AN ARCHITECTURAL DESIGN OF A NOVEL SECURED SEMANTIC META SEARCH SYSTEM (SSMSS) FOR REAL TIME MULTI-DOMAIN WEB DATA AND PERFORMANCE EVALUATION OF SSMSS AGAINST BASELINE SYSTEMS

[1]SUDEEPTHI GOVATHOTI, [2]K.VENKATA RAO

[1]Research Scholar, Department of CS&SE, Andhra University, Visakhapatnam, India

[2]Professor, Department of CS&SE, Andhra University, Visakhapatnam, India

E-mail: [1]sudeepthi.chinnam@gmail.com, [2]professor_venkat@yahoo.com

ABSTRACT

In recent years, the World Wide Web allows people to share and reuse high volumes of information. Information retrieval (IR) is defined as a process to search for information from a larger set of structured documents. Search engines are the most popular Web IR tool used. Web information retrieval performs search globally from the largest collection of documents that are linked via a network and provide search services on the internet through web search engines namely: Google, Yahoo, and Bing. In this context, a critical study has been conducted on Meta Search engine, semantic web capabilities and cryptography to identify effective techniques for information retrieval. A novel Secured Semantic Meta Search System (SSMSS) is developed with the ideology of applying data security and semantic web capabilities to Meta search engine for processing real-time multi-domain web data in textual form. A critical study of some of the popular baseline search system is discussed. This paper aims at evaluating the performance of secured semantic Meta search system against four baseline search systems namely: Swoogle, Metacrawler, Dogpile and SenseBot. The relative performance results are analyzed and presented for proposed (SSMSS) search system against four baseline systems by considering the performance metrics namely: relevancy and irrelevancy score.

Keywords: *Meta search engine, Pruning, Cryptography, Semantic search, Page rank, Information retrieval.*

1. INTRODUCTION

The Contemporary developments in information system technologies have resulted in automating numerous applications in several business areas. Information has turned into a crucial requirement and a serious resource in various applications. The information retrieval from World Wide Web is a complex task even an expert user often faces difficulties. Information retrieval (IR) is the process of finding the desired information from the large document set based on a user input query. The developers of information retrieval system provide a structural retrieval of web information across an approximate trillion webpage's in global information repository.

The concept of search engine has evolved since three generations of its inception around 1996. A search engine is a web software program that is used to retrieve required information in a predominant way. It matches the user needs that are expressed in form of search query in the interface provided. It provides effectual access to data stored across webpage's based on page ranking mechanism through page relevance criteria. A search engine results page (SERP) displays the ranked list of results retrieved by the search engine. Google, Yahoo, and Bing are the most popular web search engines used by people all over the world. 'Google' is the largest spider-based search engine, recognized worldwide, set up by Larry Page and Sergey Brin from Stanford University. Its Mission is to provide universal access and organize the world's information. 'Yahoo' is the oldest web crawler-based search founded by 'Jerry Yang' and 'David Filo' [1].

Its mission is to deliver personally relevant, meaningful data with the emphasis on user's life experience and emotions. It employs its own crawler-based index, Yahoo’s index, and ranking mechanism. 'Bing' is a search engine owned by Microsoft, formerly known as MSN Search or Windows Live Search. It emphasizes a lot of efforts on visual presentation with vibrant pictures. The drawbacks of the above-mentioned search engines are it covers low web space and poor in deep web crawling.

A new concept of Meta-search engines also termed as aggregator is the most popular search tool that performs search for required information from a large set of voluminous information globally. Google, Yahoo, and Bing, the top Web Search engines are integrated for building a Meta search engine [2]. In early 1990’s Meta search system are constructed by companies or organization to compete with the traditional search systems. It utilizes the results of traditional search system and provides ranked list of results based on its own ranking mechanism. The advantage of Meta search engines are it provides broader overview of search topic. A few drawbacks in current Meta search engines are information overload, lack of getting of correct information and content that is not machine comprehensible.

The semantic web capabilities also enhance the search process and express the information in a Machine-readable format to acquire more relevant results [3]. The innovative generation of the web, the so-called Semantic Web, seems as an optimistic technology for promoting conceptual domain knowledge. Semantic web aims to change web development in such a manner that machine can add intelligence property of information showed on the web pages. Semantic Web is an extension of the traditional web that activates the information in a format that the computer understands. It is designed with the focus of machines first, humans later.

Nowadays, as web stores and transfer valuable sensitive information across various applications there is huge demand for securing digital information efficiently. Cryptography is a popular mathematical and computer science approach aims at securing digital information against various internal and external attacks. It involves two steps namely: encryption and decryption. Encryption is a process used to protect sensitive data using the combination of Public and Private Keys. It converts ordinary readable plaintext into a ciphertext format. Decryption is another process used to convert ciphertext into plain text.

In this paper a critical study has been conducted on search system. The proposed framework suggested aims at constructing an efficient secured semantic Meta search system for real time multi-domain web data. The proposed secured semantic Meta search system (SSMS) implemented by the authors in their previous research work is processed using various algorithms namely: Horizontal partitioning relevancy criteria algorithm (HPRC), Relevancy integration factor algorithm (RIF), Improved inverted indexing algorithm (I^3) [4][5] and finally secured using the Semantic trust crypto algorithm (STC). A three layered architecture has been proposed and implemented by the authors in their previous work with the intention to help the naive users to provide convenience to user.

The remaining paper is arranged as follows: In Section: 2 Motivation for implementing SSMS system, related work is presented in Section: 3. the layered wise architectural description and objectives for developing proposed SSMSS is discussed in Section: 4. A critical study and comparative analysis of baseline and search systems are illustrated in section 5. The relative performance evaluation is presented in Section 6. The prior work is discussed in Section 7 and the paper is concluded in Section: 8.

2. MOTIVATION OF A NOVEL SECURED SEMANTIC METASEARCH SYSTEM

The information technology and World Wide Web are advancing rapidly, sophisticated methods are always required to extract needed information. According to the latest statistics, the web contains 4.65 billion pages and 54% of world population is using the internet [6]. The most popular search engines namely: Google, Yahoo, Bing and MSN perform Crawler-Based Search retrieve needed information.

The motivation towards development of an automated, efficient, secured and fast responsive novel secured semantic Meta search system (SSMSS) for real time multi-domain web data is to provide a summarized report of search topic. It is developed with the ideology of applying semantic search capabilities to Meta search engine framework for processing real-time web data. The architecture suggested has three layers namely: Meta extraction, Semantic representation, and Security layer.

