Semantic Web Mining using RDF Data

V. A. Chakkarwar
Asst. Professor
Government Engineering Collage, Aurangabad, MH, India

Amruta A. Joshi
Research Scholar
Government Engineering Collage, Aurangabad, MH, India

ABSTRACT
Information on the web is increasing every minute. Redundancy in information is growing rapidly. Data mining is the technique used to extract this data as per the user's query. Technically data mining analyzing and summarizing it into useful information. Keyword search is an important tool for exploring and searching large data corpuses whose structure is either unknown, or constantly changing. So, keyword search has already been studied in the context of relational databases XML documents and more recently over graphs and RDF data. Semantic web mining aims to combine semantic web and web mining. Semantic web mining is the need of today's redundant data. In this paper major focus is on minimizing extraction of number of pages by ranking technique. Due to which the extraction of information is done exact as query fired and the top ranked pages are shown to user. Here for this three main areas are going to use such as semantic web, ontology and RDF data.

Keywords
Semantic web, ontology, RDF, XML.

1. INTRODUCTION
The *Semantic Web* is a *Web 3.0 web technology* - a way of linking data between systems or entities that allows for rich, self-describing interrelations of data available across the globe on the web.

1.1 How does it differ from the web
Today, much of the data we get from the web is delivered to us in the form of *web pages* - HTML documents that are linked to each other through the use of *hyperlinks*. Humans or machines can read these documents, but other than typically seeking keywords in a page, machines have difficulty extracting any meaning from these documents themselves.

1.2 Enter linked Data
The web contains lots of information, but typically the raw data itself isn't available - rather only HTML documents constructed from data, if a web site is generated from a database at all.
So the semantic web changes the landscape of the internet with this problem in a number of ways:

- Using artificial intelligence (getting the web to do a bit of thinking for us).
- Encouraging companies, organizations and individuals to publish their data freely, in an open standard format.
- Encouraging to use data to businesses already available on the web (data give/take).

In essence, taking all that information published in HTML documents in different places, and allowing the description of models of data that allow it all to be treated - and researched - as if it were one database. The benefits to the automated research of all the data humanity has to offer on the internet in comparison to today's tools and software are tremendous. To store these types of data there are different ways to store as by storing data either in a *hierarchy* (for example XML) or in a *relational database* (for example MySQL, MS SQL)[17].

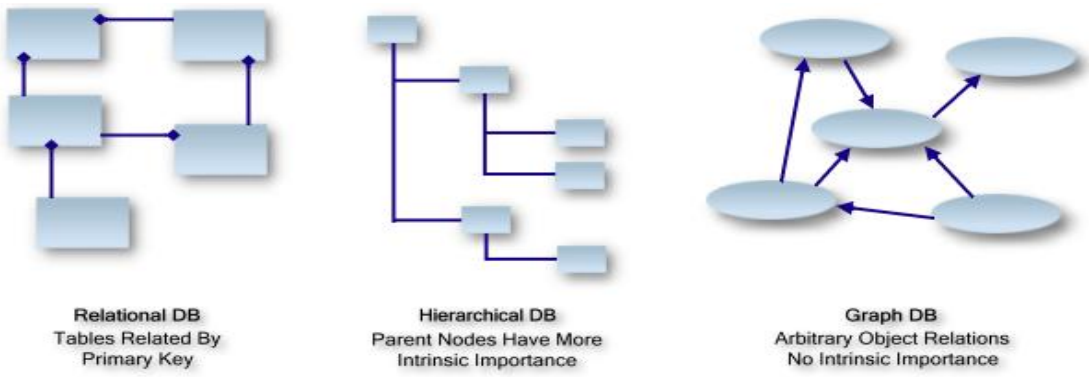

Fig 1: Graph Database represents data in different ways

1.3 Introducing the Graph Database
For most types of data storage, there is the concept of some elements of data (whether they be for example data nodes or data tables) having more precedence, or importance, over other elements.

For example, take an XML document. An XML document typically contains nodes of information each with a parent node. At the root of the document is the highest level node, which has no parent.

Take a look at the illustration above. In a data graph, there is no concept of roots (or a hierarchy). A graph consists of resources related to other resources, with no single resource having any particular intrinsic importance over another.

If the graph data model is the model the semantic web uses to store data, RDF is the format in which it is written.

