

Web Search Engine



Bosubabu Sambana, MCA, M.Tech

Assistant Professor,

**Dept of Computer Science & Engineering,
Simhadhri Engineering College, Visakhapatnam, AP-531001, India.**

Abstract:

The World Wide Web (WWW) allows people to share information or data from the large database repositories globally. We need to search the information with specialized tools known generically as search engines. There are many search engines available today, where retrieving meaningful information is difficult. However to overcome this problem of retrieving meaningful information intelligently in common search engines, semantic web technologies are playing a major role. In this paper we present a different implementation of semantic search engine and the role of semantic relatedness to provide relevant results. The concept of Semantic Relatedness is connected with Wordnet which is a lexical database of words. We also made use of TF-IDF algorithm to calculate word frequency in each and every webpage and Keyword Extraction in order to extract only useful keywords from a huge set of words. These algorithms are used to retrieve much optimized and useful results to the user.

Keywords:

TF-IDF, Semantic Relatedness, Keyword Extraction, Word net.

1. INTRODUCTION:

The grandfather of all search engines was **Archie**, created in 1990 by Alan Emtage, a student at McGill University in Montreal. World Wide Web provides us with huge amount of necessary data digitally available as hypertext Data may be WebPages, images, information and other type.

This hypertext pool is dynamically changing due to this reason it is more difficult to find useful information. In 1995, when the number of “usefully searchable” Web pages was a few tens of millions, it was widely believed that “indexing the whole of the Web” was already impractical or would soon become so due to its exponential growth. A little more than a decade later, the GYM search engines—Google, Yahoo!, and Microsoft—are indexing almost a thousand times as much data and between them providing reliable sub second responses to around a billion queries a day in a plethora of languages. If this were not enough, the major engines now provide much higher quality answers.

For most searchers, these engines do a better job of ranking and presenting results, respond more quickly to changes in interesting content, and more effectively eliminate dead links, duplicate pages, and off-topic spam. In this two-part series, we go behind the scenes and explain how this data processing “miracle” is possible.

We focus on whole-of-Web search but note that enterprise search tools and portal search interfaces use many of the same data structures and algorithms. Search engines cannot and should not index every page on the Web. After all, thanks to dynamic Web page generators such as automatic calendars, the number of pages is infinite. To provide a useful and cost-effective service, search engines must reject as much low-value automated content as possible. In addition, they can ignore huge volumes of Web-accessible data, such as ocean temperatures and astrophysical observations, without harm to search effectiveness.

The History of Search Engines:

Modern search engines are pretty incredible – complex algorithms enable search engines to take your search query and return results that are usually quite accurate, presenting you with valuable information nuggets amidst a vast information data mine [1]. Search engines have come a long way since their early prototypes, as our Internet Search Engines History infographic illustrates. From improvements in web crawlers and categorizing and indexing the web, to introducing new protocols such as robots.txt so that webmasters have control what web pages get crawled, the development of search engines has been the culmination of multiple search technologies that developed from different search engines. Alta Vista was the first search engine to process natural language queries; Lycos started strong with a system categorizing relevance signals, matching keywords with prefixes and word proximity; and Ask Jeeves introduced the use of human editors to match actual user search queries,

Most of the traditional search engines search for keywords to answer queries from users. The main focus of these search engines is solving queries with close to precise results in small period of time using much advanced algorithms. However, it shows that such search engines are incompetent in answering queries intelligently using traditional approach. The Semantic Web will support more efficient discovery, automation, integration and reuse of data and provide support for interoperability problem which cannot be resolved with current web technologies. In short it will intelligently understand the user query and search for those results that match not only the keyword but also the meaning of that query.

In this paper, we will make modification over the existing search engine by adding an additional concept of keyword extraction and semantic relatedness calculation. Semantic relatedness here is a metric which calculates the relation between words. This metric is computed with the help of Wordnet. Another metric used in the current approach is TF-IDF (Term Frequency-Inverse Document Frequency).

It is used to calculate the relevancy of each word and relevance of each document. Finally, Web search engines have no access to restricted content, such as pages on corporate intranets. What follows is not an inside view of any particular commercial engine—whose precise details are jealously guarded secrets—but a characterization of the problems that whole - of - Web search services face and an explanation of the techniques available to solve these problems.

2. BACKGROUND:

2.1 Search Engine:

Search Engine provides the gateway for most of the users trying to explore the huge information base of web pages. Search engines are a program that search documents for specified keywords on search for information on the World WideWeb and returns a list of the documents where the keywords were found. A Search Engine is really a class of programs. However, the term is often used to specifically describe systems like Google, Bing and Yahoo! Search that enable users to search for documents on the World Wide Web [2].



