CHAPTER 1

INTRODUCTION

1.1 PREAMBLE

This chapter presents a general idea of the proposed research work to be introduced in this thesis. Section 1.2 presents generations of web. Section 1.3 describes about semantic web and related technologies. Section 1.4 depicts on information retrieval, semantic search, and point out their challenges of semantic search. Section 1.5 provides motivation for present research. Section 1.6 lists the objectives of the thesis. The consecutive sections discuss the overview of the problem statement, scope and contribution of the thesis. Section 1.10 gives an organization of the thesis, it has introduced in each remaining chapters of the thesis.

1.2 GENERATIONS OF WEB

The World Wide Web has changed the way of communication, business process, searching information and entertainment. It has an impact on our daily lives. The web is the largest transformable information construct and the progress of the web and its related technologies were faster in the past two decades.

Web 1.0 is the, “readable” phrase of the World Wide Web (WWW) with flat data and is known as static web and is the first of its kind. In web 1.0, there is only limited interaction between sites and web users. It is simply an information portal where users ostensibly receive information without
providing prospects to post reviews, comments and feedback. Web 2.0 is the “writable” phrase of the WWW with communicating data and it is called syntactic web. Unlike web 1.0, web 2.0 facilitates interaction between web users and sites. Therefore, it allows users to intermingle more freely with each other. The web 2.0 supports information distribution, collaboration and sharing. For example, web 2.0 applications are Facebook, YouTube, wiki, Flickr, etc.

Web 3.0 is the “executable” phrase of WWW with dynamic applications, interactive services, and “machine-to-machine” interaction. It is known as semantic web. In web 3.0, computers can interpret information like humans and cleverly generate and distribute useful content tailored to the needs of users. One example of web 3.0 is Tivo, a digital video recorder. It is connected to a home network, including movie and television show downloads, personal photo viewing, music offering and online scheduling based on user preferences. Table 1.1 displays the summary of the three kinds of web (Umesha & Shivalingaiah 2008).