2. SEMANTIC WEB

The Semantic Web is changing the way how scientific data are collected, deposited, and analyzed [4]. In this section, a short description defining the Semantic Web is presented followed by the reasons behind the developing of Semantic Web. Next a few selective representation techniques recommended by W3C are presented and a number of successful examples from the commercial do- main that support and use the semantic data are given as well.

Semantic Web is about providing meaning to the data from different kinds of web resources to allow the ma-chine to interpret and understand these enriched data to precisely answer and satisfy the web users' requests [1],[5],[6]. Semantic Web is a part of the second generation web (Web2.0) and its original idea derived from the vision W3C's director and the WWW founder, Sir Tim Berners- Lee. According to [5] Semantic Web represents the ex-tension of the World Wide Web that gives users of Web the ability to share their data beyond all the hidden barriers and the limitation of programs and websites using the meaning of the web.

2.1 Semantic Web: Definition

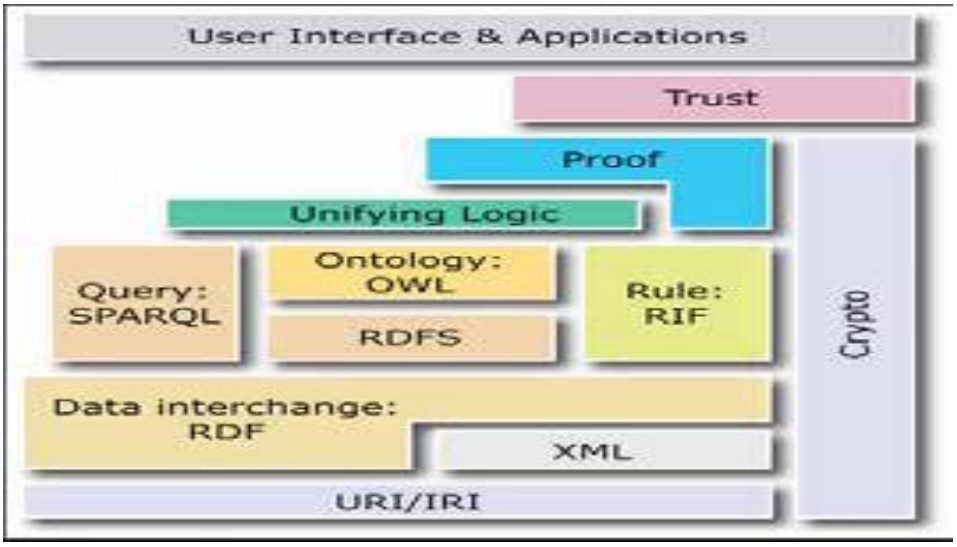

Fig 2: Layers of Semantic Data Mining

These layers are described as follows [16]:

Hyper-Text Web Technologies:-

These are the bottom layer technologies that are well known in the hypertext web domain. These technologies are used to implement semantic web applications.

1. URI
Universal Resource Identifier (URI) is used to identify the semantic web resources. This unique identification is required so as to provide manipulation with the resources in the top layers.

2. UNICODE
It helps to represent and manipulate text in various languages, thus enabling a bridging of the gap between the human languages and semantic applications.

3. XML
Extended Markup Language (XML) is the markup language that is used to create the semantic web documents in the form of structured data. Semantic Web Technologies these are the middle layer technologies, most of which has been standardized by W3C, and can be used to create semantic web applications. All are standardized by the W3C except for RIF/SWRL.

4. RDF
Resource Description Framework (RDF) is the framework that is used to express data in a meaningful way. It expresses data in the form of triples, which is easier to express info in the form of a graph.

5. RDFS
RDF Schema (RDFS) provides the schema, i.e. the vocabulary, for the RDF to maintain a proper structure of the document. It enables to maintain a proper hierarchy of classes and its properties.

6. OWL
Web Ontology Language (OWL) is used to add more meaning, constraints and restrictions to the RDF representation. It expresses the semantics of the RDF statements.

7. SPARQL
SPARQL Protocol and RDF Query Language (SPARQL) is an RDF query language and is used for querying in the database that is represented by the RDF. Querying is done so as to retrieve information by the semantic applications.

Unrealized Semantic Web Technologies
These are the top layer technologies that are not yet standardizes or are ideas that needs to be implemented to completely create semantic web applications.

8. RIF/SWRL
Rule Interchange Format/Semantic Web Rule Language (RIF/ SWRL) is used to add rules to the RDF data. This enables to represent information that cannot be directly expressed by the OWL.

