CHAPTER - 1

Introduction

This thesis presents design of a new Model of the Meta-Search Engine for getting optimized search results. The focus is on new dimension of internet search process.

1.1 INTRODUCTION

This chapter introduces the web, search engine, search engine optimization, meta-search engine and meta-search engine optimization. It discusses how search engine crawls the web, performs indexing and retrieves search results for user input search text. It also discusses how meta-search engine fetches results from the web for input search text.

1.2 THE WWW

WWW (World Wide Web) is an information retrieval system. It is mostly used for information retrieval through an experimental process, it is being developed rapidly. In an intellect WWW competes with the Gopher information system developed early in 90’s, which was used basically for information retrieval in its initial edge. WWW supports various kinds of information than Gopher information system like, text, images, videos, etc., to the user depending on their requirements. WWW can be used on a simple workstation.

The WWW is like an encyclopedia, a telephone directory, a record collection, a video shop, etc., all rolled into one and accessible through any computer. It has become so successful that it is synonymous with the Internet; but in reality the two are quite different. The Internet is like a network of electronic roads. The WWW is just one of many services using that network, just as many different kinds of vehicles use the roads. On the Internet, the WWW is just happens to be by far the most popular. [18]

The WWW is a collection of billions of documents written in a way that enables web pages to cite each other using hyperlinks. These documents, or
web pages, are typically a few thousand characters long, which are served through the Internet using the HTTP (Hypertext Transfer Protocol) to client computers, where they can be viewed using browsers. [19]

The capability to quickly fetch a large number of web pages into a local repository and to index them based on keywords is required in many applications. Large-scale programs that fetch tens of thousands of web pages per second are called crawlers, spiders, web robots or bots. Crawling is usually performed to subsequently index the documents fetched. Together, a crawler and an index form key components of a web search engine. [19]

WWW is based on hypertext, which means, among other things, that when the user is navigating on the ocean of information one can pick up an interesting word or expression within a text and request for more information about it. This does not apply to all words in a text but only to those who have been properly designated as such by the producer of the information and which are displayed on screen. In practice the use of WWW is still largely similar to the use of simpler, menu-driven information systems. [56]

While WWW starts or retrieves a large amount of information from somewhere far away. It may happen that some information is accessible just when user needs it, because some computer on the globe is not working or some cable has been cut off. The available information is not a representative sample from the pool of information known to mankind but the result of efforts of a set of people who wish to make information available in this way. [56]

The WWW is used for various purposes other than seeking information. Major focus is on communication and Information retrieval.
1.3 SEARCH ENGINES
Search Engines have their origins in Information retrieval (IR) systems, which create a keyword index for the given input search text and act in response to keyword queries with a ranked list of resultant hits.

Search Engines are a kind of tools which are designed to search information on the Web. The search engine results are generally presented in a vertical sequence often referred to as search results pages. Links available on those pages are referred as hits. Search engines like Google, Tineye, etc. also use advanced concept of data mining for effective results.

Search engine basically works on steps like, Web crawling, Indexing, Searching, etc. They work by storing information about web pages in databases, which they retrieve from the web. Figure 1.1 shows search engine architecture.

![Search Engine Architecture](image)

**Figure 1.1: Search Engine Architecture**

Web crawling is possible using a web crawler. It is a software program that traverses web pages, downloads them for indexing, and follows the hyperlinks that are referred on the downloaded pages. [11]
Indexing basically relates to search index. It is a data repository containing all the information the search engine needs to match and retrieve web pages. Data structure is used to organize the index here. It is very much like an index at the back of a book. It contains all the words appearing in the web pages crawled, listed in alphabetical order, called the index file. And for each word it has a list of references to the web pages in which the word appears is called posting list. [11]

Often, more information is stored for each entry in the index such as the number of documents in the posting list for the entry, that is, the number of web pages that contain the keyword, and for each individual entry in the posting file. It is possible to store the number of occurrences of the keyword in the web page and the position of each occurrence within the page. This type of information is useful for determining content relevance. [11]

The search index also store information pertaining to hyperlinks in a separate link database, which allows the search engine to perform link analysis, which is used as part of the ranking process of web pages. The database containing links can be used to reconstruct the structure of the web for good coverage. [11]

Search engine basically consist of set of instructions based on algorithms. The search engine provides user interface between information seeker and the search index database. The search engine gets the search terms and processes for search query.