FIGURE 1: EXAMPLE OF SEARCH ENGINE

2.2 Cloud computing:

Cloud computing is continuously developing as a standard for sharing the data over the remote storage in an online cloud server. Cloud services offers great amenities for the users to enjoy the on-demand cloud applications without any obligations related to data. During the data retrieving, different users may be in a cooperative relationship, and hence data distribution becomes important.

Goals of Search Engine:

- **Quality** - Means effectiveness can be defined as to retrieve the most relevant set of document for a query. Process text and store text statistics to improve relevance be used.
- **Speed** - Means efficiency may be defined as a process queries from users as fast as possible For it specialized data structure should be used.

How web Based Search Engine Works?

Web based search engine works by saving the information of many web pages, which they retrieve itself. These pages are retrieved by a web crawler which is also called spider which follows every link on the site. Search engine is a term used for information retrieval. Search engine match queries against an index that they create. This index contains the word in each document, pointers to their location within the document. This is called inverted file [4]

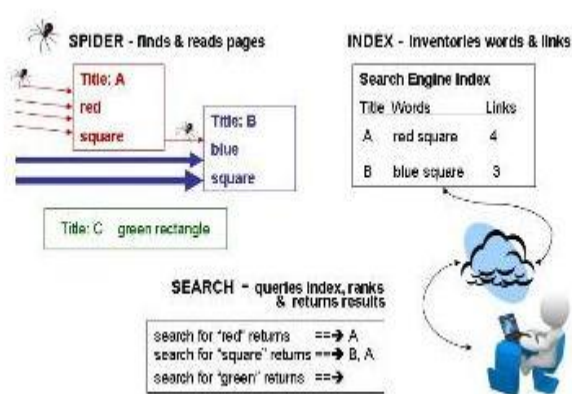


Figure.4: How do Search Engine Work

Web search engines get their information by web crawling from site to site. The "spider" checks for the standard filename robots.txt, addressed to it, before sending certain information back to be indexed depending on many factors, such as the titles, page content, JavaScript, Cascading Style Sheets (CSS), headings, as evidenced by the standard HTML markup of the informational content, or its metadata in HTML meta tags. Indexing means associating words and other definable tokens found on web pages to their domain names and HTML-based fields. The associations are made in a public database, made available for web search queries.

A query from a user can be a single word. The index helps find information relating to the query as quickly as possible. Some of the techniques for indexing, and caching are trade secrets, whereas web crawling is a straightforward process of visiting all sites on a systematic basis. Between visits by the spider, the cached version of page (some or all the content needed to render it) stored in the search engine working memory is quickly sent to an inquirer. If a visit is overdue, the search engine can just act as a web proxy instead. In this case the page may differ from the search terms indexed. The cached page holds the appearance of the version whose words were indexed, so a cached version of a page can be useful to the web site when the actual page has been lost, but this problem is also considered a mild form of linkrot.

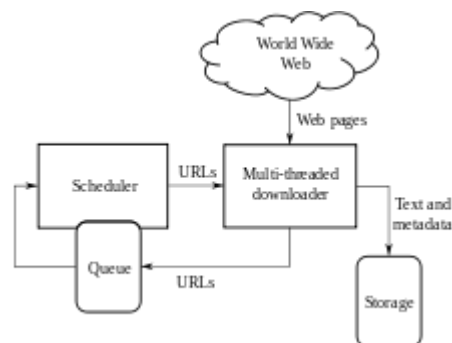


Figure.5: High-level architecture of a standard Web crawler

Typically when a user enters a query into a search engine it is a few keywords. The index already has the names of, the sites containing the keywords, and these are instantly obtained from the index. The real processing load is in generating the web pages that are the search results list: Every page in the entire list must be weighted according to information in the indexes. Then the top search result item requires the lookup, reconstruction, and markup of the snippets showing the context of the keywords matched. These are only part of the processing each search results web page requires, and further pages (next to the top) require more of this post processing. Beyond simple keyword lookups, search engines offer their own GUI- or command-driven operators and search parameters to refine the search results.

These provide the necessary controls for the user engaged in the feedback loop users create by filtering and weighting. While refining the search results, given the initial pages of the first search results. For example, from 2007 the Google.com search engine has allowed one to filter by date by clicking "Show search tools" in the leftmost column of the initial search results page, and then selecting the desired date range.^[16] It's also possible to weight by date because each page has a modification time. Most search engines support the use of the Boolean AND, OR and NOT to help end users refine the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search, which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases you search for. As well, natural language queries allow the user to type a question in the same form one would ask it to a human. A site like this would be ask.com.