The search engines are broadly classified namely: traditional search engines, Meta search engines and semantic search engines came into existence. The all above mention search engines are assessed based on various performance metrics as mention in the related work presented in section 2.

The semantic representation layer developed by the authors applies semantic search capabilities to Meta search engine framework. Ontologies is one of the semantic capability that plays a crucial role in development of semantic search engines. A review of various ontology development tools protégé, DataMaster and OntoBase is discussed in the section 2.

The security layer is implemented through a cryptography technique by conducting a critical study on various cryptography security approaches namely: Blowfish, advanced encryption standard and triple data encryption standard. The aim of construction of SSMSS is to help the naive users to reduce the overhead of semantically searching multiple repositories and increasing web information space coverage and convenience to user.

3. RELATED WORK

Dallas Knight et.al [7] assessed nine search engines in five categories namely: relevance, popularity, usability, website quality, and search engine features. Kamlesh Kumar Pandey et.al [8] reviewed the performance of index-based search engines namely: Google and Yahoo. Google and Yahoo's accuracy and relative recall are also estimated. Bernard J. Jansen et.al [9] investigated Dogpile.com which is a major Meta search engine. Sareh Aghaei et.al [10] illustrated the four Web generations and their evolution: Web 1.0, Web 2.0, Web 3.0 and Web 4.0. IDipmala T. Salunke et.al [11] reviewed different partitioning techniques to manage big data and a collaborative approach is used for handling critical databases. Muhammad Shoaib et.al [12] presented a framework to enable extraction of knowledge from multiple domains related repositories and an integrated agriculture information framework (IAIF) is proposed.

Chen Hongye et.al [13] proposed a semantic web framework and function design agricultural comprehension service system to improve recall rate and accuracy rate in information retrieval. Kgotatso Desmond Mogotlane et.al [14] reviewed two Protege plug-ins namely, DataMaster and OntoBase to build Ontologies from a relational database. They are analyzed further to match their structures using database-to-ontology mapping principles. Sheetal Shimpikar et.al [15] compared and summarized various text summarization techniques used for Indian regional languages. Nikita Munot et.al [16] discussed the taxonomy, advantages and disadvantages of extractive and abstractive text summarization methods. TingyuanNie et.al [17] reviewed and analyzed two popular encryption algorithms: DES and Blowfish that plays a crucial role in information security systems. Saikumar Manku et.al [18] designed and analyzed the Blowfish encryption algorithm for information security. Karthik.S et.al [19] compared the most common encryption algorithms: Data encryption standard, Blowfish, advanced encryption standard and triple data encryption standard.

4. ARCHITECTURAL DESIGN OF PROPOSED SSMS SYSTEM

The overall architectural design of proposed SSMS system is developed with the ideology of applying cryptographic technique and semantic search capabilities to Meta search engine framework for processing data obtained from real-time World Wide Web in multiple domains. The interface is provided with flexible instructions and menu driven operations to enhance the searching experience and provide semantically rich information as per the user needs. A layered architecture is suggested and is implemented in three layers namely: Meta extraction layer, Semantic representation layer, and Security layer as shown in figure 1:

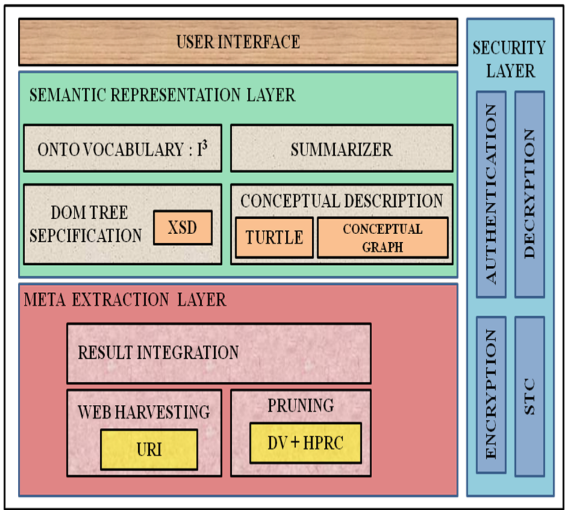

Figure 1: Architectural design of Proposed SSMS System

The Meta extraction layer accepts an input query from the user and simultaneously redirects it to multiple search engines and produces integrated results. The semantic representation layer emphasizes on applying semantic web capabilities to Meta extraction layer to promote conceptual domain knowledge. A "Text summarization" facility is provided as an added advantage to naive users by generating a summarized report of search topic. Finally, the Security layer, focus on securing the proposed system through cryptographic approach by applying "Semantic trust crypto algorithm”. A brief overview of layer wise details of ‘SSMSS’ is illustrated as follows:

Layer 1: Meta Extraction

The Meta extraction layer emphasizes developing a Meta search engine framework using “Horizontal partitioning relevancy criteria” and “Relevancy integration factor” algorithms. It retrieves the web data from multiple repositories with respect to user search query. It enhances the search capability of the user by integrating the data crawled from multiple repositories. The inclusion of steps like Pruning and Result integration plays a major role and has a great impact on the performance of the search system. The development of Meta extraction layer is carried out in three steps namely: Web harvesting, Pruning, and Result integration that are illustrated as follows.

Step 1: Web harvesting:

Web harvesting also is referred to as screen scraping or Web Crawling is defined as a process of fetching and extracting the content of the Web page using HTML parsing techniques. An input query in the form of the keyword or phrases is forwarded to selected member search engines. In this step member search engine selection and specifying category of search keyword facility is provided to the user. The retrieved results are fused/merged and stored in the fusion database with the following attributes: <Title, URL, Description, Category, Keyword> for further processing.

Step 2: Pruning

Pruning is a crucial step which has a huge impact on the performance of the proposed framework. Pruning is necessary as data retrieved from member search engines may have Null values and irrelevant data. These discrepancies may affect the performance as a result pruning is performed by applying data validation method followed by applying proposed “Horizontal partitioning relevancy criteria algorithm” (HPRC Algorithm) [4].

Step 3: Result Integration

A mathematical approach “Relevancy integration factor” algorithm is proposed to analyze and perform reordering of the pruned data obtained as per the user’s interest. The three parameters considered for calculation of relevancy integration factor are Re-rank, Weigh of page Pi and Keyword relevancy score of page Pi respectively. The average score of the above-mentioned three parameters is calculated and represents the relevancy integration factor value for a page P_i[4]. The highest Relevancy integration factor value (RIF) represents the highest precedence page retrieved for a given query.