Table 1.1 Summary of Web 1.0, Web 2.0 and Web 3.0

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Characteristics</th>
<th>Web 1.0</th>
<th>Web 2.0</th>
<th>Web 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>Generation of Web</td>
<td>Static Web</td>
<td>Syntactic Web</td>
<td>Semantic Web</td>
</tr>
<tr>
<td>3.</td>
<td>Originator</td>
<td>Tim Berners Lee</td>
<td>Tim O’Reilly</td>
<td>Tim Berners Lee</td>
</tr>
<tr>
<td>4.</td>
<td>Technology</td>
<td>Read only</td>
<td>Read and write</td>
<td>Read, write and execute</td>
</tr>
<tr>
<td>5.</td>
<td>Language</td>
<td>HTML</td>
<td>XML</td>
<td>RDF/OWL</td>
</tr>
<tr>
<td>6.</td>
<td>User Interaction</td>
<td>Information sharing</td>
<td>Interaction</td>
<td>Immersion</td>
</tr>
<tr>
<td>7.</td>
<td>Number of users</td>
<td>Million of users</td>
<td>Billion of users</td>
<td>Trillion of users</td>
</tr>
<tr>
<td>8.</td>
<td>Digital natives</td>
<td>Ecosystem</td>
<td>Participation</td>
<td>Understanding itself</td>
</tr>
<tr>
<td>9.</td>
<td>Technical sophistication</td>
<td>Connect information</td>
<td>Connect people</td>
<td>Connect knowledge</td>
</tr>
<tr>
<td>10.</td>
<td>Development</td>
<td>Brain and Eyes (Information)</td>
<td>Brain, Eyes, Ears, Voice and Heart (Passion)</td>
<td>Brain, Eyes, Ears, Voice, Heart, Arms and Legs (Freedom)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pushed web, text/graphics based flash</td>
<td>Two way web pages, 2D portals Wikis, video, shading. Personal publishing, podcasts</td>
<td>3D portals, avatar representation, Interoperable profits, multiuser virtual environments, Integrated games, education and business</td>
</tr>
<tr>
<td>---</td>
<td>-------</td>
<td>--------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>12.</td>
<td>Web Application</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Companies publish content, the people consume. For example, Cable News Network - CNN</th>
<th>People can utilize the other people published contents, companies build platforms that let people publish content, for other people. (e.g. RSS, Flickr, Adsense)</th>
<th>People construct applications that people can interact with, Companies develop platforms that allow people publish services by leveraging the associations between people (e.g. Google maps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.</td>
<td>E-Commerce evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Search engines retrieve macro contents, but many times results are inaccurate.</th>
<th>Search engines retrieve tags with micro contents manually. Web 2.0 tags everything: picture, links, events, news, Blogs, audio, video and so on.</th>
<th>Search engines retrieve micro content texts which were tagged automatically. This implies translating billions of web1.0 macro contents into micro contents. The results more precise search because tagging can solve part of the ambiguity that homonyms and synonym introduce into the process of search.</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.</td>
<td>Process</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>AltaVista, Google</th>
<th>Google, HAkia personalized, DumpFind</th>
<th>Intellidimension, Falcon SWoogte, Thinkglue, SWSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.</td>
<td>Search Engines</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>The web in the beginning when it was first developing web 1.0</th>
<th>Allow a much more sophisticated user interaction with web pages social networks and wiki web 2.0</th>
<th>Thought to be the future where the web is more interactive with users, leading to a kind of artificial intelligence web 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.</td>
<td>Version</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The rapid growth in information on the web has made the searching more difficult to find relevant information. The users normally suffer from troubles in finding accurate information from the web. At present, the user gives a set of keywords as input to a search engine, which returns as a list of
pages that are related to the keyword terms or phrases. Subsequently the result in the page sets, create too much irrelevant information, the relevant information required for a user is retrieved only after a few keyword adjustments. Therefore, the users are expected to have effective search techniques to find the relevant information easily and precisely from the web. However, most of the existing search engines provide information most of which may not be relevant to the user queries. Therefore, semantic web becomes active and popular in research field, since they help to find relevant information from the web.

1.3 SEMANTIC WEB

Berners-Lee et al (2001) introduced semantic web as an extension of the syntactic web. The main goal of semantic web is to make web contents not only human readable, but also machine readable and processable. Berners-Lee defined “the semantic web as a meaningful web and it is a mesh of information linked up in such a way as to be easily processable by machines on a global scale”. It supports more efficient searching, automation, integration, reuse of data and it brings structure to the meaningful content of web pages, generating an environment where software agents, roving from web page to web page, can gladly carry out sophisticated tasks for users. The semantic web provides a number of technologies to improve human and computer collaboration on the internet. It is developed to overcome the following problems of current web technologies.

- The current web content lacks a proper structure about the representation of information.
- Vagueness of information resulting from a poor interconnection of data.
- Automatic data transfer is lacking.
- Usability to deal with vast number of users and content, ensuring trust at all levels.
• Inability of machines to understand the provided information due to lack of a universal format.

1.3.1 Semantic Web Architecture

The semantic web is a meaningful web in which information is given well-defined meaning. The semantic web architecture is shown in Figure 1.1 referred to as the semantic web layer cake by the WWW Consortium (W3C). It has several layers arranged as protocols stack in computer networks (namely ISO/OSI stack protocols). Every layer presents a set of duties and a set of technologies. The semantic web architecture is explained as follows:

Figure 1.1 Semantic web architecture
**Uniform Resource Identifier (URI) and UNICODE:** The URI & UNICODE are responsible for encoding any character and at the same time it is also responsible for uniquely identifying different resources. Unicode technology is suitable for the job of encoding and URI is suitable for representing and identifying uniquely different resources.