9. Cryptography
This is to ensure that the statements coming from semantic web applications are from proper sources and this can be implemented using digital signatures of RDF documents.

10. Trust
Trust for statements support: the premises come from trusted sources and relying on formal language to retrieve new information.

11. User Interface
This is the top most layer that will enable the humans to use the semantic web applications.

3. SEMANTIC WEB MINING

This section provides a more explained introduction to the Semantic Web Mining followed by few examined problems facing mining the semantic data with their possible solutions (proposed by researchers) and then selective cases examined where obstacles faced traditional, data mining, and Semantic Web systems (and applications), where using the Semantic Web Mining could possibly help to tackle them and proving its usefulness in different domains. A summary of the reviewed research papers is provided at the end.

3.1 Semantic Web Mining Definition

The huge growing in the quantity of semantic data and knowledge in different fields, as the circumstance in bio-medical and clinical scenarios, could possibly create a perfect and important target in the mining process [2],[3]. The Semantic Web Mining came from combining two interesting fields: the Semantic Web and the data mining [1][16].

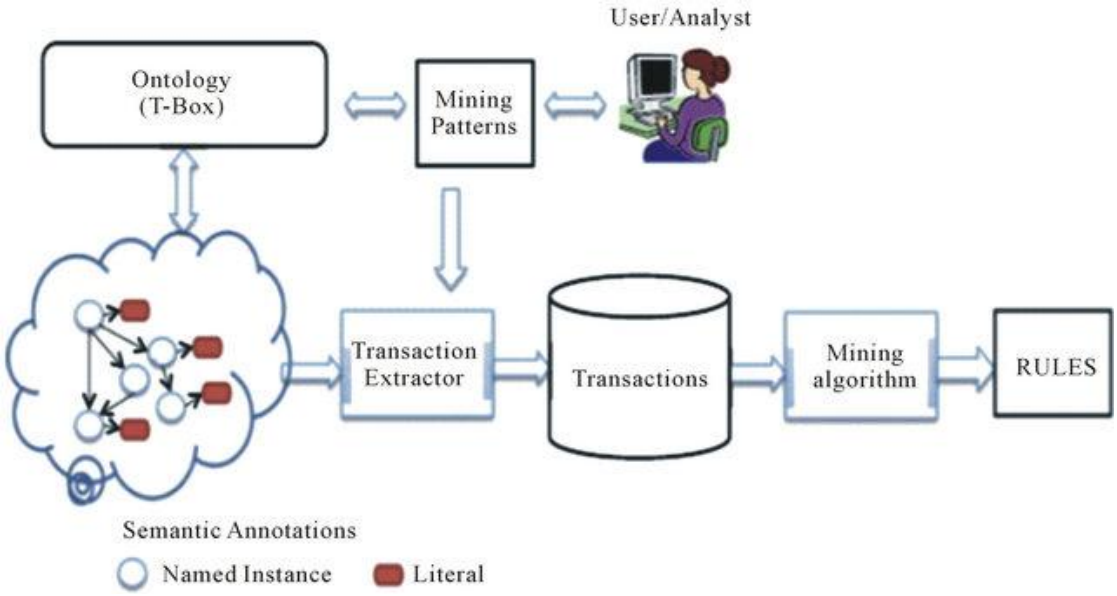

Fig 3: Semantic web mining architecture example

Mining Semantic Web ontologies provides a great possibility to get better results to its domain [3],[11], discovers new and valuable insights data from the semantic annotations [12], solves problems that deals with complex and heterogeneous data [3],[9] and improves in easy, and effective ways the results of the web mining [10],[13].

There is a need to apply and adapt data mining techniques to extract information and knowledge efficiently and effectively, represented in Semantic Web data, and to enhance the way these data are used. The requirement for a shift mining data to mining of semantic data came from adoption of Semantic Web concepts and representations in many different areas such as communities, blogs, search engines and portal, leading to fast growth in the amount of semantic data which shows statistical results from Falcons and Swoogle Se-mantic Web search engines [5][16].

The Semantic Web portal service provided by Twine is another example which reveals this need. Twine saves users' information and interest using RDF and OWL; Twine has more than three millions semantic tags and millions of relations [5].

4. PROPOSED SYSTEM

The proposed system includes following modules

- User Interface Design
- Search Web RDF Data
- Filtering & Table Extraction
- Validation & Storing into RDF Database
- User's Integrated Output

4.1 Module Explanation

- *User Interface Design*

To connect with server user must give their username and password then only they can able to connect the server. If the user already exits directly can login into the server else user must register their details such as username, password, Email id, City and Country into the server. Database will create the account for the entire user to maintain upload and download rate. Name will be set as user id. Logging in is usually used to enter a specific page. It will search the query and display the query.