Layer 2: Semantic Representation

The semantic representation layer emphasizes on applying semantic search capabilities to prune reordered data obtained from Meta extraction layer to promote conceptual knowledge base domain. This layer aims at converting information into a standardized machine-processable format using the following capabilities namely: DOM Tree specification, Conceptual Description, Onto Vocabulary generation. In order to provide added advantage to help naive users summarized report of semantically rich information is generated using "Text Summarizer" tool. The Semantic representation layer is carried out in four steps namely: DOM Tree Specification, Conceptual Description, Onto Vocabulary and Summarizer. The details are given below:

Step 1: DOM Tree Specification

Extensible markup language (XML) is the base for the semantic web representations. The attributes maintained in database are Title, URL, Description, Category of Keyword specified and Input Query Keyword. The attributes are tabulated in Table 1.

Table 1: Attributes of retrieved records

Attributes
Title
URL
Description
Category of Keyword
Input Search Query

In this Conversion Step the pruned reordered data is converted into Document Object Model notation using XML Schema Definition (XSD). The conversion is performed by considering each individual field into namespace convention.

An algorithmic approach for DOM tree specification conversion is explained below:

Input: *Pruned Reordered Database PRD= {T_1, T2…T_n} ∀ T_i {i=1, 2…n}*

Output: *DOM Notation of PRD= {T_1, T2…T_n} ∀ T_i {i=1, 2…n} in a semi-structured format.*

Parameters: *PRD=Pruned Reordered Database.*

T_i= Selected tuple.

Method:

Step 1: Define the standardized DOM notation.

Step 2: For each tuple in PRD, fetch the record having title, URL, and Description

Step 3: Add the namespaces for all the records traced.

Step 4: Save the representation of DOM Notation

Step 5: Repeat steps 2-4 until all the records are transformed into DOM notation.

The data converted into DOM notation is passed as an input for the next step for further processing.

Step 2: Conceptual Description

The Conceptual Description is the basic building block in semantic search, promoting conceptual modelling of web data. Conceptual Description defines Vocabularies that have a set of properties, or metadata elements. Resource discovery enables search engines to more easily discover resources on the Web. It enables users to better describe the content and content relationships available at a particular Web site. It recognizes things using Web identifiers (URIs) and depicts assets with properties and property estimations. A web resource can be identified with a URI. A Property is a Resource that has a name, for example, "creator" or "homepage" and property estimation is the estimation of a Property. The mix of a Resource, a Property, and Property estimation frames a Statement (known as the subject, predicate, and object of a Statement).

The Document Object Model notation generated in step 1 is converted into Conceptual Description syntax using two data serialization ways namely: Turtle notation from the convention specified in " http://www.w3.org/1999/02/22-rdf-syntax-ns " and graphical form representation using Graphviz environment which is an open source software that is used for generating graphs. The process of representing the data collected in the form of relational database to Turtle notation <Subject, Predicate, Object> and representing the hierarchy of data in the form of Graphical representation [5].

Step 3: Onto Vocabulary

Ontology is defined as a formal specification of conceptualization of the domain of Interest. Ontology Generation is classified into two stages namely Indexing and Vocabulary Maintenance, which is helpful for the next level of text summarization.

In this vocabulary generation step, description tags of relevant records are given as input to generate onto vocabulary by excluding the stop words and maintaining Term frequency (TF) of each phrase. The concept of forwarding indexing technique of inverted indexing mechanism and extract the words from the description tag by maintaining its Term frequency i.e., (TF). The forward inverted indexing in the traditional search engine will not exclude stop words which leads to time-consuming in updating the index of every document for searching process. The words which are extracted are compared with the set of common words that may exist which can also refer to as stop words. Ontology generation is performed by applying the proposed improved inverted indexing algorithm. (I^3). A step wise procedure of obtaining the Ontology database from the description attribute by maintaining the keywords along with their frequency is illustrated as follows:

1. Access the description field from the Pruned reordered database.
2. For every depicted content expel stop words
3. Identify the degree and frequency of word
4. Calculate the Word score and List the keywords along with their Term frequency (TF)
5. Repeat step 2& 3 until all phrases from description attribute is extracted.

Step 4: Text Summarizer

Text Summarizer step emphasizes creating a short, accurate, and fluent summary of relevant web pages in an effective way by considering the semantics of words. 'Intellexer API' is used to generate the report with respect to input query by helping the naive users to obtain a broad overview of search topic. Intellexer Summarizer aims at generating document summary by analyzing the extracted information. It is a popular cloud service application that employs NLP processing, machine learning methods and text analysis tools using JSON or XML notation.

Layer 3: Security

The third layer security focuses on securing the proposed Search system through the cryptographic approach in an efficient way that is vulnerable to attacks. The proposed "Semantic trust crypto algorithm" is employed in this step to maintain security standards. It is carried out in three steps namely: encryption, authentication, and decryption.

Step 1: Encryption

The pruned reordered database in the form of plain text is given as an input to encryption. The plain text is converted into ciphertext in encryption step by applying proposed "Semantic trust crypto algorithm".

Step 2: Authentication

The authentication process ensures that the ciphertext is decrypted into plaintext/original data only by the authorized users. A secure verification code is delivered via mailing service using SSL (Secure Socket Layer) protocol to the authenticated person can only perform decryption.

Step 3: Decryption:

The ciphertext obtained in the encryption step is given as an input to this phase. Decryption is the process of converting the ciphertext into original plaintext using the proposed "Semantic trust crypto algorithm".

4.1 Objectives of Proposed SSMS System

The objective of the proposed Semantic Meta Search System is illustrated as follows:

- ❖ To study advanced technologies used in Meta search engine, Semantic web and Cryptography.
- ❖ To propose a novel Secured framework for Semantic Meta Search System for real time multi domain web data.
- ❖ To prune the fused data by applying data validation method and proposed "Horizontal partitioning and relevancy criteria algorithm" (DV+HPRC) [4].
- ❖ To suggest a mathematical approach for reordering the pruned results using proposed "Relevancy integration factor algorithm" [4].
- ❖ To promote the standardized way of expressing information by applying semantic capabilities namely: DOM tree Specification (DOM), Conceptual Description and Ontologies [5].
- ❖ To generate a vocabulary of Ontology words by applying proposed "Improved Inverted indexing algorithm" (I^3) [5].
- ❖ To develop a secure and robust SSMSS using proposed "Semantic trust crypto algorithm" (STC).
- ❖ To conduct performance analysis of the proposed system against existing baseline system namely: Swoogle, Metacrawler, Dogpile and Sensebot.