The usefulness of a search engine depends on the relevance of the result set it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another.^[14] The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: one is a system of predefined and hierarchically ordered keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This first form relies much more heavily on the computer itself to do the bulk of the work.

Most Web search engines are commercial ventures supported by advertising revenue and thus some of them allow advertisers to have their listings ranked higher in search results for a fee. Search engines that do not accept money for their search results make money by running search related ads alongside the regular search engine results. The search engines make money every time someone clicks on one of these ads.

Features of Web based Search Engine:

Following are the basic features for evaluating web based search engine [5].

- ✦ **Web Indexes**-When a web search request is generated. It is the web index generated by web robots or spiders. The combination of web indexes affects the performance of a web search engine. Three main key points to design of web index are coverage, update frequency and the part of indexed web page.
- ✦ **Search Capability**- Search Engine must provide Phrase searching, truncation Search capacity finds its Performance efficiency, throughput.
- ✦ **Retrieval Issue** -This issue proceed on three Key points- Precision, Recall and response time.
- ✦ **Write Option**-Write option or output option provides the deal with actual content of output.
- ✦ **User effort** -User effort means the documentation and interface. Good prepared documentation and good interface play a different role in users' selection of web search engine. User will only user the search engine when the interface is user friendly only.

Quality of Good Search Engine:

- Ability to produce the most relevant result to any given search.
- A true search engine is an automated software program that moves around the web collecting WebPages to include in its catalog or database.
- It searches when user requests information from a search engine has its own catalog or database of collected WebPages, so you will get different results. Hits by using different search engines.

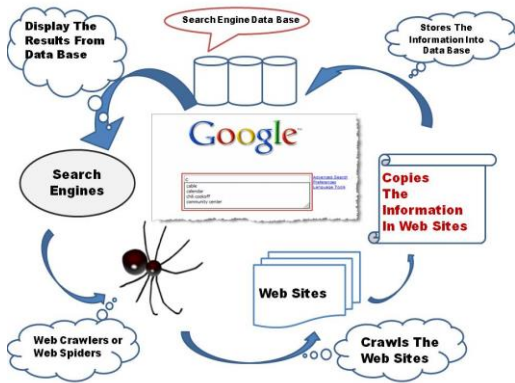


Figure.6: Index and working Process

Compression:

Indexers can reduce demands on disk space and memory by using compression algorithms for key data structures. Compressed data structures mean fewer disk accesses and can lead to faster indexing and faster query processing, despite the CPU cost of compression and decompression.

Early termination:

The query processor can save a great deal of computation if the indexer creates indexes in which it sorts postings lists in order of decreasing value. It can usually stop processing after scanning only a small fraction of the lists because later results are less likely to be valuable than those already seen. At first glance, early termination seems to be inconsistent with skipping and compression techniques, which require postings to be in document number order.

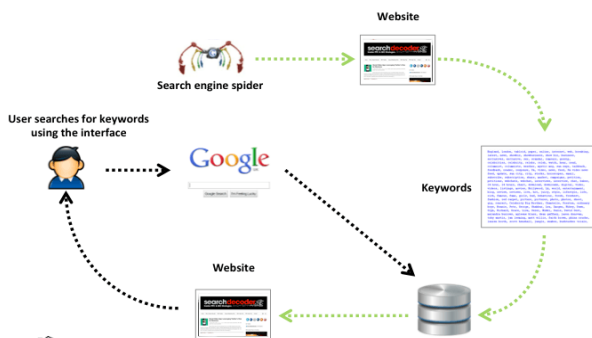






Figure.7: Information Retrieval Process in web search

Timeline (full list) [3]		
Year	Engine	Current status
1993	W3Catalog	Inactive
	Aliweb	Inactive
	JumpStation	Inactive
	WWW Worm	Inactive
1994	WebCrawler	Active, Aggregator
	Go.com	Inactive, redirects to Disney
	Lycos	Active
	Infoseek	Inactive
1995	AltaVista	Inactive, redirected to Yahoo!
	Daum	Active
	Magellan	Inactive
	Excite	Active
	SAPO	Active
	Yahoo!	Active, Launched as a directory
1996	Dogpile	Active, Aggregator
	Inktomi	Inactive, acquired by Yahoo!
	HotBot	Active (lycos.com)
	Ask Jeeves	Active (rebranded ask.com)
1997	Northern Light	Inactive
	Yandex	Active
1998	Google	Active
	Ixquick	Active also as Startpage
	MSN Search	Active as Bing
	empas	Inactive (merged with NATE)
1999	AlltheWeb	Inactive (URL redirected to Yahoo!)
	GenieKnows	Active, rebranded

