

CHAPTER 2

LITERATURE REVIEW

Shikha Goel *et al.* “Search Engine Evaluation Based on Page Level Keywords” (2012), this Paper highlights the approach of Search Engine evaluation which is based on page level keywords. Page level keywords are the keywords found in individual pages of website. Page level keyword is an impotent factor to measure the relevance of Search Engine results. A user create a query and Search Engine designer design the database for this query and later the queries are run by the users to calculate the page level keywords and the results are calculated. Keywords are grouping of words that user use to find products on Search Engine. A keyword can be any word on page but a stop word can't be a keyword for web page. Page level keywords include Title, header, and first word of title page, anchor text, page H1 tag, Meta description, image file name, ALT tags, and Page's URL string. Search Engine parse all the keywords from web page and create a database to store them into database. User's query indirectly evaluate on the dataset instead of web page. A relevant keywords occurrence count by Search Engine. A web page with higher occurrence of keywords takes high ranking in the search results [1].

Phyo Thu Thu Khine *et al.* Keyword Searching and Browsing System over Relational Database (2011), this paper highlights the searching of keywords in relational database for increase the searching speed of desired keywords. A user doesn't need the knowledge of database schema or SQL. A user submit a list of keywords them system search for the relevant records and ranking them on their occurrence basis. Three systems DISCOVER, BANKS and DBXplorer user Keyword Based Search over relational database. A query gives a set of keywords and finds rows in relation database for keywords. Indexing Relational Database: Indexing is used to speed up the retrieval of records. Indexing is useful when database has large number of Text fields. Each value in such a column considered as a small text documents that can be used for Keyword Based Search. Query Cleaning: System takes a query as input and produces a 'clean' query output. This is achieved by filtering the stop words from query. These words are meaningless. So the result occur with them may not satisfy the user. Keyword matching: Once the cleaned query is produced, this system can match the keywords. System matches the query keywords with database tuples. A keyword

matching algorithm may different for multiple keywords queries. Record Scoring: After query results, calculation of the score for each result is need. The record is determines which record is relevant to user query. This process is also called ranking process of documents in result [2].

Oladosu O. A. et al. “Researchers Social Interaction over a Responsive Knowledge Based Network” (2015), The study has been all about finding an easier, safer and educative ways of sharing opinions and interacts with other researchers in various part of the globe which can easily be referred to as Researchers Social Interaction (forum). In addition to chatting, this is a new and dynamics means of sharing ideas with different researcher all over the world within a flip of time. It has been discovered that Researcher Repository or Chat forum is one of the newest way in which Researcher and book authors of this world use now to reach out to their colleagues and numerous scholars in order to get feedback and responses concerning the entire areas of research topics to share of their idea and experience. [4]

John B. Killoran et al. “How to use Search Engine optimization techniques to Increase website visibility” (2013), This research paper highlights the two key points the ‘key concepts’ ‘key lessons’ to increase the website visibility using Search Engine Optimization Techniques. Professional point to some content on the web for themselves and or clients on websites they maintained. They use website title, key phrase, name of website, or name of organization that owns it. What contributes to Search Engine rankings? What can web content creators and webmasters do on their pages, sites and the web to make their content and sites easier to find by audience. Search rankings enable web content creators to continually monitor the exact measure of their competitive fitness and lacks for any given queries. Some websites content creators create their sites mainly directing their site for Search Engine ranking rules, not for human audience. Search Engine increase site ranking on the basis of their authenticity, topicality and quality, especially popularity. Two main factors affect SEO are, Trust – of which page rank is only the well know component, Relevance – a website topically matches with a particular query [7].

Zhou Hui, et al. “Study on Website Search Engine Optimization” (2012), this research paper highlights the Search Engine work principal. This research paper highlights the following contents of Search Engine optimization, Principal of Search Engine, Factors affecting search ranking, Method of website Search Engine optimization. Search Engine optimization points out the web page coding, web page content, web constitution, which describe how Search Engine work on web pages ranking strategies. Search Engine includes

following three categories: Full text Search Engine, Meta Search Engine, Directory Search Engine. Factor effecting search Ranking: three major factors include Webpage correlation: - determine the degree of keyword matching, keyword density, keyword distribution, and webpage tag labels. Link weight: - determine the internal link weights and the external link weight. Time-based factors: - includes the website and web page age, link age and the length of domain name registration time. Methods of websites Search Engine optimization are Keyword Optimization, Website content optimization, Website structure optimization, Webpage optimization, Link Optimization includes, Internal Links, Outgoing links, Incoming Links. Website submission includes, Submit to Search Engine, Submit to open directory library. [8]