- *Search Web RDF Data*

In this module, allows the complexity of the querying into different data sources to be hidden to the end user. A user query is an instantiation of a given view by the end user, by specifying, among the set of query able attributes of the view, which are the selection attributes and their corresponding searched values, and which are the projection attributes. An important feature of a user query is that searched values may be expressed as continuous or discrete RDF data sets. A RDF set allows the end user to express his/her preferences which will be taken into account to retrieve not only exact answers.

- *Filtering & Table Extraction*

Recent propositions in the Semantic Web community propose to extract, filter, annotate and query Web data tables, but they have not been designed with the same objectives as ours.

Table Seer for instance allows a set of predefined metadata to be extracted from Web data tables, but it does not compare the schema of the Web data tables with preexisting schemas defined in ontology. We can also cite Web Tables which proposes a system to identify relational tables in a huge amount of tables included in RDF documents and to index them, this in order to query and rank them.

- *Validation & Storing into RDF Database*

In this module, when a query is asked by the end user into the RDF data warehouse which contains RDF graphs generated by our annotation method to annotate XML data tables, the query processing has to deal with RDF data values. More precisely, it has 1) to take into account the certainty score associated with the relations represented in the data tables and 2) to compare a RDF data set expressing querying preferences to a RDF data set, generated by our annotation method, having a semantic of similarity or imprecision.

- *User's Integrated Output:*

The originality of our approach in flexible SPARQL querying is that we propose a complete and integrated solution which allows one 1) to annotate Web data tables with the vocabulary defined in an OTR, 2) to perform a flexible querying of the annotated tables using the same vocabulary and taking into account the fuzzy degrees generated by the annotation method according to their associated semantic.

4.2 Technique used or algorithm used

As evaluating data reliability is subject to some uncertainties, we propose to model information by the means of evidence theory, for its capacity to model uncertainty and for its richness in fusion operators.

Backward search technique

This technique uses different intuitions, which is more scalable and lends significant pruning power without sacrificing the soundness of the result.

Algorithm 1: BACKWARD

Input: $q = \{w_1, w_2, \ldots, w_m\}$, $G = \{V, E\}$
Output: top-k answer $\mathcal{A}(q)$

1 Initialize $\{W_1, ..W_m\}$ and m min-heaps $\{a_1, ..a_m\}$;
2 $M \leftarrow \emptyset$; // for tracking potential $C(q)$
3 **for** $v \in W_i$ **and** $i = 1..m$ **do**
4 **for** $\forall u \in V$ **and** $d(v,u) \leq 1$ **do**
5 $a_i \Leftarrow (v, p \leftarrow \{v,u\}, d(p) \leftarrow 1)$; // enqueue
6 **if** $u \notin M$ **then** $M[u] \leftarrow \{\text{nil}, ..(v,1).., \text{nil}\}$;
7 **else** $M[u][i] \leftarrow (v,1)$; ↑the i-th entry
8 **while not** *terminated* **and** $\mathcal{A}$ **not** *found* **do**
9 $(v, p, d(p)) \leftarrow \text{pop}(arg\min_{i=1}^{m}\{\text{top}(a_i)\})$;
10 **for** $\forall u \in V$ **and** $d(v,u) = 1$ **and** $u \notin p$ **do**
11 $a_i \Leftarrow (u, p \cup \{u\}, d(p) + 1)$;
12 update M the same way as in lines 6 and 7;
13 **return** $\mathcal{A}$ (if found) **or nil** (if not);

4.3 Results and Comparisons

Table 1 Characteristic of Different Search Engines

Search Engine	Characteristics
Google	Page rank technology
Yahoo	Yahoo Slurp and Bingbot as crawler
RDF Search	RDF data and domain ontology

Google is an internet-related services and products. Search engine is one of its popular invented products using PageRank technology. Yahoo is the second larger search engine after Google. Previously, Yahoo was using Google's search engine to obtain results before shifting to Yahoo Slurp and the latest crawler is Bingbot. Google and Yahoo Search have been giving ultimate benefits to internet searchers since 1997 and 2001 respectively. However, some different features in RDF search have advantages in terms of ontological concept, categorization. It is capable to give categorized and personalized results.

