

A Survey : Information Retrieval on Web Using Semantic Web

Mahesh. D.Titiya* Dr. Vipul. A.Shah**

Abstract : In today's world tremendous volume of data in World Wide Web and unstructured nature of data which makes retrieval of valuable information to the people a tiresome task. The information on the World Wide Web cannot be process by machine easily because it is only in human readable form. The idea of Semantic Web and its importance has grown tremendously in the recent years. The use of semantic web and ontologies in information system has become more and more popular in the various fields such as web technologies, database integration, natural language processing etc. Ontologies can be also be used to better organize information resources and assist users in retrieving relevant information. The main purpose of study paper is to investigate recent status of semantic web and flow condition of the Semantic Web with the significant spotlight on querying of data from the web.. This paper incorporates basic segments on the Semantic web and its layered engineering then it discusses the different techniques on semantics and ontology based search. It also discusses information based query and knowledge based search that facilitate the search using semantic web concept. This paper supplements recent studies with latest frameworks for semantic web search with detailed and recent specifications.

Keywords : Semantic Web, Information Retrieval, Ontology, Ontology Learning, Concepts

1. INTRODUCTION

The data and the services on the World Wide Web are fast growing. It is shared by people and the applications across the globe. The modern information system is moving from “data processing” towards to “concept processing”, meaning that the basic unit of processing is less and less an atomic piece of data and it is becoming more a semantic concept which carries an interpretation and exists in a context with other concepts. Semantic Web has become popular as it is providing the common framework for sharing of data across different communities and among different applications. Semantic Web is an extension of the current World Wide Web (www) in which well defined meaning is given to each of the information which provides better enabling to computers and people to work in cooperation. The vision of the semantic web is the idea of having the data on the web are defined with meaning and linked in such a way that it can be used by machine not just for displaying to the user but also for the automation, integration and reuse of the data across various applications. To reach the goal of semantic web, the resource of the web should be annotated with semantic information. Also the semantic web and its application heavily depend on the ontologies to structure of data for comprehensive and transportable machine data. Thus the Semantic Web success is also depending on the quality of ontologies developed for the applications. Appropriate ontologies should be needed for each of the domain for providing basic semantic tool to construct the semantic web.

This paper incorporates starting segments on the Semantic web with layered architecture then it discusses the several methods to search information based on semantic and ontology. It also talks about the different query languages and knowledge based system which provides foundation for search using semantic web. The paper accompaniments recent surveys with new methods for semantic web search with detailed and recent specifications.

* Department of Computer Engineering Government Engineering College, Rajkot, Gujarat, India Email- mdtitiya@gmail.com.

** Department of Instrumentation and Control Dharmsinh Desai University, Nadiad, Gujarat, India Email- vashahin2010@gmail.com

We have also explain the tools ,technology and detailed specification about Semagix Freedom, TAP, OWLJessKB, DLDB, Sesame, Jena and KAON. The semantic web and its layered architecture explained in section 2 and 3. Section 4 depicts diverse sorts of semantics for the Semantic Web. Several techniques for searching are explain in section5 and retrieval of relevant information based on ontology is explained in section 6. The various types of semantic based query language explained in section7 and section 8 and section 9 explains recent and advance technology for the semantic web knowledge base. The conclusion is at last section

2. THE SEMANTIC WEB

The most quoted definition of the Semantic Web is given in the following: “The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.” [20]

The primary thought of the Semantic Web, proposed by Tim Berners-Lee, is to incorporate machine interpretable metadata with the existing information on the web in such a way that world wide web data are not just for displaying purpose but for enabling automation, integration, discovery and reuse of data among various applications. The power of semantic web with sophisticated artificial Intelligence techniques which is used in knowledge representation and acquisition of knowledge from the web. The recent web technologies and use of hypertext provides the meaning to web data which enables the development of more powerful web applications. The web data is looks like a global database. In the semantic web the languages are developed for representing information which leads to the information retrieval from the web an easy task.

3. THE LAYERED STRUCTURE OF SEMANTIC WEB

The Semantic web layered structure describes the layers of the semantic web and their components and its relationship of layers which is shown in figure1 [30]. We give a brief description of each of the layer of the semantic web.