Based on user search query search engine examines its index and provides a list of hits of matched pages. So, results with titles and links are displayed on screen based on indexing.

Search engines crawl the web from time to time and index the web pages. However, it is virtually impossible for any search engine to have the entire web indexed. Most of the time, a search engine can index only a small portion of the vast set of web pages existing on the Internet. Each search engine crawl the web separately and creates its own database of the content.
Therefore, searching more than one search engine at a time enables to cover a larger portion of the World Wide Web. [32]

Crawling the web is a long process, which can take more than a month whereas the content of many web pages keep changing more frequently and therefore, it is important to have the latest updated information, which could be present in any of the search engine. [32]

Crawling is directly intended for collection of URLs from the web.

The only way to collect URLs is to scan collected pages for hyperlinks to other pages that have not been collected yet. This is the basic principle of crawlers. They start from a given set of URLs, progressively fetch and scan then for new URLs and then fetch these pages in turn, in an endless cycle. New URLs found thus represent potentially pending work for the crawler. The set of pending work expands quickly as the crawl proceeds, and implementers prefer to write this data to disk to relieve main memory as well as guard against data loss in the event of a crawler crash. There is no guarantee that all accessible web pages will be located in this fashion; indeed, the crawler may never halt, as pages will be added continuously even as it is running. Apart from URLs, pages contain text; this is submitted to a text indexing system to enable information retrieval using keyword searches. [19]

The crawler’s role usually ends with dumping the contents of the pages it fetches into a repository. The repository can then be used by a variety of information retrieval systems and services which may, for instance, build a keyword index on the documents. [19]

The scope of a search engine depends on relevance of the result based on user search query. There are chances that there may be hundreds of web pages include common word or keyword or phrase. Some of them may be most relevant and some of them may not be. Actual requirement is desired result at proper position on search web page. So, most of search engines are using their ranking methods for this. Question here is how search engine
decides which hit is better than another? For doing these search engine uses one of the method from available ranking methods.

Search Engine provides better support to the user to search information from the WWW (World Wide Web) and to present results in a list form. That list basically contains information related to web pages, images or other type of information related files.

Search engines are one of the primary ways that Internet users find Web sites. For that reason a Web site with good search engine listings may see a dramatic increase in traffic. Here, main challenge is to provide efficient search techniques with adaptive approach. People basically require good listings. Unfortunately, many Web sites appear poorly in search engine rankings or may not be listed at all because they fail to consider how search engines work. [32]

In the current scenario information seekers are facing problems with number of links returned by individual search engines. So, information seekers waste their time in accessing unwanted links. Reason here is on indexing process of search engine. They perform indexing on the bases of keywords.

In particular, submitting to search engine is only part of the challenge of getting good search engine positioning. [32]

1.4 SEARCH ENGINE OPTIMIZATION

Search Engine Optimization is the art and the skill of getting URL to appear significantly in organic search engine results when a search process submits a query relevant to that URL.

Search Engine Optimization refers to improving results provided by Web Search Engine. In Web Search Engine it provides results list, whereas Search Engine Optimization provides some specific links for information related to web pages, images, etc for marketing purpose only.
Search Engine optimization (SEO) encompasses a wide variety of tasks that improve a website's presence on search engines. [12] Search engines are a primary channel for finding information on the Web, and top-ranking websites bring in more visitors than sites that rank poorly.

