic web which is needed in web applications such as web browsing, E-learning, e-commerce, etc. They show that the previous work on ontology IR, like SHOE, On to Broker, On to Seek, Web KB, Corese, focused on ontology knowledge representation language, but in this paper they rather focus on the query processing point of view to enhance the query result.

In [21] Christian Bizer, et. al described the extraction of the DBpedia knowledge base and gave an overview of application that facilitate the Web of Data around DBpedia. The DBpedia project leverages the gigantic source of knowledge by extracting structured information from Wikipedia and making this information accessible on the Web. The DBpedia knowledge base has several advantages over existing knowledge base: it covers many domains, it represents real community agreement, it automatically evolves as Wikipedia changes, it is truly multilingual and it is accessible on the web. The DBpedia

project showed that a rich corpus of diverse knowledge can be obtained from the large scale collaboration of end-users. The DBpedia knowledge base has the potential to revolutionize the access to Wikipedia. The utility of the knowledge base as interlinking hub for the Web of data is demonstrated by the increasing number of data sources that decide to set RDF links to DBpedia and the growing number of annotation tools that use DBpedia identifiers.As a future work, the authors give three directions: Cross-language info box knowledge fusion, Wikipedia article augmentation, Wikipedia consistency checking.

In [4] Rupal Gupta and Sanjay Kumar, focus on the role and usage of SPARQL. They also make a Comparison of SPARQL with SQL and present an execution analysis of SPARQL with some tools. The authors use Two tools for execution Query , the first tool is Jena with ARQ processor which is a command line interface and the other is TWINKLE which is a GUI interface .They show that querying SPARQL query using TWINKLE tool is more convenient than with Jena ARQ processor, and they give some important advantages of TWINKLE over Jena and they propose as a future work to design some methodologies which can give an optimization concept for efficient data retrieval from SPARQL.

LINKED DATA

It is a promising technology for storing and providing structured data. Linked data uses the principles and technologies of the semantic web to publish and interlink data, then all entities are references by URIs (Uniform Resource Identifier) using the standard web protocol HTTP (Hypertext Transfer Protocol) [8,9]. The particular strength of Linked Data is that applications can use it straightforward [18].

The best principles that are advocated by the LOD community for collaboratively interlinking and publishing structured data over the web [10,11,19]:

- Use HTTP URIs so that the names can be looked up.
- Use URIs as names for things.
- Return useful information upon lookup for those URIs.
- Include links by using URIs that reference to remote documents.

There are a number of challenges for LOD, the first is that the Linked Data is structured, the second is that the Linked Data is dynamic, the third is the linked data is uncertain, finally the Linked Data is distributed.

MACHINE LEARNING

Machine Learning (ML) is the study of how computers can learn patterns in empirical data [12,17]. It refers to the changes in systems that perform tasks associated with AI like Recognition, Planning, Robot control, Diagnosis, Prediction[13,14,15].ML methods treats the description of the relationship between observed variables and group memberships as a "black box" and don't assume a probabilistic data model. The main goal of ML is to convert observational data into a model that can be used for prediction of unseen data [12, 15].

SYSTEM ARCHITECTURE

We propose a new approach SVRDBML (Searching Vectorised RDF Data Based Machine Learning Algorithm) that employs the Semantic Web and Linked Data Techniques to overcome some of the classical web limitation. In the new Web, the content of the webpage is the RDF (Resource Description Framework). The language that retrieve the information from the RDF is called SPARQL (SPARQL Protocol And RDF Query Language) .This language is an improvement of the keyword-based search but it requires an exact match between the query structure and the RDF content and there is no support for probabilistic reasoning that can give the consistency between the query and the retrieved content.

Our proposed approach converting the RDF content into a matrix of features and treat the queries as a classification problem. It contain two phases: the Offline and Online phase. The main idea of offline phase is to read multiple webpages and convert them into RDF/XML format and check the syntax of each file then convert it into RDF/Ntriples format. Finally collect all corrected files into a directory. Figure (1) explain the part one of the offline phase.