5. Critical Study of Some Popular Baseline Search Systems

This section presents a critical study and architecture for few of the popular search systems. The details are illustrated as follows:

5.1 Swoogle

Swoogle is a crawler-based semantic search engine developed by the college of Maryland, Baltimore nation with the help of US DARPA (The Defense Advanced Research Projects Agency) and National Science Foundation offices [20]. Swoogle is one of the web search tools in the semantic web which works in view of the methods like crawler record based and recovery framework. It is developed in JAVA (JDK 1.4.2) and utilizing Apache tomcat server. The front end was designed using PHP Hypertext Pre-processor which is a server-side scripting language and utilizes a MySQL database (4.1.16). The Documents are written in RDF/XML are a segment of the semantic web reports which are recorded by swoogle.

Swoogle offers different administrations like program interface and REST Full networks administrations. The program interface provides administration to the clients and through REST full convention it gives administrations to programming specialists. It is an executed framework that finds, examines and records information encoded in semantic web archives on the Web. The data produced by the traditional search engine is metadata, which gives abstract data about the retrieved results. Swoogle provides various services like browser interface and REST Full webs services to the users, whereas through REST full protocol it provides services to software agents.

It is a web searcher for Semantic Web Ontologies, chronicle terms and data appropriated on the Web. It reasons about these chronicles and their constituent parts (e.g., terms and triples) and records huge metadata about them. It provides web-scale semantic web information to assist human clients and programming frameworks in finding important records and triples by means of its inquiry and route administrations. It gives an adjustable calculation motivated by Google's Page Rank calculation. It is a search engine for Semantic Web ontologies, documents, terms and data published on the Web. It is an implemented system that discovers, analyzes and indexes knowledge encoded in semantic web documents on the Web. The swoogle home page is shown in figure 2.

Searching over 10,000 ontologies

Swoogle © 2004-2007, ebiquity group at UMBC
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License

Figure 2: Swoogle home page

The architecture of swoogle was divided into four major components they are SWD discovery, metadata creation, data analysis an interface. The first component semantic web document (SWD) discovery is used to find out the matched SWDs to our input query throughout the website. The usage of the second component, i.e., metadata creation is used to generate the abstract metadata for discovered SWDs in both semantic and syntactic level. The third component is the information investigation finds information based on page rank calculations and indexing strategies. The last component is an interface that provides information to a Semantic Web people group. It uses a customized algorithm for ranking the query results. The architecture of swoogle is shown in figure 3:

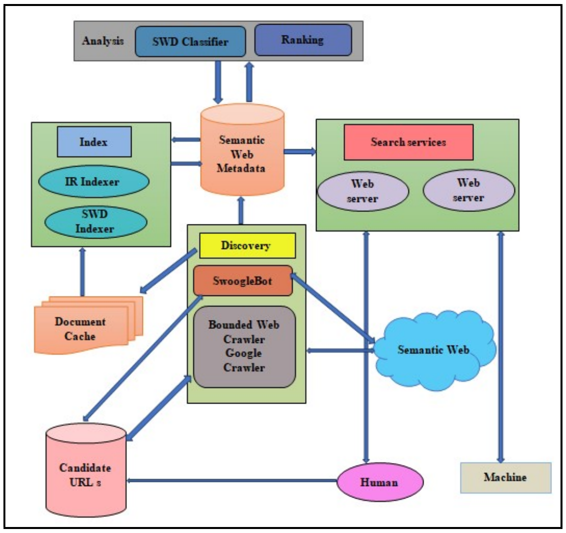

Figure 3: Swoogle architecture

5.2 Metacrawler

Metacrawler also called as Internet Spyder is developed by Erik Selberg under the direction of teacher Oren Etzioni. It is developed with the aim to provide stable administrations and deliver answers for different inquiries produced by the Search Engine. It is one of the popular Meta search engines, provides the results obtained from various search engines like Google, Yahoo, Bing, Ask.com, About.com etc. Metacrawler search engine helps the user to search multiple types of data like images, videos, news, and business information. The concept of Meta search engine became popular during the year 1990's. It is initially supported by various search engines to retrieve the report, but now it is turned as one of the top search engines like Yahoo, Google. The Metacrawler search engine home page is shown in figure 4.

metacrawler
Web
News
Images
Videos
© 2018 InfoSpace Holdings LLC
About • Privacy Policy • Terms of Use • Contact Us

Figure 4: Metacrawler home page

The techniques used in Metacrawler are a) Query handling b) Duplicate elimination of URL's and c) Truth Extractor. The specific goal in the development of Metacrawler is to identify the accuracy of information by utilizing the concept of Truth Extraction calculation. The performance of the Metacrawler search engine is evaluated based on two parameters namely: relevance ratio and retrieval effectiveness. Relevance ratio gives the performance of the search engine based on the web results Retrieval effectiveness is defined as the ability of the system to retrieve relevant documents when a query was issued.

Metacrawler architecture has four components namely: User Interface, Aggregation Engine, Parallel Web Interface, and Harness. The User Interface is the layer that interprets client inquiries and choices into the proper parameters. The user inquiries are sent to the aggregation engine, which is in charge of acquiring the underlying references from each administration, post-preparing the references, wiping out copies, and ordering and yielding the outcomes back to the user interface legitimately. The Parallel Web Interface downloads the HTML pages from the Web and acquires results. The harness has control modules that are compiled and represent a specific service for each module. It is the core place which has service-specific information. It receives control information also sends some status information back to the Aggregation Engine that is used to measure progress. The engine is designed in modular fashion so that modules can be added, modified, and removed without impacting the performance of search system. The architecture of Metacrawler provides several advantages against traditional search services, like flexibility, versatility, and convenience as the Web develops and changes is shown in figure 5.