Search Engine Optimization has proven to be one of the most cost-effective form of marketing available. But using search engine optimization, one can retrieve information from one search engine only. It cannot access various search engines at a time. There is need of Adaptive Search Engine which will give flexibility to user to have choice of particular search engine from single user interface. [32]

Search engine optimization means ensuring that Web pages / Web sites are accessible to search engines and are focused in ways that help improve the chances they will be found. [32]

Most of web search engines are commercial ventures supported by advertising revenue and, as a result, some employ, the practice of allowing advertisers to pay money to have their listings ranked higher in search results. Search engines don’t take money from user to search results, but makes money by running search related advertisements alongside the regular search engine results. The search engines make money every time someone clicks on one of available advertisements. [47] Search engines are programmed to rank websites based on popularity and relevancy of search results.

For better search results Search Engine Optimization has to be employed. It is the process of affecting the visibility of link in search engine’s search result page. Hits retrieved by search engine in a natural way are known as organic search results which are un-paid. Presently, search engines are using search engine optimization for marketing purpose.
1.5 META-SEARCH ENGINES

Information Retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents. There are common characteristics in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval.

Meta-Search Engines are powerful tools that search multiple search engines simultaneously. Meta-Search Engine is a search tool that sends user requests to several other search engines and aggregates the results into a single list or displays them according to their source (on specific page and position based on rank). Meta-Search Engine enables users to enter search criteria once and access several search engines simultaneously. Meta-Search Engine operate on the premise that the web is too large for any one search engine to index it all and that more comprehensive search results can be obtained by combining the results from several search engines. If this facility is not there then user has to use multiple search engines separately.

It is known that existing meta-search engine do not have its own database. So, they do not retrieve search results from physical database. It works like middle agent between end-user and search engines. Instead, they fetch user search request, pass it to multiple search engines and retrieves results based on specific method.

After collecting all results meta-search engine will remove duplicate hits as per ranking algorithm, and results will be combined into a single merged list. Whatever ranking algorithm used here should satisfy user’s information need by displaying specific number of hits on screen in decreasing order of the rank. For giving assurance that results will meet user’s information need, there is a need to evaluate ranking algorithm also. Still optimization techniques are required to provide better result. For that one can think about good ranking strategies.
Meta-search engines do not search the web itself; rather they search existing search engine’s indexes. By searching more than one search engine’s index simultaneously, meta-search engines access more of the web in a single search. However, meta-search engines often do not search the best search engines; because of the fees such search engines charge. Also search engines that are busy with too many other searches at the exact moment information seeker conduct his/her search are sometimes skipped, so results can be inconsistent, Meta-search results are broad, but often not as deep as a single search engine’s. Meta-searching is a good place to start when information seeker want to check the first few results from several search engines. [21]

Maximizing meta-searching is important. And which search engines are used for meta-search should be identified properly, which may return most useful results. Figure 1.2 shows meta-search engine architecture.
1.6 META-SEARCH ENGINE OPTIMIZATION

Meta-Search Engine Optimization refers to improving results provided by Meta-Search Engine. In Meta-Search Engine as it is well known it provides specific links retrieved from various search engines. So still it is possible to minimize this number of possible links provided by Meta-Search Engine to optimize and provide better search results.

In existing meta-search engine optimization is advertisement prone techniques used. To get right and preferred position on particular search page for specific search engine Pay Per Click (PPC) strategy is used and in that way they are getting gain in their revenue. In most of the cases meta-search engines also do marketing like general search engines, but optimization should be for organic search (natural search). [73]

1.7 MOTIVATION

In the current development, Information Technology is growing rapidly. World Wide Web or the Web is one of the best inventions of it. The Web currently treated as a huge source of information. Appropriate methods are always required to extract needed information from the Web. Search engines like Google, Yahoo, Bing, etc. are the tools to retrieve needed information from the Web. Moreover, there are many meta-search engine exists. As we know, using meta-search engine one can fire input query to several search engines.