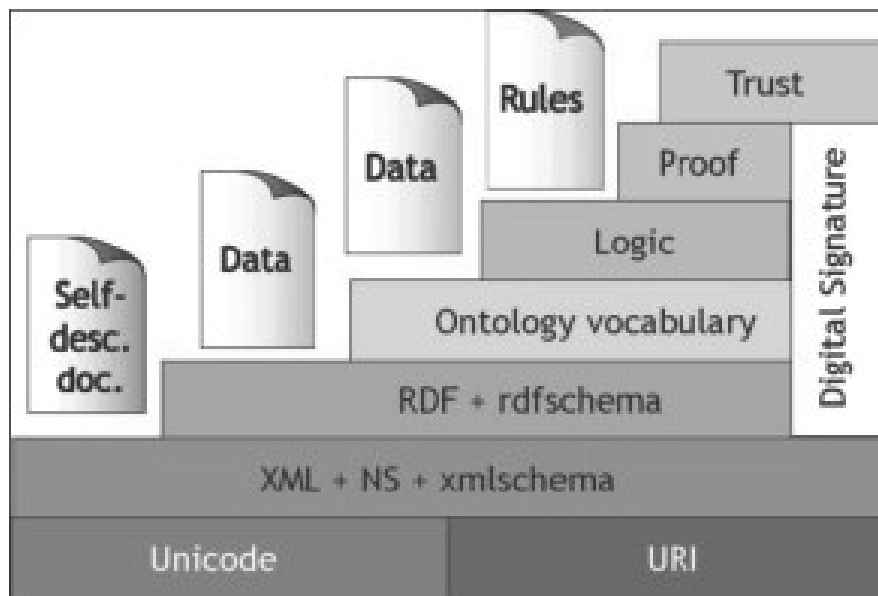


Fig. 1. The Semantic Web Layered Structure [30]

3.1. Layer 1- URIs and Unicode

In the semantic web each of the objects is identified by the Unique Uniform Resource Identifier which is assigned to the object. The subclass of the URIs is Universal Resource Locators (URLs) and Uniform Resource Name (URNs). The semantic web includes resource like images, video or audio files, people, events etc. Each resource on the web has unique URLs. There are no any standards are developed to assign URLs to resource of the web Unicode is a character set that can deal with multiple human languages.

3.2. Layer 2-XML and Namespaces

XML namespaces are used for providing name to elements and attributes in an XML document. They are defined in a W3C recommendation. An XML instance may contain element or attribute names from more than one XML vocabulary. If each vocabulary is given a namespace, the ambiguity between identically named elements or attributes can be resolved.

The Semantic web metadata uses XML syntax. Extensible Markup Language (XML) [24] is a standard text format for serializing data using tags. Since a decade XML is available and it has many technologies and tools available for XML data processing, such as DOM and SAX parsers, DTD and XML Schema validation, XPath and XQuery query languages, XML databases, etc. The extensions of the XML are XML Namespaces [23] which provide the means to identify the element in the vocabulary uniquely. Where the vocabulary consists of XML element types and attributes names. In other words the namespace mechanism defines a URI to indicate the vocabulary and an element name to indicate the element in the vocabulary. Due to overlap of vocabularies in many XML documents there is problem of collision and recognition..

3.3. Layer 3-RDF

Resource Description Framework (RDF) [15] is a language which is used for representing information in the Web. In the RDF modeling a RDF graph is used in which the nodes are represented by RDF URI reference, blank nodes or plain literals and the edges are labeled with RDF URI references. The example of the RDF graph is shown in figure 2[33]. The RDF graph contains triples, where each triple consists of a subject, a predicate and an object. Each triple translates into a statement about a resource. The RDF graph sample is shown in figure 2[31]. The RDF graph contains the triples, where every triple comprises of an object, subject and a predicate.