Arooj Fatima, Cristina Luca, George Wilson, “User Experience and Efficiency for Semantic Search Engine” (2014), this paper highlights the limitations of traditional information retrieval system such as Keyword Based Search and highlights the efficiency of Semantic Search. Traditional information retrieval system expects specific type of query from user but user has no need to be worry about nay pre defined structure. This paper produced a technique that can be used for effective information retrieval and effective user experience for Semantic Search Engine. Paper includes the Semantic Search Engine’s structure from query engine to ranking of Web Pages. This paper presents a model of query optimizer to highlights the capability of Search Engine and indexing of WebPages. Query optimizer used ontology processor to optimize the user query [9].

S.G.Choudhary *et al.* “Semantic Search algorithm based on page rank and ontology: A review” (2013), this research paper introduces some Semantic techniques and provides different algorithms based on page ranking and ontology. The algorithms based on Semantic Search are the following ways: PSSE: Personalized Semantic Search Engine use the user’s profile with the ranking score and ontology for calculation of personalized factor which helps to get more personalized result. Google Search Engine use user profile to personalized search result for their location and on the basis of their previous search interest result. The architecture of PSSE has two parts: Offline and Online. The Offline part consists of crawling and preprocessing processes. The Online phase includes query processing and result ranking. Ontology Based IR Model: The approach of retrieval is based on adaption of classical vector space model. In classical vector model keywords are assigned weight to indicate that some words are better in documents than others. Weights are computed automatically by an

adoption of TFIDF algorithm. The vector space model includes two terms Weighting algorithm, and a Ranking algorithm. [10]

Robin Sharma *et al.* “Web Page Indexing through Page Ranking for Effective Semantic Search” (2012), this research paper highlights the architecture of Semantic information retrieval to enhance the search result. And an algorithm is purposed for Semantic indexing of web pages. Semantic web search is the search of meaning for user’s query. The meaning of query is hidden in the query itself. Words like What, Why, When affects the query meaning badly. In the traditional search these interrogational words removed or ignored from query and page ranking performed on the basis of main keywords. The search results in traditional search are not according to user satisfaction. This paper purposed a system that provides a result in order to relevance of keywords. The system is divided into three modules as follows User Interface Module: User enter the query through user interface whatever wants to search and gets the corresponding results. Query Analysis: Query entered by the user will be interrupted by preprocessing [11].

P. Chahal, M. Singh and S. Kumar “Ranking of Web Documents using Semantic Similarity” (2013), this paper purposed a novel technique which makes user search data quite efficient. This technique gives a relationship or similarity between searched document and user query. It is also consider the semantic structure of document and user query. The result set obtained from this approach gives better results than prevailing approaches. The future work can be done by using deeply semantic analysis of web pages and relevance of documents [13].

Seema Rani , Upasana Garg “A Ranking of Web Documents using semantic similarity and Artificial intelligence Based Search Engine” (2014), This paper proposed system presents a improved Semantic Similarity technique to rank a web page from a set o given web pages. System is using semantic similarity algorithm along with AI technique to rank the WebPages. System is tested on 1000 web pages comes under various categories like education, computer, programming, chemistry etc. Various input queries are given as an input to the system and results are checked. Overall accuracy of the system evaluated to 95%. System can be further improved by implementing on the cloud servers and by using multithreading techniques to improve the time efficiency. System can be further checked by increasing the number of web page categories. Multithreading techniques can also be integrated in the proposed system to improve the overall performance of the system [17] .