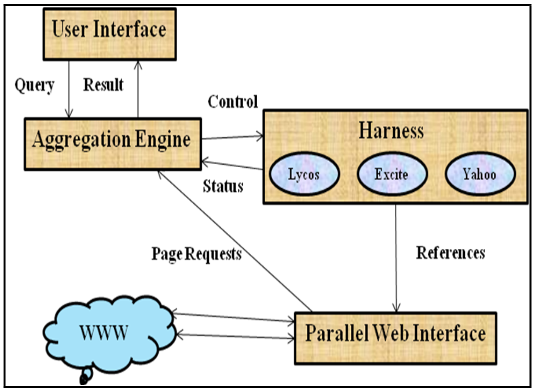

Figure 5: Metacrawler architecture

5.3 Dogpile

The InfoSpace technology developed a Dogpile search engine to retrieve information for an appropriate query. It gives high performance than other search engines like Google, Yahoo. Generally, every search engine has a unique procedure to obtain the results. Dogpile search engine initiated its operations from November 1995, and later it was handover to the Go2net which in turn taken by the InfoSpace. Result Merger is the one of an essential methodology in the working system of the Dogpile. It combines the results from various search engines and displays the top ten or more results without any post-processing process.

The extreme complexity is involved in merging algorithm, re-ranking the results which are obtained from the various search engines.

The architecture of Dogpile consists of an integration of various search engines with the Dogpile web crawler. Result merger process is used by this search engine. . If the user searches for a file and posts some query, the query request is forwarded to various search engines such as Google, Yahoo, Bing and Ask.com. When a searcher submits a query, Dogpile.com simultaneously submits the query to multiple other Web search engines then collects the results from each Web search engine emphasize on elimination of redundant results and cumulates the results and proprietary algorithm is used to combine ranked list of results. The homepage of the Dogpile search engine is shown in figure 6.

Figure 6: Dogpile home page

It combines the search results obtained from multiple sources and the topmost results are displayed as per the significance. The user retrieves the information from the Dogpile with more relevancy than Google, Yahoo, etc as it combines the results of various search engines. The architecture of Dogpile Meta search engine is shown in figure 7.

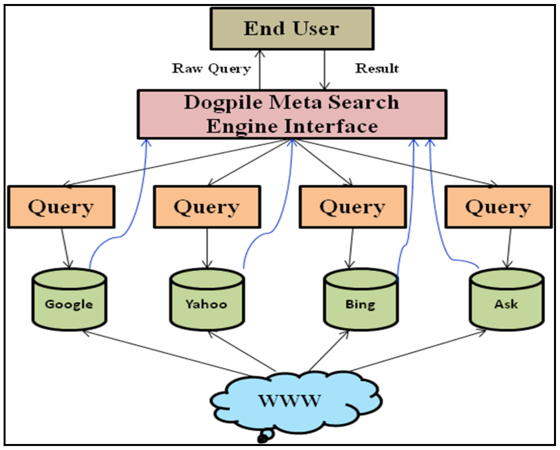

Figure 7: Dogpile architecture

5.4 SenseBot

Sensebot is a new semantic search engine, which presents the text summary of an input query when compared to other search engines. When the user submits an input query to a SenseBot search engine it produces top results from those results it presents abstract information which is relevant. The concept of text mining and semantic property is used to inspect the web pages. Multi-document summarization technique is used to present a summarized report of the top 10 results. Natural Language Processor and text mining are employed to produce efficient results. An API is used to perform the search process and analyze the webpage in order to produce the result. The home page of SenseBot is shown in figure 8.

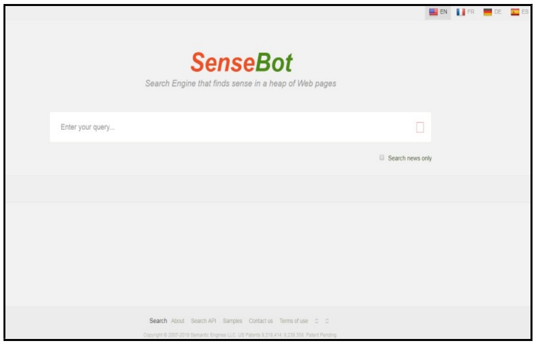
EN
FR
DE
ES
SenseBot
Search Engine that finds sense in a heap of Web pages
Enter your query...
Search news only
Search
About
Search API
Samples
Contact us
Terms of use
Copyright © 2007-2018 Semantic Engines LLC. US Patents 9,218,414; 8,239,358. Patent Pending.

Figure 8: SenseBot home page

The architecture shows the working model of the Sensebot which is used by the users to receive the information. An input search query is submitted by the user to the Sensebot user interface. It forwards the query to multiple search engines and retrieves the results. Sensebot produces web link as the summarized report to the end user. The architecture of SenseBot is shown in figure 9.

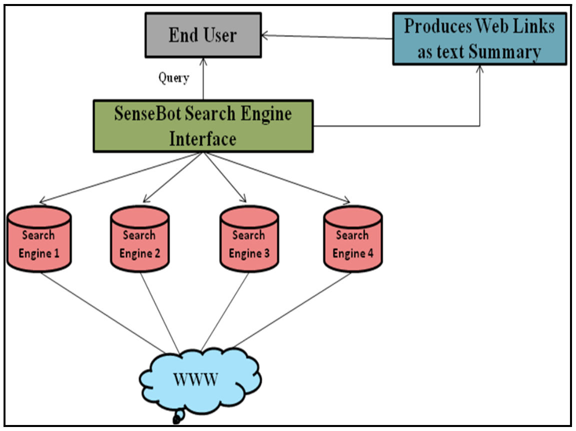

Figure 9: Sensebot architecture

5.5 COMPARATIVES ANALYSIS OF SSMSS AGAINST FOUR BASELINE SYSTEMS

This section presents a comparative analysis of Proposed SSMSS against four Baseline systems based on some of the features like algorithm used, evolution, key components and results are presented. The proposed system effectively overcomes the conventional baseline systems in its performance that can be discussed in detail in section 6.0. The comparison is illustrated in detail in table 2 as follows:

Table 2: Comparative Analysis of Baseline against Proposed Search Systems

Analysis Factor	BS1: Swoogle	BS2: Metacrawler	BS3: Dogpile	BS4: Sensebot	Proposed: SSMS
Evolution	It is developed as a research project at the University of Maryland by Li Ding advised by Professor Tim Finin	It was developed by Erik Selberg under the guidance of professor Oren Etzioni	Info Space technology developed a Dogpile search engine.	Semantic Engines LLC situated in New York developed sensebot	It is Proposed by authors
Key technology	Swangling Technique	Query processing, Duplicate elimination URL and Truth Extractor	Result Merger	Text Mining, multi-document summarization	Pruning, Text Summarization, Result integration, Security
Algorithm	Rational surfer model	Truth Extraction algorithm	Merging algorithm	Multi-document summarization	HPRC, RIF,I3,STC
Results	Semantic Web Document	Entity	Entity	Text Summary in the form of Web Links	Summarized Report

6 PERFORMANCE EVALUATION

An input search query in the form of the keyword is given by the user to search engine interface and the following parameters are calculated like the count of results retrieved, relevant and irrelevant are obtained. The efficiency of Web Search system is measured by considering two metrics namely: Relevancy Score and Irrelevancy Score with respect to the results obtained to user input search query group. The definition of relevance and irrelevancy score are defined as follows:

Relevancy Score: Relevancy Score is defined as the ratio of the total count of relevant results obtained to the total count of the retrieved results.