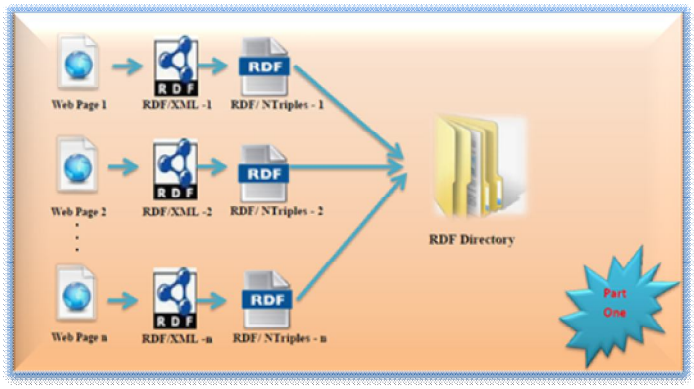

Figure 1 part one of the offline phase

In the second part (part two) , that explained in figure (2), the directory has been reading one file at a time and model is extracted for each file and from that model the system

extract the RDF statements (subject, predicate, object). The statements , which contain all the information in the file , are converted into a matrix format that contain '0' or '1', where '0' means no statements for that triples and '1' means existence of the triples. For all the WebPages (RDF file) the system create two general set :new_ predicate set and new_ object set after removing the less importance features. The new_ predicate set and new_ object set save all the predicates and objects respectively in all WebPages (without duplication). Then each matrix for every webpage is converted into Vectors that contain the label and the vector values. The label is the name of the webpage and the values is the object values for the statement that the label is the subject for it. The Vector for all WebPages are saved in a big CSV (Comma Separated Value) and the system also create other CSV files to use them in the online phase. Table (1) explain the content of the CSV File.

Table 1 Multiple Vectors with multiple Values

	Predicate- 1	Predicate- 2	. . .	Predicate- n
Web Page label-1	*Object Value*$_{(1,1)}$	*Object Value*$_{(1,2)}$	. . .	*Object Value*$_{(1,n)}$
Web Page label-2	*Object Value*$_{(2,1)}$	*Object Value*$_{(2,2)}$	. . .	*Object Value*$_{(2,n)}$
. . .	. . .	. . .	. . .	. . .
Web Page label-m	*Object value*$_{(m,1)}$	*Object Value*$_{(m,2)}$	. . .	*Object Value*$_{(m,n)}$

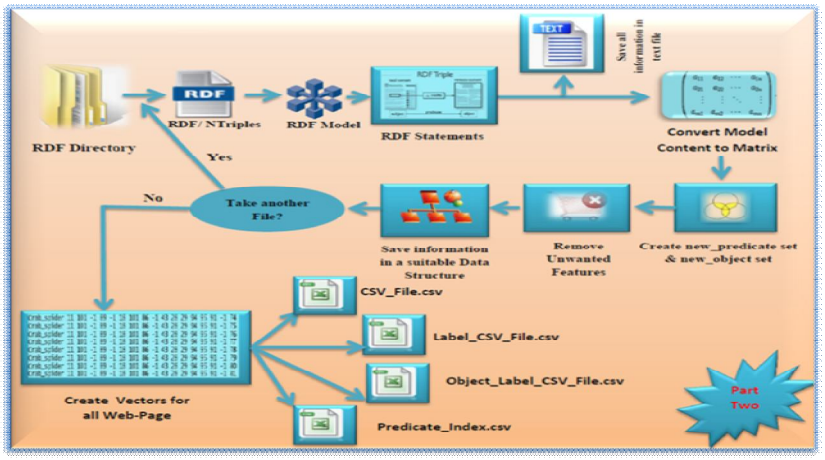

Figure 2 Part Two of the offline phase

In the online phase , the system read the CSV files that contain the vectors for all WebPages (files) and randomize its contents in order to increase the complexity of these vectors, then the system create the Array_ CSV and create the General _Set which contain the training set and the testing set. The system pass the training set into multiple classifiers. If the classifiers pass the accuracy test that done by the system using the testing set , it add to the set of good classifiers , else the system check for some enhancement and eliminate it if continue to give bad classification accuracy. Finally the system is ready to take the query from a user as a Vector of values and each remaining classifiers give an estimated answer and the system make a voting and return the result to the user. The Figure (3) show the block diagram of the online phase.