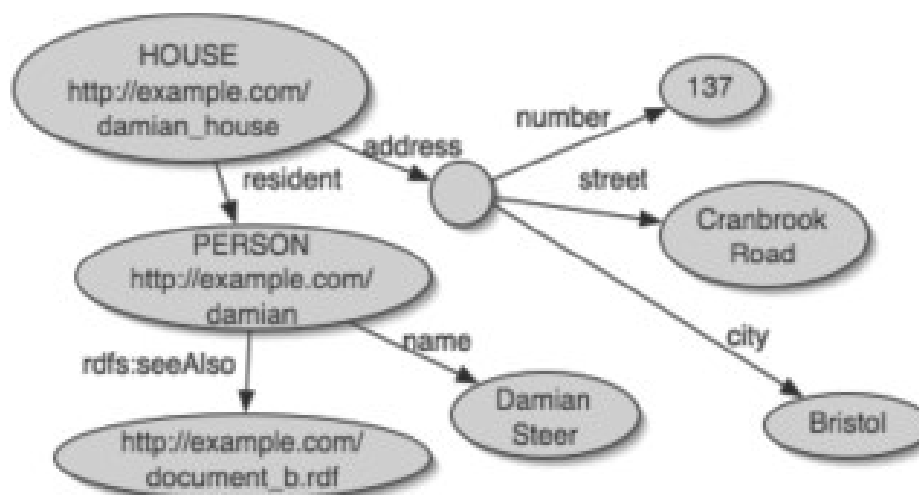


Fig. 2. RDF Graph Example [31]

3.4. Layer 4-RDF Schema

RDF Schema (RDFS) [25] is a language used to describe RDF vocabularies. RDFS describes class hierarchies and property hierarchies and labels to URIs. It also describes the constraints, domain and range of properties etc. Although RDF Schema and XML Schema [34] are both “schemas”, but used for different purposes: RDFS is used for inference and XMLS is used for validation.

3.5. What is Ontology?

The ontology definition given by Gruber [36]: Ontology is a “formal specification of a conceptualization”, and is shared within a specific domain. In different words ontology is a document which is defined in a language like RDF Schema which describes the vocabulary of the terms and concepts and their relationship is used for the specific domain.

3.6. Layer 5-Ontology Languages

The Ontology Language DAML+OIL [32] and Web Ontology Language (OWL) [13], are more complex than RDFS and provide more capabilities to define ontologies. Web Ontology Language(OWL) extends RDFS and successor of DAML + OIL.OWL has the following features: property characteristics (Transitive Property, Symmetric Property, Functional property,inverse Of, InverseFunctionalProperty), property restrictions (allValuesFrom, someValuesFrom, cardinality, has Value), ontology mapping (equivalence between classes and properties equivalent Class, equivalent Property; identity between individuals), set operators (intersection Of, union Of, complement of etc. The OWL language has three expressive sublanguages: OWL Lite is used for classification hierarchy and for simple constraint features, OWL DL which guarantee maximum expressiveness and reasoning capabilities and decidability.OWL Full allows mixing of OWL with RDF Schema

3.7. Layer 6 - Rules

To discover the knowledge rules are used in the semantic web. Presently there are no any standard for creating rule in the semantic web. In languages like Prolog the rule statement “female(X): “daughter(X,)” might mean that an object is a female if this object is a daughter of some other object. Backward chaining or top-down resolution is used for the reasoning. The same idea is used in “deductive databases”, where reasoning rules are stated in languages with Prolog-like syntax. Closed World Machine (CWM) is the example of a data processor for the Semantic web.

3.8. The Other Layers

The other layers of the semantic web represent reasoning in the semantic web. For eg. Description Logic is used for reasoning. Encryption and Signature of XML data which provides proof and trust in the semantic web.

4. SEMANTICS FOR THE SEMANTIC WEB

There are three types of semantics for the semantic web: the implicit, the formal and the powerful. The implicit semantics are not stated explicitly and it extracted by visualizing or extracting data from the patterns. For example the occurrence of the keyword, hyperlinks. Position in the concept hierarchy etc. semantic allows the finding relevance data to some semantic context. However it is not machine processable and also not possible to name a relationship between concepts. The formal semantics is presented in some well-formed syntactic language. The formal semantic have the following features: (1) the notions of model and model theoretic semantics language expressions are interpreted in models which reflect “structure of the world”, and (2) the principle of compositionality expression meaning is a function of the meanings of expression’s parts and of the way they are syntactically combined. Examples of such languages are RDF, OWL, and Description Logics. This type of semantics is machine processable. The major drawback of the formal semantic is that it becomes impractical as the knowledge size is increases as the knowledge is added from the different sources. The powerful semantic has the features of implicit and formal semantics. It derives the relationship using fuzzy logic and probabilistic theory. The relationship derived which is valid one. The major drawback is that probabilities need to assign in powerful semantic. In summary the current web mostly exploit the implicit semantic which is present in the web pages. In the semantic web major focus is the formal and powerful semantic.