Ping-Tsai Chung *et al.* "A web server design using Search Engine optimization techniques for web intelligence for small organizations" (2010), This research paper highlights the ideal model for developing web servers for small organizations by including Search Engine optimization techniques. As we now today is a globalization era, a new way of business and marketing is developed by the use of Internet web services. Search Engine optimization is the process of improving the visibility of a website I the Search Engine result via "natural" search result. Small organization such as schools, banks, government agencies, libraries, retailers, restaurants, post offices could build their web services in higher quality for improving their business in surviving in today's competitive world. The web server design model includes the following three major components are The Best set of keywords, The Top-K Keyword proximity organization search, similar keyword search and image search for attracting new customers. Choosing best set of keywords is based on the comparative study result of Search Engines in keyword. Search in one area such as Information Technology area, or Business area, or Healthcare area in Wikipedia categories. Top-k keyword proximity organization search depends on the following two steps are Accurate Rate: - The information is perfectly found from the search result, it does not contain information that is unrelated. Recall Rate: - The information has similar related topics from the key word that you are searching. Similar keyword searches and image searches: - In similar keyword search experiments, we make a predefined table which list a keyword with its similar keywords. For example "Healthcare" and "Health care" are similar. Also "computer programmer" and "Programmer" are similar keywords. Design and analysis of web services affected by the followings Get web pages be in the search indexed, Choosing Best set of Keywords, Optimizing the web server, Attract Links. Quality of services of web servers directly depends on Trust and Reputation, Security, Enterprise Application Integration [18].

Fuxue Wang *et al.* "An Empirical Study on the Search Engine Optimization Technique and Its Outcomes" (2011), This research paper highlights the Effectiveness of SEO Techniques and Page Interest of Search Engine optimization techniques and figure out which technique strategy is most effective. Effectiveness of a page is defined by Page Interest. Page Interest is evaluated through visit on page. Page interest is also affected by another factor also such as Page view and Bounce Rate. The error 404 page not displayed when someone wants to find a page and that is not available on the site. It is possible that user don't want to visit on the same page. So to customize the 404 error pages might increase the chance to those visitors to continuously suffering the web site. Page size: - is a sum of the file sizes for all the elements that make up a page including defining HTML. Search Engine appreciating smaller

pages, page recommended size is 150 kilobyte. Title Length: - Most Search Engine use the web page title as main content of sites for the search result pages. The recommended length size is 80 characters. Keyword density: - keyword optimization defined as the percentage of keyword accounting for the total number of words in the page. /keyword density is important because Search Engine use this information to categories the theme of web site. The recommended keyword density is between 2% and 8%. [19]

Santosh Kumar Ganta, *et al.* “Search Engine Optimization through Web Page Rank Algorithm” (2010), highlight the Search Engine optimization with ranking algorithm. Focus on Semantic Search rather than keyword search. Page relevance is measured through a probability-aware approach that relies on several graph-based representations of the involved entities. Describe the prototype of a relational based Search Engine. Describe the mixed approach of semantic relation and ontology. Semantic Web will provide adequate instruments for improving search strategies and enhance the probability of seeing the user query satisfied without requiring tiresome manual refinement. Page relevance is measured through a probability-aware approach that relies on several graph-based representations of the involved entities [20].

Joeran Beel, *et al.* “Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co.” (2010), highlight the concept of Search Engine optimization. Based on three recently conducted studies, guidelines are provided on how to optimize scholarly literature for academic Search Engines in general and for Google Scholar in particular .and a complete guide of design a research paper to publishing policies. Also highlight the searching algorithms in Google Scholar. Google Scholar is one of those Search Engines that combine several factors into one ranking algorithm. The most important factors are relevance, citation count, author name, and name of publication [21].

George S. Spais, “Search Engine Optimization as a dynamic online promotion technique: the implication of activity theory for promotion manager” (2010), the paper highlight the possibility of an extension of Bedny’s perspective of ‘activity’ theory as a framework for the elaboration of new online promotion channels, such as the Search Engines. This concept was approached as a framework for Search Engine Optimization (SEO) contextual issues, which can be used to aid the design and analysis of the SEO promotion technique investigations. SEO’s objective is to increase web visitor counts by ranking an organization’s web site very high as a result of searches using the most appropriate keywords

describing the promotion offer through a web site. Therefore, effective SEO promotion technique can result in higher sales and greater revenues [22].