Comparison is done between RDF Search, Google and Yahoo. Google is giving 968 millions of results when 'C language' keyword is entered as shown in Figure 9 and Figure show 526 millions of results provided by Yahoo search engine.

Question asked for various search engines:

1) C language
2) Apple fruit
3) Apple phone

The result derived for keyword "C language" below

Table 2 comparison between Semantic and non-semantic search engine

Search engine	Pages found	Time(in millisecond)
Google	96,80,00,000	0.31 sec
Yahoo	526,000,000	0.29 sec
Bing	52,70,00,000	0.30 sec
RDF Search Engine	8	0.05 sec

Table 3 comparison between various semantic search engines

Search engine	Pages found	Time(in millisecond)
SenseBot	3	0.11 sec
Rengine	15	0.09 sec
DuckDuckGo	26	0.06 sec
RDF Search Engine	8	0.05 sec

5. KEYWORD AND SEMANTIC BASED SEARCH RESULTS

Below images shows some more results derived from different queries.

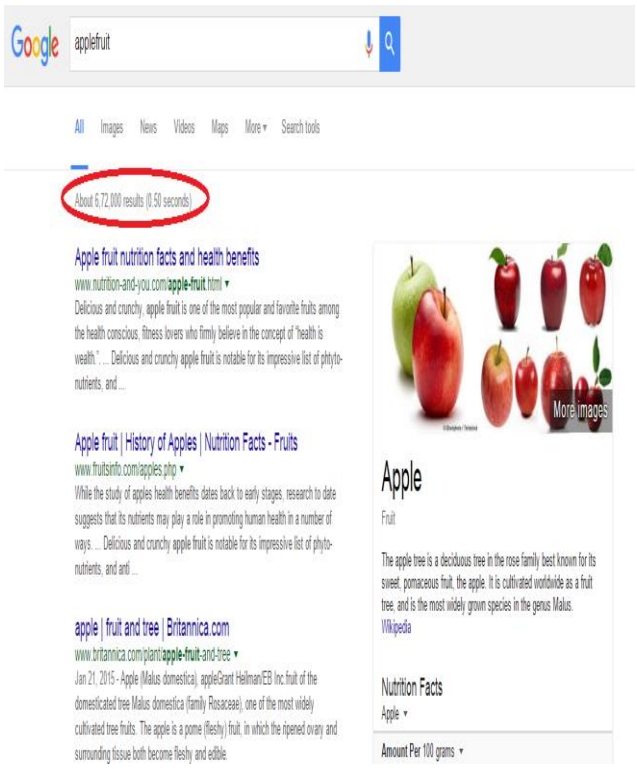

Fig 4: Result of Keyword 'apple fruit' in Google search

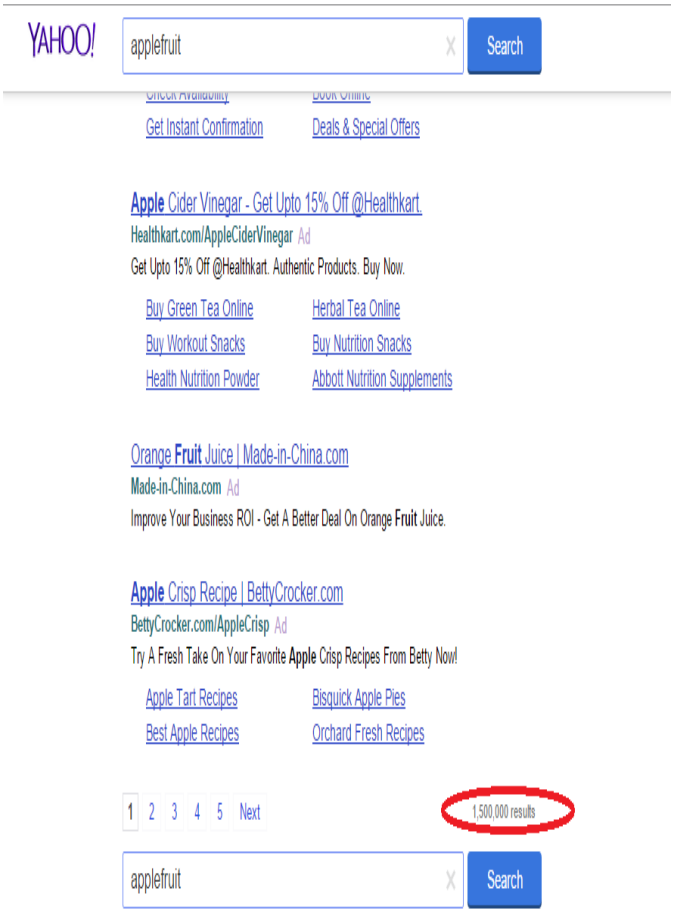

Fig 5: Keyword 'apple fruit' in Yahoo search

Apple fruit nutrition facts and health benefits - Nutrition and You

Apple fruit | History of Apples | Nutrition Facts - Fruits

apple | fruit and tree | Britannica.com

Apple fruit images - All-free-download.com

Apple fruit - All-free-download.com

15 Health Benefits of Apples - Best Health

Fig 6: Keyword 'apple fruit' in RDF search

6. APPLICATIONS

6.1 Semantic Web applications:

RDF is the cornerstone of The Semantic Web, yet there still very few commercial RDF apps. In the latest issue of Nodalities, a magazine about the Semantic Web by UK Company Talis, there is an article by Talis CTO Ian Davis about the state of Semantic Web applications.

6.2 RDF application development for IBM data servers

An RDF store in the DB2 database server is a set of user tables within a database schema that stores an RDF data set. A unique store name is associated with each set of these tables. Each RDF store has a table that contains metadata for the store. This table has the same name as the store.

7. CONCLUSION

Here in this research the problem of scalable keyword search on big RDF data and proposed a new summary-based solution. The research gives a concise summary at the type level from RDF data during query evaluation, this leverage the summary to prune away a significant portion of RDF data from the search space, and formulate SPARQL queries for efficiently accessing data. Furthermore, the proposed summary can be incrementally updated as the data get updated. Experiments on both RDF benchmark and real RDF datasets showed that the solution is efficient, scalable, and portable across RDF engines.