5. TYPES OF SEARCH

There are two kinds of searches Navigational Searches: In this class of searches, the user providing word or combination of words or phrase for finding the relevant documents. These words are not denoting as a concept. The user is using search engine such as google to perform the navigation to reach to particular intended document. Research Searches: In many of the cases the user provides the phrase to search engine which denote the object for which the user is trying to search information. The user does not know about the document for which he/she is trying to get to. Rather, the user is trying to search for multiple documents which together will give him/her the information he or she is trying to find. Example: A search query like “World Wide Web track 2pm Panel” does not

define any of the concept. The user is just trying to find the web pages containing all the words of search query. On the other hand, search queries like “Isaac Newton” or “Dublin Ohio”, denotes a person or place. The user is likely doing a research search on the person or place denoted by the search query [26]. Both types of searches can be improved by exploiting significant domain ontology and annotations.

6. ONTOLOGY BASED INFORMATION RETRIEVAL MODEL

In this section, we explore the changes which semantic web brings to the information retrieval (IR) model. The view of the classical IR model is shown in Figure 3 [18]. In the IR model, a query is formulated from information which is in need matched over document representations (*e.g.* index structures).

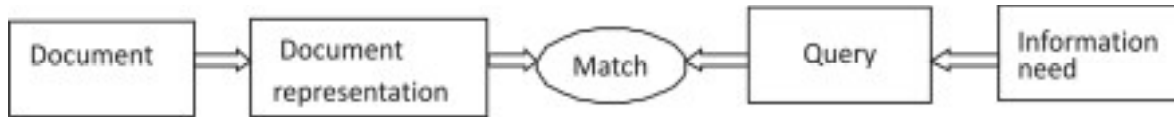


Fig. 3. A Classical Information Retrieval Model [18]

7. KNOWLEDGE BASED SYSTEMS FOR THE SEMANTIC WEB

There are many knowledge base systems are in existence such as RDF(S), DAML + OIL and OWL which can be used as a basis for semantic web for semantic web repositories. We are providing brief description of several such systems, describing their general architecture and functionality, also the detail given on storage design, querying of data and support for the reasoning. We provide brief introduction into several such systems, describing their general functionality and architecture, further details on storage design, querying and reasoning support.

7.1. Semagix Freedom

Semagix Freedom [7, 34] is the business level system which is providing basis for the semantic web application development. It mainly includes Classification of content automatically, Extraction of Ontology driven metadata supporting for process of complicated query using ontology., Construction of ontology, Summarization of content, information creation and Summarization, Querying of knowledge base content, Ontology exporting in RDF/RDFS with rules which are not articulated in RDF/RDFS. The architecture of system is shown in Figure 4 [9]. It provides the modeling tool to build ontology. The role of Knowledge agents to automatically maintain the ontology. Semagix Freedom is providing modeling means to construct ontology, Knowledge Agents (KA) without human intervention maintain the ontology. The Content Agents (CA) extracts the data which are unstructured or semi structure or structured in nature. The data are collected from different source. There is no any programming language or special toolkit required for Knowledge Agents and the Content Agent.

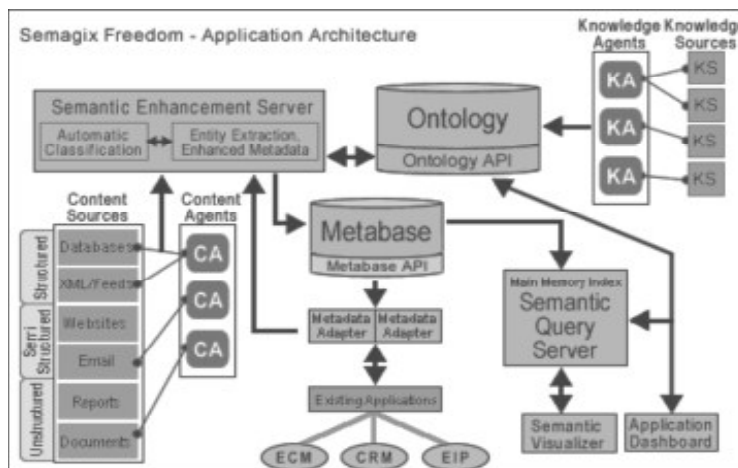


Fig. 4. The Architecture of Semagix Freedom[9]