**XML:** XML (eXtensible Markup Language) has the capability of being used as a base syntax for other technologies developed for upper layers of the semantic web architecture. It is processed by machines to help the communication between human and machine.

**RDF:** Resource Description Framework (RDF) is a language used to describe documents written in a structured & machine descriptive syntax language to ensure that the documents are written correctly according to recommendations.

**RDF Schema:** RDF Schema contains the language that provides meaning to the semantic web architecture by representing metadata to be accessible and processable by machines.

**Ontology:** Ontology is defined as the collection of terms used to describe a domain of knowledge that is shared by people, databases, and applications. It is an explicit and formal specification of a conceptualization. Ontology’s are the backbone technology for the Semantic Web. Web Ontology Language (OWL) is one of the languages for creating ontology which is very similar to RDF, but it is a weightier language with better machine-interoperability than RDF. It describes the exact nature of resources and relationships between them. The created OWL for the particular application is stored in the repository.

**SPARQL:** SPARQL (Simple Protocol And RDF Query Language) is a protocol and query language for semantic web data resources. SPARQL aims to let people work upon the data without bothering them about its internal organization. Hence, SPARQL has been designed to provide an interface to query disparate data having multiple dimensions spread across diverse data
sources. It encompasses capabilities to integrate vast amount of the existing Web 2.0 data with other semantic web utilities and services. This allows for navigation through different levels of hierarchy. Therefore, SPARQL enables people to share, merge, and reuse data globally, fulfilling the goal of the Semantic Web.

**Rules layer:** Rule layer aims to support an inference, in addition to allow queries and filtering. Rule-languages and rule-based systems have played important roles in the history of computer science and the information technology. With the development of the Web, in particular, the Semantic Web, more research is needed in this area as shown by the “Rule” component in the layer cake.

**Logic Framework:** The logic Framework layer provides the answer for the question of why this piece of information is taken or appears to the user. This layer includes predicate logic and quantifiers so as to facilitate deductions. Knowledge Interchange Format (KIF) is the language used to specify logic in this layer.

**Proof:** The Proof layer is assumed to answer agents about the question of why they should believe the results. The proof layer main aim is to provide explanations about the answers given by automated agents that consume the provided information. At present, there is no technology recommended by W3C to this layer, but there is an attempt for developing a proof language by the knowledge systems laboratory at Stanford University.

**Trust:** Trust in the semantic web area which can be concluded as to be sure that the information provided, is valid and the degree of confidence in the resource providing this information.

**User Interface:** User interface enables to use semantic web applications for humans.
1.3.2 Semantic Web Tools

This section provides some tools related to semantic web such as Jena, Protégé, Bossam, Netica-J, Pallet and semweb4j. The tools possible to compile a list of hundreds of components that in one way or another can be used in building or extending semantic webs.

**JENA:** Jena is a Java framework for building semantic web applications developed by HP labs. It has a rule-based inference engine and provides a programmatic environment for RDF, RDFS, OWL and SPARQL. The Jena framework includes an RDF application programming interface (API), reading and writing RDF in RDF/XML, N3 and N-Triples, OWL API, In-memory too persistent storage and SPARQL query engine. It also includes a document manager that assists with the process of managing imported ontology documents. The RDF API provides statement centric and resource centric methods for manipulating an RDF model as a set of RDF triples and a set of resources with properties respectively. It also has cascading method calls, built in support for RDF containers, integrated parsers and writers for RDF/XML, N3 and N-TRIPLES.

**Protégé:** Protégé is a free, open-source platform to construct domain models and knowledge-based applications with ontologies. Ontology is an explicit specification of a conceptualization. The depictive primitives are mostly classes or sets, attributes or properties, and relationships or relations among class members. The descriptions of the expressive primitives include information about their meaning and constraints on their logically consistent application. Ontology is a set of concepts - such as things, events, and relations - that are specified in some way such as specific natural language in order to create an agreed-upon vocabulary for exchanging information. Domain Ontology stands for the specific meanings of terms as they apply to
that particular domain (ex. chemicals, Gene Ontology, Tour ontology). The research uses OWL based ontology to improve the efficiency of the search.