RS_i = Relevancy score

RET_i = Total count of results retrieved

REL_i = Total count of relevant results

$$\boxed{RS_i = \frac{REL_i}{RET_i}} \text{ ----------- (1)}$$

Irrelevancy Score: Irrelevancy Score is defined as the ratio of the total count of irrelevant results obtained to the total count of the retrieved results.

IRS_i = Irrelevancy score

RET_i = Total count of results retrieved

$IREL_i$ = Total count of irrelevant results

$$\boxed{IRS_i = \frac{IREL_i}{RET_i}} \text{ ----------- (2)}$$

The four baseline systems that are considered for evaluating the performance are Dogpile, Metacrawler, SenseBot and Swoogle. Among the four baseline systems Dogpile and Metacrawler are Meta search engines whereas swoogle and SenseBot are semantic search engines. An input search query is categorized into three query groups namely: a) Simple Single word Query, b) Simple Multi word Query and c) Complex Multi Word Query. The search systems are tested and the following parameters are recorded namely: count of retrieved, relevant, and irrelevant results.

The notations for the below table 3 are as follows:

Baseline System 1 (BS1): Swoogle
Baseline System 2 (BS2): Metacrawler
Baseline System 3 (BS3): Dogpile
Baseline System 4 (BS4): Sensebot
Proposed System (P): Proposed (SSMSS)

Table 3: Result/analysis of SSMSS against Baseline Search System

Input Query Keyword	Total Count of Results Retrieved					Count of Relevant Results					Count of Irrelevant Results					Relevancy Score					Irrelevancy Score				
	BS1	BS2	BS3	BS4	P	BS1	BS2	BS3	BS4	P	BS1	BS2	BS3	BS4	P	BS1	BS2	BS3	BS4	P	BS1	BS2	BS3	BS4	P
Simple Single Word Query																									
Pineapple	66	72	73	72	73	20	23	25	27	29	46	49	48	45	44	0.30	0.32	0.34	0.38	0.40	0.70	0.68	0.66	0.63	0.60
Rice	81	75	82	77	80	21	22	23	24	28	60	53	59	53	52	0.26	0.29	0.28	0.31	0.35	0.74	0.71	0.72	0.69	0.65
Sridevi	73	73	73	74	75	19	21	23	25	27	54	52	50	49	48	0.26	0.29	0.32	0.34	0.36	0.74	0.71	0.68	0.66	0.64
Aadhaar	70	69	70	71	72	22	23	26	28	30	48	46	44	43	42	0.31	0.33	0.37	0.39	0.42	0.69	0.67	0.63	0.61	0.58
Ubuntu	74	72	72	73	69	20	22	25	27	30	54	50	47	46	39	0.27	0.31	0.35	0.37	0.43	0.73	0.69	0.65	0.63	0.57
Ilayaraja	66	65	69	69	69	20	21	24	26	29	46	44	45	43	40	0.30	0.32	0.35	0.38	0.42	0.70	0.68	0.65	0.62	0.58
Onida	73	70	76	68	73	23	25	28	27	29	50	45	48	41	44	0.32	0.36	0.37	0.40	0.40	0.68	0.64	0.63	0.60	0.60
Simple Multi Word Query																									
Golkonda Fort	77	78	78	75	78	20	23	26	24	28	57	55	52	51	50	0.26	0.29	0.33	0.32	0.36	0.74	0.71	0.67	0.68	0.64
Digital Agriculture	68	68	73	68	68	23	25	25	26	27	45	43	48	42	41	0.34	0.37	0.34	0.38	0.40	0.66	0.63	0.66	0.62	0.60
Organic Farming	76	81	80	77	78	21	28	30	28	31	55	53	50	49	47	0.28	0.35	0.38	0.36	0.40	0.72	0.65	0.63	0.64	0.60
Big Boss 2	65	67	69	70	70	20	24	27	29	30	45	43	42	41	40	0.31	0.36	0.39	0.41	0.43	0.69	0.64	0.61	0.59	0.57
China Trip	67	68	70	70	71	19	22	23	25	27	48	46	47	45	44	0.28	0.32	0.33	0.36	0.38	0.72	0.68	0.67	0.64	0.62
Thailand Tourism	67	65	71	72	70	20	23	25	24	26	47	42	46	48	44	0.30	0.35	0.35	0.33	0.37	0.70	0.65	0.65	0.67	0.63
Microsoft phones	64	64	69	64	68	19	21	24	23	26	45	43	45	41	42	0.30	0.33	0.35	0.36	0.38	0.70	0.67	0.65	0.64	0.62
Complex Multi Word Query																									
Windows And Linux Environment	67	68	73	68	73	20	24	27	25	30	47	44	46	43	43	0.30	0.35	0.37	0.37	0.41	0.70	0.65	0.63	0.63	0.59
Evolution of Web	60	64	73	69	71	20	22	25	23	26	40	42	48	46	45	0.33	0.34	0.34	0.33	0.37	0.67	0.66	0.66	0.67	0.63
QCS Engineer Job Vacancies	65	71	75	72	74	21	23	28	24	29	44	48	47	48	45	0.32	0.32	0.37	0.33	0.39	0.68	0.68	0.63	0.67	0.61
Internet of	65	69	72	70	70	22	24	26	25	29	43	45	46	45	41	0.34	0.35	0.36	0.36	0.41	0.66	0.65	0.64	0.64	0.59

The total count of results retrieved, relevant and irrelevant for the above three mentioned query groups are aggregated and presented in table 4 as follows:

Table 4: Relative Performance analysis of the SSMSS against baseline systems

SEARCH SYSTEM	INPUT QUERY GROUP								
	SIMPLE SINGLE WORD QUERY			SIMPLE MULTI WORD QUERY			COMPLEX MULTI WORD QUERY		
	RETRIEVED	IRRELEVANT	RELEVANT	RETRIEVED	IRRELEVANT	RELEVANT	RETRIEVED	IRRELEVANT	RELEVANT
SWOOGLE	503	358	145	484	342	142	442	296	146
METACRAWLER	496	339	157	491	325	166	464	303	161
DOGPILE	515	341	174	510	330	180	498	315	183
SENSEBOT	504	320	184	496	317	179	472	304	168
SSMSS	511	309	202	503	308	195	493	295	198

Graphical representation of relative Performance analysis of the SSMSS against the baseline system is shown in fig: 10.