The retrieved content is further enhanced by Semantic Enhancement Server which recognizes concern documents characteristics such as date, currency, etc. The metabase is saved in the form of table and its snap residing in the main memory to boost questioning. Semantic Query Sever which gives the facility for the querying with the use of HTTP and Java APIs returns the result in XML

7.2. Semantic Search and TAP

Semantic Search is a use of the semantic web to execute request and TAP is an application framework for building semantic web applications. TAP is having the following characteristics:

- **Interface for querying :** Getdata which is supported by repositories of Semantic Web repositories to performing search based on semantic.
- **Scrapping :** TAP is providing supports for interpretation of request which is generated by query interface Getdata. So that a customer can imagine that the site offers a GetData interface to its information.
- **Publishing :** On the server side, an Apache HTTP server module is running which is utilized for uncovering information through the GetData interface. Information is distributed to the client as RDF comments, TAP Apache incorporates and mapping of RDF documents to the chart structure to maintain a strategic distance from RDF parsing when questioning is performed.
- **Registries and caching :** The different servers which monitors which URL has values for which legitimate ties about which classes of assets. A customer can guide the question to the registry which diverting to proper sites. Registries additionally reserve the reactions to GetData asks for productivity. GetData is a giving interface for questioning which is based on SOAP convention .SOAP convention utilized for performing Remote Procedure Call. By method for GetData, a customer program can get the characteristics of resource from a RDF chart. Each GetData question is a SOAP message tended to that URL. The message determining two contentions: the asset whose properties are retrieved and the properties that are being to retrieve [25].

7.3. Sesame

Sesame [27] is a RDF system which support for the RDF Schema inferencing. It has the components for questioning of information in three dialect (SeRQL, RDQL, and RQL), parsing and writing in various sentence structures and supporting for the MySQL, PostgreSQL, It conveyed as a RDF database, with determination in RDBMS, or as a Java library for building various applications [27].

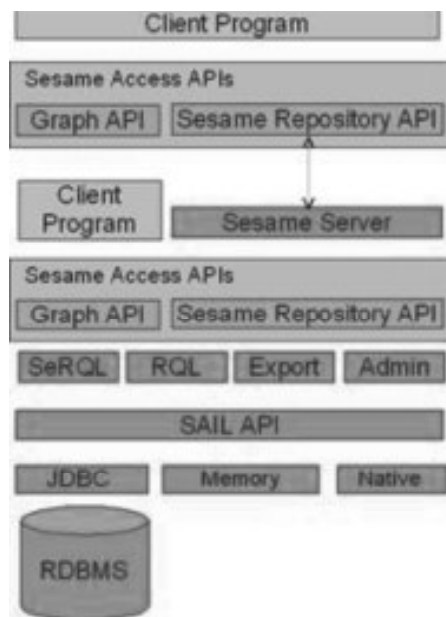


Fig. 5. Sesame Architecture[27]

The Sesame design is shown in figure 5[27]. Customer programs use sesame access API to get to the information from the server locally or remotely through HTTP and RMI Protocol. Useful modules resemble Query Module (SeRQL, RQL, RDQL), Export module utilized for sending out information into RDF(S) format. An Admin Module has the regulatory functionalities, for example, storage of data and interface utilizing SAIL API. The administrator module main layer of Sesame which gives interface to programming that uniquely distinguish abstract storage device and take care in deduction of learning [28]. As we have expressed before Sesame can store RDF information into relational database, for example, MySQL, Oracle and SQL Server or into object relational database such as PostgreSQL. Following we depict the relational schema for these two storage approaches.[28]

7.4. DLDB

DLDB [19] is a kind of knowledge base system which expands relational database management system with inferencing using ontology. It also provides support for RDFS, DAML+OIL, OWL, and MS Access for storing annotation and knowledge inferencing. The relational schema is constructed by utilizing ontology: A table is produced for every class with property using ontology. Further, for every table, a view is created which is having duplicate data and additionally logical view is created which stores the inferred data..The DLDB Queries are using system API and queries are expressed in knowledge interchange format. The queries are converted into structure query language for generated views.