**Bossam:** Bossam is a rule inference engine for the semantic web. It supports reasoning over OWL ontologies, SWRL ontologies, and RuleML rules. Bossam includes several expressivity features such as URI references as symbols, second-order logic syntax, URI-based Java method attachment, and support for both negation-as-failure and classical negation. Bossam rule engine used for loading, inference, and querying over the set of documents. The set of documents can contain any combination of the RDF(S) documents (in RDF/XML or in N3), OWL documents (in RDF/XML or in N3), Bossam rule documents and SWRL (+OWL) documents (in OWL or in RDF/XML).

**Pellet:** Pellet is an open source OWL Description Logic (DL) reasoner. Pellet has support SWRL and provides standard and cutting-edge reasoning services for OWL ontologies. Pellet is the leading choice for systems applications that need to represent and reason about information using OWL DL. It combines optimizations for nominal, conjunctive query answering, and incremental reasoning. Pellet OWL DL reasoner is a core component of ontology-based data management applications.

**Netica-J:** The Java version of Netica-J is used to estimate a user’s preference in a Bayesian Network. It computes probabilities to be entered directly, perhaps primarily coming from an expert, and it can detect probabilities from data. It will not deal in templates directly, but it has the amenities for libraries and on-the-fly construction that such a program requires. Netica-J use functions to measure a probability, build, learn from data, modify, performance-test, transform, save and read nets, along with a powerful inference engine. It can manage “cases” and sets of cases, in addition can connect directly with most database software. Netica-J uses the fastest known
procedure for exact general probabilistic inference in a compiled Bayesian network. In this research Netica-J is used to measure user preferred activities.

**Semweb4j:** the toolkit used to build semantic web-enable applications.

### 1.3.3 Semantic Web Languages

This section describes the third generation (semantic web) of information retrieval techniques. The third generation searching technique uses various semantic languages such as XML, RDF OIL, DAML+OIL, OWL SWRL, CL and KIF. It is designed to combine the scalability of existing internet search engines with new and improved relevancy models (Antoniou & van 2004).

**XML:** The emerging web is developed using XML rather than HTML. XML (eXtensible Markup Language) is a general-purpose specification for creating custom markup languages. More willingly than waiting for standard bodies to adopt tag set developments, or for browser companies to implement each other's standards, with XML, the user is allowed to create their own set of tags like

```
<lecturer>
  <name>John</name>
  <depart>ECE</depart>
</lecturer>
```

If the user is interested in knowing “Name of the faculties in University”, they returned result is semantically unsatisfactory in XML (Maheswari & Karpagam 2010).

```
<University>
  <lecturer>John</lecturer>
  <professor>Hussain</professor>
  <faculty>Chandru</faculty>
</University>
```
Human can predict the answer as John, Hussain and Chandru. But for the machine it is very difficult to understand that lecturers and professors are also categories of faculty. This form of information makes usage of the semantic model of the particular domain, and cannot be represented in XML efficiently, hence RDF is designed.

**RDF:** RDF creates semantic data as machine accessible information. It is the standard model for data transaction on the Web. It has features that facilitate data integration even if the underlying schemas differ and it explicitly supports the evolution of schemas over time. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as two ends of the link is referred to as a “triple”. Triple is the simple data model, it allows structured and semi-structured data to be integrated, exposed, and shared across several applications. The linking structure forms a directed, labeled graph, where the edges denote the named link between two resources presented by the graph nodes.

![RDF Graph representation of nodes and properties](image)

**Figure1.2 RDF Graph representation of nodes and properties**

Figure 1.2 shown as RDF graph, where the nodes are represented as ovals denote classes and the edge represents the properties. The class and properties can be linked to some other resources. RDF and RDFS allow the illustration of some ontological knowledge. The main modeling primitives of RDF/RDFS concern the organization of vocabularies in typed hierarchies. The hierarchies are subclass and sub property relationships, domain and range
restrictions, and instances of classes. On the other hand a number of other features are missing as follows (Maheswari & Karpagam 2010).