		Yellowee.com
	Naver	Active
	Teoma	Inactive, redirects to Ask.com
	Vivisimo	Inactive
2000	Baidu	Active
	Exalead	Active
	Gigablast	Active
2003	Info.com	Active
	Scroogle	Inactive
2004	Yahoo! Search	Active, Launched own web search (see Yahoo! Directory, 1995)
	A9.com	Inactive
	Sogou	Active
2005	AOL Search	Active
	GoodSearch	Active
	SearchMe	Inactive
2006	Soso (search engine)	Active
	Quaero	Inactive
	Ask.com	Active
	Live Search	Active as Bing, Launched as rebranded MSN Search
	ChaCha	Active
	Guruji.com	Inactive
2007	wikiseek	Inactive
	Sproose	Inactive
	Wikia Search	Inactive
	Blackle.com	Active, Google Search
2008	Powerset	Inactive (redirects to Bing)
	Picollator	Inactive

	Viewzi	Inactive
	Boogami	Inactive
	LeapFish	Inactive
	Forestle	Inactive (redirects to Ecosia)
	DuckDuckGo	Active
2009	Bing	Active, Launched as rebranded Live Search
	Yebol	Inactive
	Mugurdy	Inactive due to a lack of funding
	Scout (Goby)	Active
	NATE	Active
2010	Blekkio	Inactive, sold to IBM
	Cuil	Inactive
	Yandex	Active, Launched global (English) search
2011	YaCy	Active, <u>P2P</u> web search engine
2012	Volunia	Inactive
2013	Halalgoogling	Active, Islamic / Halal filter Search
2013	Egerin	Active, Kurdish / Sorani Search engine

Search engine	Market share in September 2015-16	
Google	69.24%	
Bing	12.26%	
Yahoo!	9.19%	
Baidu	6.48%	

Search engine	Market share in September 2015-16	
AOL	1.11%	██████████
Ask	0.24%	██████████
Lycos	0.00%	

4. RESEARCH WORK:

Existing System:

Problems Facing by Current Search Engines:

- ✦ Crawlers are not able to analyze the content of keyword in web page before they download it.
- ✦ User submits his request for retrieval of information without mentioning the content in which he otherwise desire.
- ✦ Crawler treats user search request in isolation.
- ✦ There is a requirement to prepare separate files for each web document.
- ✦ Augmentation is required in HTML document.

Types of Search Engine:

According to functioning three types of search engine

- 1. Crawler Based Search Engine:**
They create their listings automatically. Spider builds them. Computer algorithm ranks all pages. These types of search engines are heavy and often retrieve a lot of information. For complex search it allows to search within the results of previous search and enable you to refine search results.
- 2. Human Power Directories:**
These are designed by human selection means they depend on professional to create listings. These never contain full text or webpage they link to.
- 3. Hybrid Search Engine:**
These are different from traditional text oriented search engine such as Google or directly based searched engine such as Yahoo in which each program operates by comparing a sets of metadata.

4. Content/topic
5. Web search engines
6. Selection-based search engines
7. Meta search engines
8. Semantic search engine
9. Desktop search tools
10. Web portals and vertical market websites that have a search facility for online databases
11. Deep Web Search Engines

Search Engine Optimization (SEO):

Search Engine Optimization is the procedure of improving the visibility of a website or webpage in search engine via the natural or unpaid searched results. Optimization may target different types of search like image search, local search, video search, academic search, new search, industry specific vertical search .It can also be define as the process of affecting the visibility of a website or webpage in search engine. In search engine optimization updating or modification of all variables to get a better location in the search engine takes place. We start with Search Engine Optimization and how it can be used to formulate internet marketing. Strategy as well as Technical aspects of SEO.

- A) Using SEO as a marketing strategy it can be described as a method of getting our website to rank higher in search engine as Google; Yahoo, Means that if user likes to search for a list of optimized keywords the chances are that the visitors see your site on first few places may be good.
- B) Parameters for evaluating SEO of websites- Page Rank- Page rank of each page depends on the page rank of pages pointing to it.
- C) To enhance our site page rank few key ideas are inbound links, outbound links, Dangling links, domain and File names and broken links.