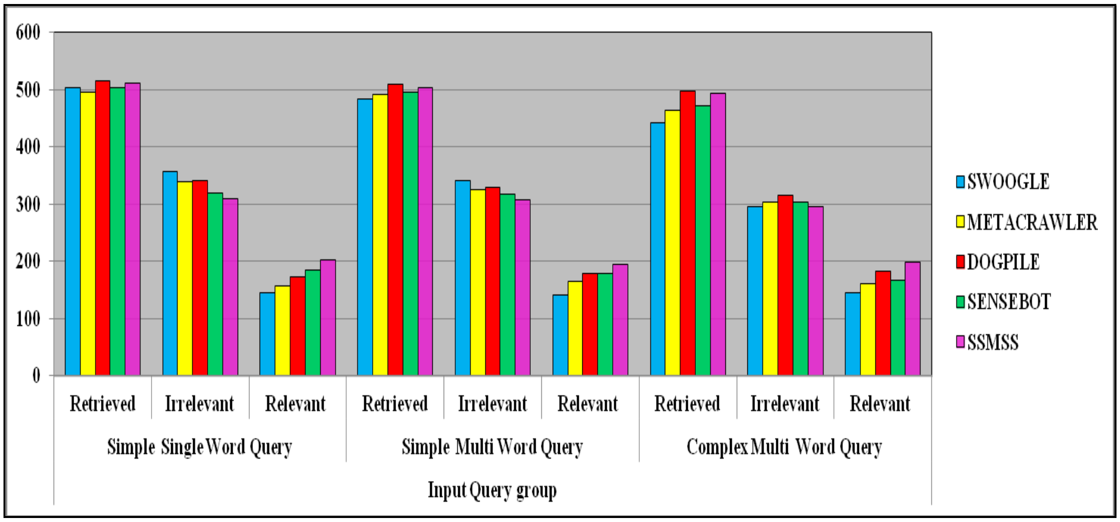

Figure 10: Graphical representation of relative Performance of proposed and baseline systems

The overall relevancy score and irrelevancy score is calculated by applying the romulas as specified in 1& 2 for the proposed SSMS system and four baseline systems as shown in table 5:

Table 5: Overall relevancy and irrelevancy score

Search System	Irrelevancy score	Relevancy score
Swoogle	0.69	0.31
Meta Crawler	0.67	0.33
Dogpile	0.65	0.35
Sensebot	0.64	0.36
SSMSS	0.6	0.4

Graphical representation of overall relevance and irrelevance score analysis for the SSMSS against the baseline system is shown in fig 11:

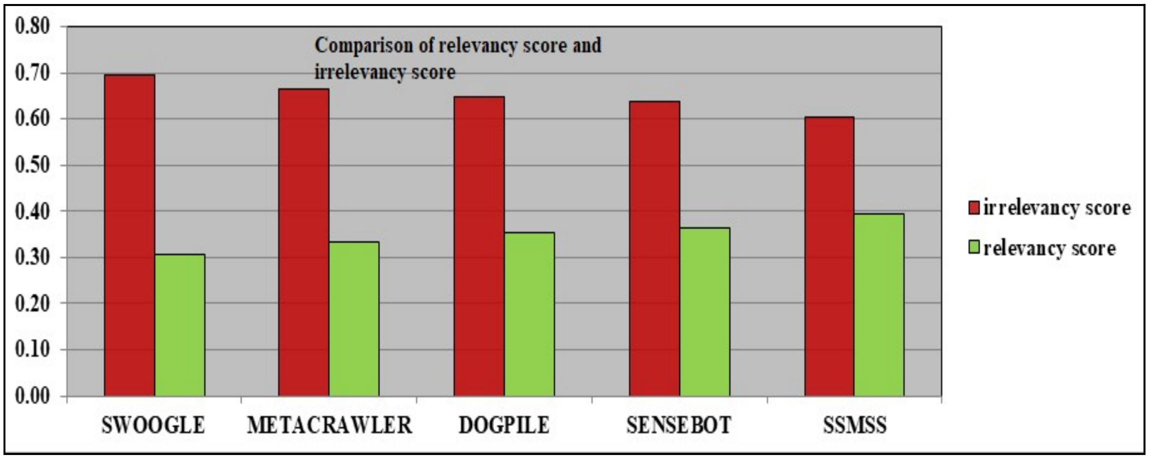

Figure 11: Graphical representation of Overall Relevancy and Irrelevancy score

In this study, relevancy score and irrelevancy score of four Baseline System against proposed SSMSS are observed. The result of this study concludes that the overall relevancy score obtained for the proposed system is 0.39 is high when compared to other baseline systems. Similarly, the irrelevancy score obtained for the proposed SSMSS is 0.60 which is less when compared to other search systems.

7 PRIOR RESEARCH WORK

A brief survey and broad overview of various Meta search engines and the key technologies involved in the development of Meta search engine is presented by the authors [21]. A review of some of the best semantic search engines namely: Hakia, SenseBot, swoogle e.t.c that yields unique search experience to user is discussed by the authors [22]. The proposed secured semantic Meta search system (SSMSS) is developed in three layers namely: Meta extraction layer, semantic representation layer and security layer. The Meta extraction layer emphasize in developing a Meta search engine framework using proposed "Horizontal partitioning relevancy criteria" and "Relevancy integration factor" algorithms for efficient data retrieval from multiple repositories is presented by authors [4].

The semantic representation layer applies semantic web capabilities namely: XML, conceptual description and ontologies to Meta search engine framework for accessing semantic rich information [5]. A secured semantic Meta search stack using proposed "Semantic Trust Crypto" algorithm is implemented in security layer. A comprehensive investigation is conducted for some of the popular cryptography approaches [23]. The prior work conducted so far provided as basis for performing relative performance evaluation of proposed secured semantic Meta search system (SSMSS) against four baseline systems namely: Swoogle, Metacrawler, Dogpile and SenseBot that is presented in this paper.