**Local scope of properties:** RDFS range indicates the range of a property. Consider a sentence “Animal eat plants” this rule is used for all classes of animals. Thus, in RDF Schema it is impossible to apply range restrictions only to certain classes. For instance, according to the previous rule it says cows eat only plants, while other animals may eat meat and so it is an invalid statement.

**Disjointness of classes:** Sometimes there is a need to say that classes are disjoint. For example, male and female are disjoint. But in RDF Schema it is used to denote subclass relationships, for example, female is a subclass of person.

**Boolean combinations of classes:** Sometimes there is a prerequisite to build newfangled classes by combining other classes using union, intersection and complement. For example, class person denotes to be the disjoint union of the classes male and female such definitions does not allow in RDF Schema.

**Cardinality restrictions:** Restrictions on how many distinct values a property may or must be taken is represented as cardinality ratio.

**Special characteristics of properties:** Sometimes it is useful to say that a property is transitive (like “greater than”), unique (like “is mother of”), or the inverse of another property (like “eats” and “is eaten by”).

RDF does not allow specifying these kinds of features and RDF/RDF(S) is not sufficient as ontology language, since more expressive languages for semantic web is needed, hence OIL is developed.
**Ontology Interchange Language (OIL):** OIL provides modeling primitives used in frame-based and Description Logic oriented Ontologies. OIL has formal semantics, reasoning support, modeling primitives and automatic consistency checking. The main contribution of OIL is that it is able to describe structured vocabulary with well-defined semantics.

**DAML+OIL:** DARPA agent markup language (DAML) + OIL is a semantic markup language for Web resources. It is an ontology language that describes classes and properties, in addition to the set of axioms that state features of these classes and properties. It has sufficient expressiveness. The meaning is defined by standard model-theoretic semantics based on interpretations. The interpretations consist of domain of discourse and an interpretation function.

**OWL:** The OWL (Ontology Web Language) is derived from DAML+OIL for applications that need to process the content of information instead of just presenting info to people. OWL helps greater machine interpretability of Web content which is supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. OWL has three steadily expressive sublanguages: OWL Lite, OWL DL, and OWL Full. It encourages complement, disjoint, functional properties, furthermore it also allows the overlapping of classes and properties. OWL is also used for the specification of ontology.

**Semantic Web Rule Language (SWRL):** SWRL is a proposed language for the semantic web that can be used to express rules as well as logic, mingling associating OWL DL or OWL Lite using a subset of the Rule Markup Language (RML).

**Common logic (CL)** is a context for a family of logic languages that can be accurately exchanged and transmission of knowledge in computer-based systems.
**Knowledge Interchange Format (KIF):** KIF is a language for the interchange of knowledge among disparate programs. It has declarative meanings. That means the semantics of expressions in the representation can be understood without appeal to an interpreter for manipulating those expressions and it is logically comprehensive, that means it provides for the expression of arbitrary sentences in the first-order predicate calculus. KIF supports the definition of objects, functions, relations and non-monotonic reasoning rules.

1.3.4 **Semantic Web Applications**

Currently, many researches are going on the semantic web. They are artificial intelligence, Natural Language Processing, soft computing, database, human-computer interaction, software engineering, E-Learning and information retrieval.

In this thesis, semantic web techniques have been applied to enhance search performance in information retrieval systems. This research work focuses on the design of effective methodologies for the search of relevant information and also to enhance the search efficiency of semantic web based search engines.