By doing literature survey, it is identified that existing meta-search engine works on a fixed number of search engines, and some of them are domain specific also. Moreover, existing meta-search engines do not allow the user to select search engine of own choice, since there is no dynamic management of search engine’s URLs. Also, existing meta-search engine fires input query to fixed number of search engines with stop words. Hence, optimum result is not obtained. Moreover existing meta-search engine are subject to time outs when search processing takes too long time. Also, there is a need of good ranking method in meta-search engine.
Moreover, to improve efficiency and to present results according to user like / dislike input there is a need to use indexing and database concept in meta-search engine. Also, most of the existing meta-search engine uses various marketing strategy as optimization technique to gain revenue. So, there is a need to focus on meta-search engine optimization also. These challenges have motivated to work in this area.

1.8 PROBLEM STATEMENT
As a part of research a new model of meta-search engine is proposed here to overcome above problems of existing meta-search engine.

1.9 OBJECTIVE
The main objective of this research work is to design and implement a new model of the meta-search engine for refining the search-results up to the desired level. The developed new model of meta-search engine,

i. Removes problem of ambiguous results with or without stop words in input stream.

ii. Provides user various search options like Aggregate search, LIKE search and Selection search for searching text. And Direct search for searching text, image and video link.

iii. Uses database concept for indexing purpose to improve response time.

iv. Allows user to select search engine(s) for search process.

v. Provides facility to user to rate links of search results.

vi. Improves effectively in resultant list using new ranking formula.

vii. Provides listing in an organic way.

viii. Allows administrator to manage search engine’s URLs dynamically.

ix. Allows administrator to manage stop words dynamically.

x. Allows administrator to update indexed links for searched terms.
1.10 STRUCTURE OF THESIS
The thesis comprises of eight chapters. An overview and objective of thesis is presented in following eight chapters.

Chapter 1: Introduction
This chapter gives introduction about the web, search engines, meta-search engines, working of meta-search engines and optimization of meta-search engines.

It discusses features required in a new meta-search engine. Moreover, it discusses the problem considered in this research area with its objectives. It also gives layout of the thesis.

Chapter 2: Literature survey and scope of research
This chapter introduces prior studies covering search engines, meta-search engines, difference between search engines and meta-search engines, ranking formulas, etc.

It also explains the features of existing meta-search engines, their limitations and need of design of new meta-search engine.

Chapter 3: Study of Related Technology
This chapter gives brief overview about each of used technology. Following technologies have been used for development of the new meta-search engine:

i. Operating system: ubuntu 09.04, ubuntu 12.04
ii. Open source server-side scripting language: PHP
iii. Scripting language: HTML
iv. Open source back-end: MySQL
v. Client-side scripting language: JavaScript
vi. Other technologies: CSS (Cascading Style Sheet) and Ajax
Chapter 4: New Model of the Meta-Search Engine
This chapter describes the model of designed meta-search engine. It also discusses different modules and features of the model like different search facilities: Default aggregate search, Selection search, LIKE search, and Direct search.

In addition to these, it also introduces briefly about Administrator control module as well as searching related to images and video content.

Chapter 5: Outcome of the model and Performance Analysis
This chapter describes the working of different features of the newly designed meta-search engine and discusses various outcomes through different searching options for text based search. Moreover, it discusses performance analysis of the model with existing meta-search engines, using various tools.

Chapter 6: Types of searches
This chapter discuss in detail about different search facilities available in new model of meta-search engine like: Aggregate search, Selection search, LIKE search and Direct search with their benefits.

Chapter 7: Administrator Control
This chapter discusses about various administrator control in detail about addition and deletion of individual search engine page URLs. It also discusses procedure for updation of databases by administrator.

Chapter 8: Conclusion and Future Scope
This chapter concludes the newly developed meta-search engine and its impact. It also gives ideas about future scope in this area. Research work in this era can be continued with different optimization techniques related to ranking formulas.

Designing Model for Meta-Search Engine
1.11 SUMMARY
This chapter gives introduction of the web, search engine, search engine optimization, meta-search engine and meta-search engine optimization. It also discusses working of search engine and meta-search engine. Moreover, it discusses challenges in developing a new meta-search engine. In the last objective of the proposed research work has been discussed. It also furnishes layout of thesis.