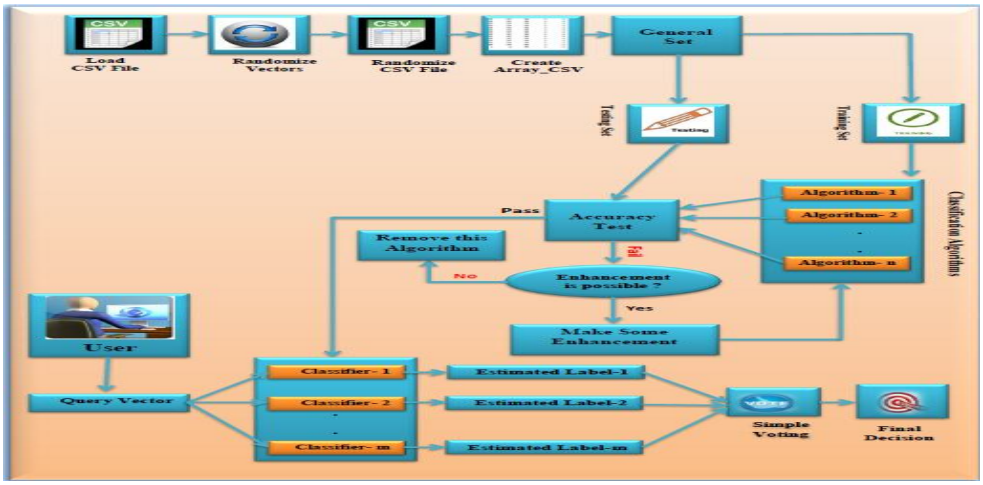

Figure 3 Online phase Block Diagram

6- SYSTEM IMPLEMENTATION

We implement our system using a real dataset from the LOD. We take the BBC wildlife website and take '327' animals WebPages in the form of RDF/XML files. The total number of predicates for all these animals are '21' predicates. These features are the index for all Vectors in the CSV file. Figure (4) show these predicates. The total number of objects in these files are '6654' objects. Figure (5) give an excerpt for the RDF/XML file for king cobra (a sample of animals).

```
Predicate 0 = http://xmlns.com/foaf/0.1/depicts
Predicate 1 = http://purl.org/ontology/wo/genus
Predicate 2 = http://purl.org/ontology/wo/kingdomName
Predicate 3 = http://purl.org/ontology/wo/scientificName
Predicate 4 = http://purl.org/ontology/wo/commonName
Predicate 5 = http://purl.org/ontology/wo/familyName
Predicate 6 = http://purl.org/ontology/wo/phylumName
Predicate 7 = http://purl.org/ontology/wo/genusName
Predicate 8 = http://www.w3.org/2002/07/owl#sameAs
Predicate 9 = http://purl.org/ontology/wo/kingdom
Predicate 10 = http://xmlns.com/foaf/0.1/depiction
Predicate 11 = http://purl.org/ontology/wo/family
Predicate 12 = http://purl.org/ontology/wo/phylum
Predicate 13 = http://purl.org/ontology/wo/name
Predicate 14 = http://purl.org/ontology/wo/adaptation
Predicate 15 = http://purl.org/ontology/wo/livesIn
Predicate 16 = http://purl.org/ontology/wo/distributionMap
Predicate 17 = http://purl.org/ontology/wo/species
Predicate 18 = http://purl.org/ontology/wo/speciesName
Predicate 19 = http://purl.org/ontology/wo/superfamily
Predicate 20 = http://purl.org/ontology/wo/superfamilyName
```

Figure 4 features for all animals

```
htt
"K
<?xml version="1.0" encoding="utf-8"?><rdf:RDF
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema "#
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns "#
xmlns:owl="http://www.w3.org/2002/07/owl "#
xmlns:foaf="http://xmlns.com/foaf/0.1 "/
xmlns:dc="http://purl.org/dc/terms "/
xmlns:dctypes="http://purl.org/dc/dcmitype "/
xmlns:skos="http://www.w3.org/2004/02/skos/core"#
xmlns:xsd="http://www.w3.org/2001/XMLSchema "#
xmlns:po="http://purl.org/ontology/po "/
xmlns:wo="http://purl.org/ontology/wo<"/
<rdf:Descriptionrdf:about="/nature/species/King_Cobra">
.
.
.
<dfs:label></wo:Phylum><wo:Kingdomrdf:about="/nature/kingdom/Animal#kingdom"><rdfs:label>animalia</rdfs:label>
</wo:Kingdom
<rdf:RDF<
```