1.4 **INFORMATION RETRIEVAL**

The process of retrieving information from the web is called information retrieval (IR). The Information retrieval is the activity of obtaining pertinent information to an information need from a collection of information resources. Searches can be based on metadata or full-text or other content-based indexing. IR is used to digital libraries, information filtering, multimedia search and most visible in web search engines.
Since the 1940s, the problem of Information Retrieval (IR) has charmed increasing attention, particularly because of the intensely growing availability of documents. IR is the method of finding out relevant documents from collection of documents, based on a query raised by the user.

1.4.1 Information Retrieval Models

The classical information retrieval models are shown in Figure 1.3. It can be mainly divided into set theoretic models, algebraic models and probabilistic models. Set theoretic models have four main types which is Boolean, case-based reasoning (CBR), fuzzy set and extended Boolean models. Algebraic models consist of vector space model, generalized vector space model, latent semantic indexing and neural network model. Probabilistic model covers probabilistic, inference network, and brief network model.

![Information Retrieval Models Diagram](image)

Figure 1.3 Information retrieval models
In the set theoretical model, documents and queries are represented as a set of index terms. Similarly, in algebraic model documents and queries are represented as vectors in a t-dimensional space. In probabilistic model documents and queries representations are based on probability theory.

Following the definition in Baeza Yates & Ribeiro Neto (1999) an IR model is a quadruple \([D, Q, F, \text{sim}]\), where:

- \(D\) is the set of (logical representations of) documents.
- \(Q\) is the set of (logical representations of) queries.
- \(F\) is a framework for modeling documents, queries, and their relationships.
- \(\text{sim}: Q \times D \rightarrow R\) is a ranking function that defines an association between queries and documents, where \(R\) is a totally ordered set (commonly \([0,1]\), or \(P\), or a subset thereof). This ranking and the total order in \(R\) define an order in the set of documents with regard to the query.

To build a model, first it is necessary to think of how to represent documents and user information needs. Given the representations, the framework in which they can be modeled is then conceived. This framework should provide the intuition for constructing a ranking function. For the classic Boolean model, the framework is composed of sets of documents and the standard operations on sets. The framework classic vector-space model is composed of a \(t\)-dimensional vector space and linear algebra operations on vectors. Similarly the framework in classic probabilistic model is constructed of sets, standard probability tasks, as well as the Bayes' theorem.

The information in the Web mostly contains large unstructured text. The classical information retrieval techniques have become inadequate for the increasingly vast amount of text data. Finding relevant information to a user
query is a difficult task. It is difficult to formulate effective queries for analyzing and extracting useful information from the documents. Semantic web based search is a new and promising way of improving search performances on the Web. This research mainly deals with the semantic web based search technique functioned in an information retrieval system.

1.4.2 Semantic Search

Semantic search is a searching technique in which a search query aims to find keywords with the intent and appropriate meaning of the searching words. Semantic search provides more meaningful results by evaluating and understanding the search expression and finding the most relevant results in a website, database or any other data repository. The semantic search tries to solve problems like interoperability, improvement of searching methods, reliability in documents, among others, by making formally explicit the semantics of the data.

1.4.3 Challenges in Semantic Search

Semantic search has been a rapidly developing technique to improve search outcomes. The semantics of the input expression have revealed to produce better results than the traditional methods. Current search engines sometimes provide irrelevant results for user requirements. Today researches in the field of semantic search are in the opening stage; the present web search problem has a limited recall and precision values due to synonyms, homonyms, multiple languages, spelling mistakes, heterogeneity, duplications and user perceptions. In addition to the existing web search method provides no means to specify the relation between a resource and a term. For example, sell/buy. For this reason the development of reliable, precious and fast methods for semantic search is still an open challenge.
1.5 MOTIVATION FOR PRESENT RESEARCH

Search requester / users in the IR environments are heterogeneous. The requester/user could be a human, organization, student, agent or anyone. If search requesters are human or organizations, from the perspective of search requests, they can be divided into two different user groups in web search field. First, the search requesters without domain knowledge regarding their search query, and second the search requesters with domain knowledge with regard their search query. In this instance, to quickly and precisely find the search information this can best fit their user requirements. Retrieving exact information is an important issue in an information retrieval system. More than hundred publications in the area of semantic web based search have been made during the period 2001-2013. Inspired by these initiatives of the research community across the world, the author chose to carry out research in the area of semantic search.