7.5. DAMLJessKB and OWLJessKB

DAMLJessKB[23] and OWLJessKB[7] are the logical reasoner having the capabilities of depiction, based on memory for DAML and Ontology language respectively. The primary tools for both the system shell which perform constrain base reasoning and knowledge stored in the memory. The fundamental methodology of DAMLJessKB and OWLJessKB is mapping of RDF XML linguistic structure into a collections of triples, mapping into information in constraint based system. The constraints are retrieved from the semantics of the dialect. The role of RDF parser is to generate the triple stream from the XML document. The triplets are inserted into the rule based system which generates the rule with extra information for the knowledge base.[36].

7.6. Jena

Jena [29] utilizing structure of java for constructing Semantic Web applications, which gives an automatic foundation to RDF, RDFS and OWL, along with constraint based inference engine. Jena stores RDF diagrams in primary memory or in a database and it underpins two types of questioning: triple match and RDQL [7]. Following we describe RDF storage, questioning and inference of knowledge Jena. The evolution of RDF storage from Jena1 to Jena2. The Jena1 utilized the triplets which are in standardized structure and it is stored. The storage table stores all the triples statements re subject, predicate, object. The Jena1 utilizes the standardized mapping for MySQL and PostgreSQL and Oracle. However Berkley Database utilized denormalized outline which is putting away triples in a solitary table. This methodology appears to particularly effective in space, but still three way join is required to get triple. The schema used by Jena2 is demoralization of data. The literals and resource table are utilized to accumulate the values which have URIs and literals. As the length exceeds above threshold value when their length exceeds some threshold it can take more space. However it shows better performance in the retrieval of the data. Jena2 provide foundation for the property class tables.

8. SUMMARY TABLE OF KNOWLEDGE BASED SYSTEM FEATURES

We have summarized our discussion of existing knowledge base system in Table 1. We have also given detailed evaluation of systems such as Sesame, DLDB, DAMLJessKB and OWLJessKB with the criteria such as inference support, scalability and update support etc.

Table 1. Summary of features of Different Knowledge based System

<i>Knowledge Based System</i>	<i>Types of Storage</i>	<i>Database Schema</i>	<i>Support of update</i>	<i>Support for Inference</i>	<i>Supported Query Language</i>	<i>Measure of Scalability/ Performance</i>
Semagix Freedom	Relational Database System	Relations of Tables	Yes	Yes	Ontology and Metabase API used	More than million of instance; 64 concurrent users can query, 10 ms for query resolves.
TAP	File system, Relational Database System, (MySQL and, Berkley DB)	Triple store (Subject, Object, Predicate)	No	No	Get Data (Resource Description Graph)	Relation contains thousands of records
Sesame	Main Memory, ORDBMS (Postage SQL), RDBMS (MySQL, Oracle, etc.)	Triple store (Subject, Object, Predicate)	Yes	Yes (it saves into Database)	SeRQL, RQL, RDQL	Tested on 3 (Database B) and 1 (memory) million triples (58.1 MB) is too long (order of weeks); linear scalability on querying
DLDB	Relational Database Management System for access	Specific to application (A different tables for each class instance and property.	Yes	Yes (saves into views and FaCT DL)	Knowledge interchange formate API is used.	Tested on 6.8 million triples (583MB), linear scalability on data loading, repository size and querying.
DAML Jess OWL Jess KB	Memory	Not Applicable	Yes	Yes; Jess reasoner	Jess	Can load up to a few hundred thousand Triples; bit slow; soundness problem.
Jena	Memory, ORDBMS (Postgre SQL), RDBMS (MySQL, etc.)	Triple store (Subject, Object, Predicate) Data mining based on application specific	Yes	Yes; built-in or external reasoner	Triple match, RDQL	Tested on small data sets-ten thousands of triples; querying is, in general, slower then for Sesame [39]
KAON	Memory, file system, ORDBMS (Postgre SQL), RDBM(SQL Server, Oracle, etc.)	Metamodel structure, Triple store (Subject, Object, Predicate)	Yes	Yes ; external reasoner only	RDF-QEL (Datalong-like)	Can load upto thousands triples. Beyond that it is slow