Search Engine Optimization Technique:

Basically three techniques for search engine optimization are there

- **Directory Submission:** It is the important technique in Search Engine Optimization to create in coming links to a website through indexed page and category. Different directory provides free service to website. Directory submission request information regarding URL, title, keywords.
- **Keyword Generation:** All search engine optimization need some words to elaborate information based on these words. Keywords should be of your organization on subject. This process can be proceeding by different online tools like word tracker, yahoo keyword selector tool, Google Ad words.
- **Link Exchange:** To start up any website for any business we need reciprocal link exchange with other websites. It is the procedure to take place link on other website and other website place links on our site.

exchange links. In it import link, outbound links, internal link are used.

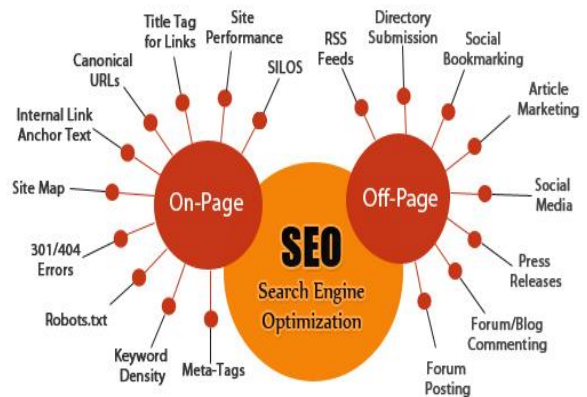


Figure.8: Overview of SEO

Tools of Search Engines Optimization:

SEO tools are the operators that optimize the search engine functionality Basics tools are-

- **Keyword Tool-** Include keyword research tools, keyword density analysis tool, and competitor analysis tool. It can used for website classification and regulate keywords deployment columns. Example – Keyword selector tool, external tool
- **Link Tool-** These tools include link popularity spider simulator, by which ranking of website can be increased.
- **Usability Tool-** This tool test pages display effects in different resolution, different operating system, and different browser. These include HTML and CSS validation, Firefox extension, and Page speed test.
- **Keywords Strategy-** When choosing keywords, it must be related with products, area, service.
- **High duality incoming Link-** Submit the website to search engine directories, find websites to

Disadvantage:

Present System does not have the option of granting/revoking data access in Proper manner for user/clients needs.

Proposed System:

In this paper, we are proposed in new method in this web Search Process

Proposed Architecture:

This architecture should be capability of systems working in a Distributed manner. In it all processing should be processed on idle computer. The distributed architecture should not increase Network traffic. All the systems connected in a Network should be operated using some firewalls. The module should be as much as easy to plug and play.

- ▣ **Crawler Unit:** - It will crawl a website. It will need to use secondary memory to store the web pages downloaded before analysis. The web pages should be saved on each host system, rather than transferred to the control unit to minimize the network traffic. In crawler unit the technique of Data mining that is cluster may be applied by which similar data elements, similar URL's may be kept as a cluster. Cluster helps us to crawl the different pages. Different types of clustering algorithm may be used to crawl the useful URL's.
- ▣ **Control Unit:** - When a crawler requests a job or sends some data elements the control unit will live on a web server and will be used by it.

It will need to save the commands that the user like to be process. This can also be understudied by this example as doc file is saved on server.

■ **Messaging System:** - To satisfy the necessity of crawler and control unit the crawler must be able to download and process websites with less transmission with control unit-The crawler unit starting a crawl, sending a message to control unit shows that it is ready to execute a new request. The control unit sending an instruction to the crawler showing a site to crawl and type of processing to be performed on downloaded website.

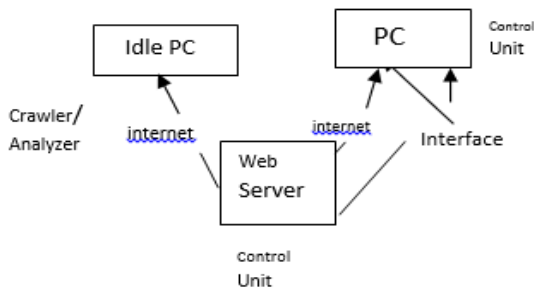


Figure – A

The problem is that control unit outside the network can not initiate communication with component inside but only can send information in response to a request for it as per Figure-A. This architecture can be successful for any component. Where the control unit is on a public access web server. The architecture described here is employed to design a system for the job of analyzing the link structure of web sites. This program had not run quickly enough to consider necessary number of websites and so it has been individually setup and run on a number of computers parallel. One more feature that was built into the crawler was a option of types of checking for duplicate pages to be used in a website crawl.