Figure 5 An excerpt of RDF/XML for King Cobra

Figure (6) give an excerpt for the Vectors of 'King Cobra' that have the webpage http://www.bbc.co.uk/nature/life/King_Cobra. The label for each vectors in Figure (6) is "King_ Cobra" whereas the other values is each vector are the object indexes for that statement. The value '-1' means there is no object values for this feature (predicate). All the Vectors for all files are collected in one CSV file and the system give every label an index because all the classifiers need a numerical value to make the classification. Figure (7) explain an excerpt of the CSV file for all WebPages.

```
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 129 711 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 129 710 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 129 141 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 129 131 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 129 147 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 129 143 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 318 711 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 318 710 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 318 141 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 318 131 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 318 147 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 318 143 3715 3727 3726 -1 -1
King_Cobra 3720 6 16 3725 2612 4 21 3723 3718 27 3719 31 32 3716 270 711 3715 3727 3726 -1 -1
```

Figure 6 An excerpt of a vectors for King Cobra

```
1,0,5,16,17,18,4,21,17,26,27,28,31,32,33,-1,-1,-1,-1,-1,-1,-1,-1,
1,0,6,16,17,18,4,21,17,26,27,28,31,32,33,-1,-1,-1,-1,-1,-1,-1,-1,
1,0,7,16,17,18,4,21,17,26,27,28,31,32,33,-1,-1,-1,-1,-1,-1,-1,-1,
1,0,8,16,17,18,4,21,17,26,27,28,31,32,33,-1,-1,-1,-1,-1,-1,-1,-1,
1,0,9,16,17,18,4,21,17,26,27,28,31,32,33,-1,-1,-1,-1,-1,-1,-1,-1,
2,41,53,16,71,56,40,67,65,57,27,45,47,52,60,42,54,62,66,69,-1,-1,
2,41,53,16,71,56,40,67,65,57,27,45,47,52,60,42,59,62,66,69,-1,-1,
2,41,53,16,71,56,40,67,65,57,27,45,47,52,60,43,54,62,66,69,-1,-1,
2,41,53,16,71,56,40,67,65,57,27,45,47,52,60,43,59,62,66,69,-1,-1,
2,41,53,16,71,56,40,67,65,57,27,45,47,52,60,46,54,62,66,69,-1,-1,
.
.
```

Figure 7 An excerpt of a vectors for all files

The CSV file shown in figure (7) is loaded from the online phase and the system will extract the training set and testing set. We choose the training set is 20% of the general set and the testing set is 80% of the general set. We use multiple classifiers and the classifiers that pass the accuracy test are: kNN (k-Nearest Neighbour) for k=4, Decision Tree, Naïve Bayes, and Nearest Mean. Table (2) show the accuracy for each classifiers whereas Figure (8) illustrate the ROC (Receiver Operating Characteristics), where Error1=False Negative (FNs) and Error2 =False Positive (FPs).

Table 2 Accuracy for multiple classifiers for BBC website

No.	Classifier Name	Accuracy (%)
1.	kNN , k=3	97.78
2.	Decision Tree	94.32
3.	Naïve Bayes	77.21
4.	Nearest Mean	92.15