8 CONCLUSION

A critical study of Semantic Web, Meta search engines and Cryptography techniques is performed. A new secured semantic Meta search system (SSMSS) is developed to meet the global standards of web technology. The proposed system is processed using various algorithms namely HPRC, RIF, I^3 and STC. A critical study is conducted on some of the popular search systems. The relative performance of baseline and proposed search system is evaluated by considering the metrics namely: relevancy and irrelevancy score. It is observed that proposed system efficiently retrieves semantically rich information when compared to existing baseline system. The proposed SSMS System can be extended to retrieve image and multi-media data by generating image tags. The language conversion tools can also be included to extend the services in naive user languages. For efficient handling of user queries on the high volumes of variety data, the data management has to be enhanced by simulating a Big Data SSMSS with Hadoop file systems and map-reduce programming features.

REFRENCES:

[1]. Manika Dutta and Dr. K. L. Bansal: “A Review Paper on Various Search Engines (Google, Yahoo, Altavista, Ask and Bing)”, in proceedings of International Journal on Recent and Innovation Trends in Computing and Communication, Volume 4, Issue: 8, pp.190 – 195, 2016.

[2]. Jai Manral and Mohammed Alamgir Hossain: “An Innovative Approach for online Meta Search Engine Optimization”, in proceedings of 6th Conference on Software, Knowledge, Information Management and Applications, Chengdu, China, pp.1-7, 2012.

[3]. Dr Brian Matthews: “Semantic Web Technologies”, in proceedings of article JISC Technology and Standards Watch, pp.1-21, 2005.

[4]. G.Sudeepthi and Prof. M.Surendra Prasad Babu ,” Design and Development of Meta Search Engine Framework using Horizontal Partitioning Relevancy Criteria and Result Integration Factor Algorithms for Efficient Data Retrieval”, International Journal of Engineering & Technology (IJET-IJENS), Volume:18 Issue No:03, pp:19-27, June 2018.

[5]. Sudeepthi Govathoti and M.S. Prasad Babu, “An Implementation of a New Framework for Automatic Generation of Ontology and RDF to Real Time Web and Journal Data”, International Journal of Computer Science and Information Security (IJCSIS), Volume. 16 Issue No. 1, pp.89-94, January 2018.

[6]. https://wearesocial.com/blog/2018/01/global-digital-report-2018

[7]. Dallas Knight, Alec Holta and Jim Warrenb: "Search engines: “A study of nine search engines in four categories", in proceedings of Journal of Health Informatics in Developing Countries, Volume 3, Issue No.1, pp.1-9, 2009.

[8]. Kamlesh Kumar Pandey, Rajat Kumar Yadu and Pradeep Kumar Shukla, “A Comparative Performance Analysis of Google Index Web Search Engine and Yahoo Directories Web Search Engine”, in proceedings of International journal of advanced technology in engineering and science, Volume 04, Issue:04, pp.233-243, 2016.

[9]. Bernard J. Jansen Amanda Spink and Sherry Koshman , “Web Searcher Interaction With the Dogpile.com Meta search Engine”, in proceedings of Journal of the American Society for Information Science and Technology, Volume 58, Issue 5, pp.744–755, 2007.36

[10]. Sareh Aghaei, Mohammad Ali Nematbakhsh and Hadi Khosravi Farsani: “Evolution of The World Wide Web: From Web 1.0 To Web 4.0”, in proceedings of International Journal of Web & Semantic Technology (IJWesT), Volume 3, Issue 1, pp:1-10, 2012.

[11]. IDipmala T. Salunke and Girish P. Potdar: “A Survey Paper on Database Partitioning” in proceedings of International Journal of Advanced Research in Computer Science & Technology, Volume 2, Issue 3, 2014.

[12]. Muhammad Shoaib and Amna Basharat: “Semantic Web based Integrated Agriculture Information Framework”, in proceedings of Second International Conference on Computer Research and Development IEEE, pp.285-289, 2010.

[13]. Chen Hongye: “Study of Agricultural Knowledge Service System Model Based on Semantic Web”, in proceedings of International Forum on Information Technology and Applications, Computer Society IEEE, pp.381-383, 2010.

[14].Kgotatso Desmond Mogotlane and Jean Vincent Fonou-Dombeu: “Automatic conversion of relational databases into ontologies: a comparative analysis of protege plug-ins performances”, in proceedings of International Journal of Web & Semantic Technology (IJWesT), Volume 7, Issue 3, 2016.

[15].Sheetal Shimpikar and Sharvari Govilkar: “A Survey of Text Summarization Techniques for Indian Regional Languages”, in proceedings of International Journal of Computer Applications , Volume 165, Issue 11, 2017.

[16].Nikita Munot and Sharvari S. Govilkar: “Comparative Study of Text Summarization Methods”, International Journal of Computer Applications, Volume 102, Issue 12, 2014.

[17].Tingyuan Nie: “A Study of DES and Blowfish Encryption Algorithm”, in proceedings of Communication and Electronic Engineering, IEEE, 2009.

[18].Saikumar Manku and K. Vasanth: “Blowfish Encryption Algorithm for Information Security”, in proceedings of ARPN Journal of Engineering and Applied Sciences, Volume 10, Issue 10, 2015.

[19].Karthik .S and Muruganandam: “A Data Encryption and Decryption by Using Triple DES and Performance Analysis of Crypto System”, in proceedings of International Journal of Scientific Engineering and Research (IJSER), Volume 2, Issue 11, 2014.

[20].Li Ding ,Tim Finin, Anupam Joshi, Yun Peng, R. Scott and Cost Joel Sachs: “Swoogle: A Semantic Web Search and Metadata Engine”, in proceedings of thirteenth ACM international conference on Information and knowledge management, pp.652-659, 2004

[21].G.Sudeepthi and Prof.M.Surendra Prasad Babu, “A Survey on Meta Search Engine in Semantic Web”, International Journal of Computer Technology and Applications, Volume.2 Issue.6, pp. 3051- 3055, 2011.

[22]. G.Sudeepthi, G. Anuradha and Prof. M.Surendra Prasad Babu,” A Survey on Semantic Web Search Engine”, International Journal of Computer Science Issues, Volume 9 Issue 2, pp. 241-245, 2012.

[23].G.Sudeepthi and Prof. K.Venkata Rao,” A Comprehensive Investigation of Some Popular Cryptographic Approaches - Implementation of Secured Semantic Meta Search Stack Using Proposed “Semantic Trust Crypto” Algorithm ”, Journal of Advanced Research in Dynamical & Control Systems, Volume. 10, Special Issue. 13, pp: 1011-1025, 2018.