There are three (3) options-

1. Uses no page checking thinks that two page with different pages are different pages.
2. Use HTML page checking oppose new page which is identical HTML to later retrieved page.
3. Use weak HTML Page Checking.

Clustering is a method in which like records are grouped together. This technique is done to give the end user a high level view of what is going on in the data set. Sometimes clustering is performed not so much to keep records together as to make it easier to see when one record sticks out from the rest.

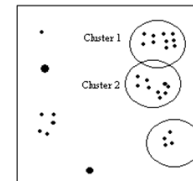


Figure – B

Clusters may be created either statistically or by using artificial intelligence methods. Clusters can be analyzed automatically by a program or by using visualization techniques as described in fig- B.

Criteria for site optimization:

For a new website to be optimized for the given keywords need to have some technical issues checked [5].

- ✿ Meta descriptions or Metadata Keywords
- ✿ Keyword analysis
- ✿ Title Tags
- ✿ Page content
- ✿ Headlines Tag
- ✿ URL structure and domain
- ✿ Images Tag
- ✿ Page Load time
- ✿ XML site Map
- ✿ Meta data using schema
- ✿ Site map
- ✿ Robot.txt
- ✿ 404 error
- ✿ Duplicate contents

Crawling Techniques:

- ➔ **Focused Crawling:**- Focused crawler is designed only to retrieve documents on a specific topic, thus reducing the amount of network traffic and downloads. The goal is to selectively seek out pages that are relevant to a pre defined set of topics. This leads to savings in Hardware and Network resources and helps keep the crawl more up to date.

➤ **Distributed Crawling:-** A single crawling process is non-useful for large scale engine that needs to fetch large amount of data rapidly. Distributing the crawling activity via multiple processes can help build a scalable, system which is fault tolerant system. Distributing the load decreases hardware requirements and at the same time increases the overall download speed and reliability.

Advantage:

Here we proposed the methods using Web Search in an efficient manner without any restrictions.

CONCLUSION:



In this work, we have identified a challenges during the architecture defined here is capable of crawling a large number of websites. It cannot process 100% automatic for jobs that involve crawling entire websites without heuristic for finding or searching duplicate pages. This design approach is suitable for the situations where a job can be divided into a disconnected crawling based job by which execution on different systems should not produce a problem. It may be non-useful if the crawls have to cross transmission each other in any case, for example to check a page from one crawl had already been found in another. Second case may be if the data mining has to be perform upon the whole data set is an integrated way.

FUTURE WORK:

In this work, though we have identified and studied and research a Search engine is a complex system on which further enhancements should be made. Some of the key ideas are like using query caching, disk allocation, sub indices, RAID techniques. More advanced algorithms are also required to decide which old pages should be re crawled and which should be new one crawl. Normal features by commercial Search engine like Boolean operators, negations, steaming use of AI should be added.

Acknowledgment:

This thesis paper is Heartily Dedicated to my parents Sri.S.Dandasi & Smt.Janaki, Mrs.Suneetha and My life Inspirer Eminent Scientist Sri.Dr.A.P.J.Adbulkalam.

Authors' Profiles:



Bosubabu Sambana working with as an Assistant Professor in Simhadhri Engineering College, Visakhapatnam. He is completed Master of Computer Applications and Master Degree in Computer Science & Engineering from Jawaharlal Nehru Technological University – Kakinada, Pursing Master of Science in Mathematics, Andhra University, and Andhra Pradesh, India. He has 4 years good teaching experience and having a good Knowledge on Space Research, Future Internet Architecture, Cloud Computing, Internet of Things/Services/Data, Computer Network and Hacking along with Computer Science Subjects. He is Published 4 Research Papers in various reputed International Journals and Magazines. He is the member of NASA, INTERNET SCOCIETY, W3C, MECS-PRESS, IAENG, IAAE and IJECSE.

REFERENCES

- [1] <http://www.en.wikipedia.org/worldwideweb> history.
- [2] web search engine in Google Search .
- [3] Search engine Time in Available: <http://www.en.wikipedia.org/>
- [4] Web search part:1,2 David Hawking,CSIRO ICT Centre.
- [5] Rajesh Singh, S.K. Gupta, IJAIEM- ISSN 2319 – 4847, Volume 2, Issue 9, September 2013
- [6] Images and Information – <http://www.google.co.in/>
- [7] All content are Available : <http://www.wikipedia.org/>