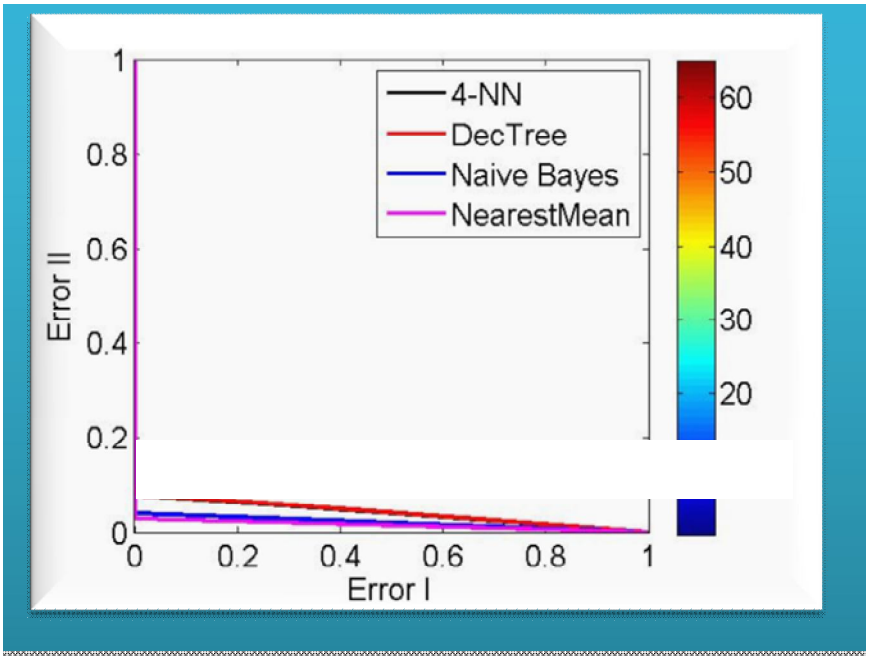

Figure 8 The ROC for good classifiers

REFERENCES

1. P. Desikan, et. all. , "Hyperlink Analysis: Techniques and Applications," 2003.
2. T. Berners Lee, et. all. , "The Semantic Web", Scientfici American, 2001.
3. C. Bizer , " The Emerging Web of linked data," IEEE Computer society, 1541-1672 , 2009.
4. R. Gupta and S. Kumar Malik, "SPARQL Semantics and Execution Analysis in Semantic Web Using Various Tools," IEEE computer society, 2011.
5. A. Hogan, et. all., "An empirical survey of Linked Data conformance," Elsevier, 2009.
6. W. IJntema, J. Sangers, F. Hogenboom and F. Frasincar," A lexico-semantic pattern language for learning ontology instances from text," Elsevier, 2012.
7. E. Meij, et. all., "Mapping queries to the Linking Open Data cloud: A case study using DBpedia," Elsevier, Web Semantics: Science, Services and Agents on the World Wide Web 9,pp. 418-433 , 2011.
8. André Freitas, et. all.,"Querying Heterogeneous Datasets on the Linked Data Web Challenges, Approaches, and Trends," IEEE internet computing, 1089-7801, 2012.
9. M.Graube, et. all., " Linked Data as integrating technology for industrial data ,"IEEE Computer Society, 2011.
10. C.Bizer, et. all.,"Linked Data - The Story So Far ," International Journal on Semantic Web and Information Systems(IJSWIS), 2010.
11. F. Bauer and M. Kaltenböck, "Linked Open Data: The Essentials, A Quick Start Guide for Decision Makers, " Book, Edition Mono, ISBN: 978-3-902796-05-9, 2012.
12. D. Stahl, et. all., "Novel Machine Learning Methods for ERP Analysis: A Validation From Research on Infants at Risk for Autism," Developmental Neuropsychology 37, pp.274-298, ISBN: 8756-5641, 2012.

13. David Milne, Ian H. Witten, "Learning to Link with Wikipedia," ACM 978-1-59593-991-3/08/10, pp. 509-518, 2008.
14. N. J. Nilsson, "Introduction to machine learning," Book, 1998.
15. S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification
16. Techniques," Informatica 31, pp. 249-268 , 2007.
17. G.Nagarajan and K.K. Thyagharajan, "A Machine Learning Technique for Semantic Search Engine," Elsevier, Procedia Engineering 38, pp. 2164-2171, 2012.
18. C. Mangold, "A survey and classification of semantic search approaches," International Journal of Metadata, Semantics and Ontology, Vol.2 No.1, pp.23-34, 2007.
19. X.Zheng, et. all., "SPARQL Query Mediation for Data Integration," MIT-ESD , 2012.
20. M.Hausenblas and M. Karnstedt, "Understanding Linked Open Data as a Web-Scale Database," Second International Conference on Advances in Databases, Knowledge, and Data Applications, IEEE computer society, 2010.
21. O.Corby, et. al," Querying the semantic web with corese search engine," 2004.
22. C. Bizera, et. al, " DBpedia - A crystallization point for the Web of Data ," Elsevier ,2009.