9. OTHER KNOWLEDGE BASED SYSTEMS

There are many other systems which supporting RDF(S)/DAML+OIL/OWL storage, querying of data and reasoning capabilities. 4Suite, DAML DB, EOR, Haystack, Inkling, Parka Database, rdfDB, RDFLib, RDFStore, RDF- Suite, Redfoot, Redland, Edutella, etc. Summary information about all knowledge based system can be found in survey paper [14], [16] and [34]. Our survey paper accompaniments with those surveys with new knowledge base systems such as Semagix Freedom, TAP, OWL- JessKB, DLDB as well as with more advance and recent specifications for Sesame, Jena and KAON.

10. CONCLUSIONS

In this survey we have reviewed papers of Metadata based searching, Semantic web search, use of semantics for search, Ontology based searching, query language for ontology based search and knowledge based system that enables semantic web search facility. The paper accompaniments existing surveys with new systems which is available with more detailed and recent specification of such systems. The conclusion drawn from this survey is follows. 1) Existing RDF(S) query languages are not complete; it is also having lack of expressivity and explanation. 2) Lack of specific standards for query language, rules, inferencing of result. Due to which the inferencing result which got from knowledge base are incompatible and it is difficult to compare, exploit, and lean. So forth. It is not always apparent which inferencing path that a reasoner should follow as a correct path.

26. Y. Guo, Z. Pan, and J. Heflin. An evaluation of knowledge base systems for large OWL datasets. In Third International Semantic Web Conference, Hiroshima, Japan, pages 274–288, 2004.
27. P. Haase, J. Broekstra, A. Eberhart, and R. Volz. A comparison of RDF query languages. In The Semantic Web - ISWC2004. Proceedings of the Third International Semantic Web Conference, 2004.
28. G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis, and M. Scholl. RQL: A declarative query language for RDF. In The Eleventh International World Wide Web Conference (WWW'02), 2002.
29. J. B. Kopena and W. C. Regli. DAMLJessKB: A tool for reasoning with the Semantic Web. In 2nd International Semantic Web Conference (ISWC2003), 2003.
30. S. Lu, M. Dong, and F. Fotouhi. The Semantic Web: Opportunities and challenges for next-generation Web applications. *International Journal of Information Research*, 7(4), July 2002.
31. A. Magkanaraki, G. Karvounarakis, T. T. Anh, V. Christophides, and D. Plexousakis. Ontology storage and querying. Technical Report No 308. April 2002. <http://139.91.183.30:9090/RDF/publications/tr308.pdf>.
32. B. Motik, D. Oberle, S. Staab, R. Studer, and R. Volz. KAON server architecture. Technical Report 421, University of Karlsruhe, Institute AIFB, 76128 Karlsruhe, Germany. 2002. <http://wonderweb.man.ac.uk/deliverables/documents/D5.pdf>.
33. M. Olson and U. Ogbuji. Versa. <http://uche.ogbuji.net/tech/rdf/versa/>.
34. Z. Pan and J. Heflin. DLDB: Extending relational databases to support Semantic Web queries. In Workshop on Practical and Scaleable Semantic Web Systems, ISWC, pages 109–113, 2003.
35. J. R. Searle. Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3, 1980. Also available <http://www.w3.org/DesignIssues/Semantic.html>.
36. T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 6(2):199–221, 1993.
37. R. Guha, R. McCool, and E. Miller. Semantic Search. In The Twelfth International World Wide Web Conference, May 2003.
38. Y. Guo, Z. Pan, and J. Heflin. Choosing the best knowledge base system for large semantic web applications. In Thirteenth International World Wide Web Conference (WWW2004), pages 302–303, 2004.
39. Y. Guo, Z. Pan, and J. Heflin. An evaluation of knowledge base systems for large OWL datasets. In Third International Semantic Web Conference, Hiroshima, Japan, pages 274–288, 2004.
40. P. Haase, J. Broekstra, A. Eberhart, and R. Volz. A comparison of RDF query languages. In The Semantic Web - ISWC2004. Proceedings of the Third International Semantic Web Conference, 2004.