7.1 Future Scope

Furthermore this research can be extended to sentence search as well as image search and video search. In image and video search this RDF technique can enhance up to limited search results same as keywords search. Hence it may enhance the quality of search results as per user query.

8. REFERENCES

[1] V. Crescenzi, G.Mecca and P. Merialdo, "RoadRunner: Towards automatic data extraction from large web sites" In VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases, Roma, Italy, pages 109–118. Morgan Kaufman, Sept. 1.

[2] D.Anglunin, "Inference of Reversible Languages", J. ACM 29(1982) 741-765.

[3] D Anglunin, "On the complexity of minimum inference of regular sets" Inform Control 39 (1978) 337-350.

[4] S.Madria, S.Bhowmick, S.Sourav, W.K.Ng & E.M.Lim "Research Issues in Web Data Mining", CiteseerX Beeta,1999,p.4-12.

[5] A.J Gerber, A.Barnard, A.J Van der Merwe"Towards a Semantic Web Layered Architecture"

[6] T.Berners-Lee, J.Hendler and O. Lassila "The Semantic Web," Scientific American. 284(5):35-43, 2001.

[7] P.DuPont "Regular Grammatical Inference from positive and negative samples by genetic search" the GIG method, In Proceedings of second International Colloquium.

[8] C.N. Hsu and M.T. Dung "Generating finite-state transducers for semi-structured data extraction from the web" Information Systems, 23(8):521–538, 1998.

[9] Lars Marius Garshol (2004) Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all on www.ontopia.net. 13 October 2008.

[10] A. Maedche, S. Stabb (2001) "Ontology Learning for the Semantic Web" IEEE intelligent Systems, Special Issue on the semantic Web, 16(2).

[11] R.Studer, V. Benjamins & D.Fensel "Knowledge engineering, principles and methods" Data and Knowledge Engineering 25 (1998) 161–197.

[12] Gruber Tom (1993): "A translation approach to portable ontology specifications". In: Knowledge Acquisition. 5: 199.

[13] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. daSilva, and J. S. Teixeira: "A brief survey of web data extraction tools." SIGMOD Rec., 31(2):84–93, 2002

[14] E.M Gold "Complexity of automaton identification from given data", Inform Control 37(1978) 337-350.

[15] T. Mitchell (1997): Machine Learning, McGraw Hill. ISBN 0-07-042807-7.

[16] amruta arun joshi, Prof. V.A. Chakkarwar, "A review on Semantic web mining", IJCSIT, Vol 5 (1), pp 431-433, 2015, ISSN No. 0975-9646.

[17] http://www.linkeddatatools.com/semantic-web-basics.