1.6 OBJECTIVES OF THE THESIS

The research work aims to enhance semantic search performance in an information retrieval system. The goal is

- To develop generic semantic search technique for requester without domain knowledge and improve search performances.
- To develop specific semantic search techniques for requester with domain knowledge to find information precisely and quickly. The search technique decreases searching time.
- To develop QoS based ranking methodology to enhance semantic search performance of both generic and specific searches.
1.7 OVERVIEW OF PROBLEM STATEMENT

Information retrieval that handles the bindings of semantic search is facing problem due to low precision, recall and long time. Recent researches have been focused on this issue to improve web search performance.

In order to combat these issues, an efficient semantic search system with analytical proof is essential to compare the proposed work to be more effective than the existing approach.

1.8 SCOPE OF THE THESIS

This thesis presents a technique that will enable a user’s query to precisely retrieve information using semantic web.

1.9 CONTRIBUTION OF THE THESIS

The proposed system in this thesis uses semantic web based efficient search techniques. Currently the existing search engines sometimes miss highly relevant results and return some irrelevant results for user requests. For example, consider a query searching for ‘laboratories in India researching on Semantic Web’ may input the following keywords: ‘Laboratory’, ‘Semantic Web’, and ‘India’. For this query, one of the most popular search engines returns about 1,33,00,000 pages. However, on the top 10 pages, only three pages are related to the desired laboratories and others are irrelevant pages just contain some of the three keywords, such as ‘W3C Semantic Web FAQ’ and ‘Semantic Web Asia’. In the irrelevant pages, the relationship ‘research on’ between ‘Laboratory’ and ‘Semantic Web’ and the relationship ‘located on’ between ‘Laboratory’ and ‘India’ are not preserved. Furthermore, the diverse implicit meanings in the relationships among
keywords are ignored in the search. This research overcomes such limitations of traditional search by enriching the search process with ontology.

In this work, novel techniques are proposed for effective web search to enhance the semantic search performance.

The generic semantic search technique is proposed to the user who does not have domain knowledge with regard to their search queries. The generic semantic search consists of ontology knowledge base (OKB), natural language processing (NLP) techniques and Semantic Case Based Reasoning (SCBR) model.

The Specific semantic search techniques are proposed to the user who has domain knowledge with regard to their search queries and find information quickly. Specific search comprises with ontology knowledge base, Bayesian network techniques and semantic query.

Moreover, a QoS based ranking methodology is proposed to enhance search performance. Finally, a performance analysis has been made for efficient semantic search system with existing methods.
1.10 ORGANISATION OF THE THESIS

Figure 1.4 shown as organization of the thesis. This thesis is organized into six chapters.

Chapter 2, “Literature Survey” describes the brief overview about various research works in this area by reviewing different publications.

Chapter 3, “Generic Semantic Search Using Natural Language Processing Technique and SCBR Model” discusses how to resolve the issue of getting required results by the search user without domain knowledge regarding their search queries or wrong queries, misspelling, spelling variants due to synonyms and to assists to get the precious results by using spell-check method, synonymer (WordNet) and SCBR model.
Chapter 4, “Specific Semantic Search Using Bayesian Network and Semantic Query” focuses on the issue of the search user with domain knowledge regarding their search queries to quickly retrieve the required results from the ontology knowledge base by using Bayesian network and semantic query.

Chapter 5, “QoS Based Ranking Methodology for Enhancing Semantic Search Performance” discusses a solution to increase semantic search performance in generic semantic search and specific semantic search techniques.

Chapter 6, “Conclusion” concludes the thesis and suggests the scope for future research.