Duygu Tumer *et al.* “An Empirical Study on Semantic Search Performance of Keyword-Based and Semantic Search Engine: Google, Yahoo, Msn and Hakia” (2009), this paper analyzes the Semantic Search performance of Search Engines. This paper includes three Keyword Based Search Engines like Google, Yahoo, Msn and a Semantic Search Engine Hakia. Different queries of different topics analyze the performance of these Search Engines. Web Search Engines are computer programs which allow users to search their desired information from websites. The most popular Search Engines are Google, Yahoo, and Msn with 71.9, 71.7% and 4.2 volume of search ratio respectively. Hakia is the publicly available Semantic Search Engine. This paper has a table with ten different types of queries. These queries were run on the both keyword-based Search Engine as well as Semantic Search Engine. Keywords were used to replace phrase. Beside the keywords phrase were used in Hakia for the main feature of Hakia Semantic Search. Paper also highlights the concepts of relevant and non-relevant documents. A document which matches with the query keywords is called relevant and which don't called non-relevant. Google, Yahoo, Msn retrieved at least one relevant document for all queries. Yahoo and Msn retrieved approximately 75.5%, 63% and 78% non-relevant documents respectively. Hakia retrieved 62.5% non-relevant documents [23].

Junaidah Mohamed Kassim *et al.* “Introduction to Semantic Search Engine“ (2009), This research paper highlights the Semantic Search Engine design and use as well as traditional Search Engine. The first generation of web ‘web1.0’ from 1990 – 2000 refers to internet at its emerging stage and produces a Producer – Consumer relationship. Web2.0 transforms the Web into a space that allows anyone to create and share information online. Web 3.0 shows more intelligence like the ‘web machine’ learns, suggests that what people like and would like to get. Semantic Search integrates the technologies of Semantic Web and Search Engine to improve the search results gained by current Search Engines. General processes of Semantic Search Engine are: - 1). The user query is interpreted and extracting the relevant concepts from the sentence. 2). the set of concepts is used to build a query that is launched against the ontology. 3). the results are presented to the user. The use of Semantic Search Engine: - Semantic Search use to improve the results of research searches in two ways, 1) traditional search results take from of a list of document and web pages. 2).The search phrase in research researches denotes one real word objects [24].

Yi Jin *et al.* "The Research of Search Engine Based on Semantic Web" (2008), this research paper highlights the possibilities to improve a traditional Search Engine to create a Semantic Search Engine. This paper describes the semantic web as well as the architecture of Semantic Search Engine. The semantic web uses the tag notes to enrich the text documentation which machine can understand. These tag notes provides the Meta-data of these texts, and will capture the meanings of these texts, which can be understand by machine. Meta-data are the fundamental building blocks of the semantic web. The researching purpose of Semantic Web is to describe the Web documents clearly. The foundation of the Semantic Web is the markup languages. The Architecture of Semantic web Search Engine focused on Retrieval and Reasoning driven processes. The system should be able to run some reasoning engine to identify the necessary facts and rules so as to achieve the desire conclusion. The semantic web tags can be used to get semantic notes. The major premise of Semantic notes is getting key words, terms and other elements from text documentation [25].

Bernard J. Jansen *et al.* "Determining the informational, navigational and transactional", (2007): this paper highlights the comprehensive classification of user intent in Web searching. Classification includes the informational, navigational, transactional level. This paper present a automatic classification model of Web search queries and identified that more than 80% queries are informational by nature and 10% are navigational and transactional. 70% of Web search user use Search Engine as their point of entry. To analysis the query classification paper collects 50 queries from computer science student. Paper classified that 26% queries are misclassified [26].

Bernard J. Jansen *et al.* "Determining the User Intent of Web Search Engine Queries" (2007), this paper highlights a method to determine the user intent underlying Web Search Engine queries. This paper analyzes the samples of queries from seven transaction logs from three different Web Search Engines containing more than five million queries. From this analysis, paper identified the characteristics of user queries based on three broad classifications of user intent. The classifications of informational, navigational, and transactional represent the type of content destination the searcher desired as expressed by their query. This paper show that more than 80% of Web queries are informational in nature, with about 10% each being navigational and transactional. This paper classified company or organization name in Navigational query. Navigational queries are short in length as compare

to other query category. Finally this paper concludes that Search Engines are used as informational tools rather than Navigational or Transactional tool [27].

Bernard J. Jansen *et al.* “Web Searcher Interaction with the Dogpile.com MetaSearch Engine” (2007), this paper highlights the analysis of Meta Search Engine Dogpile.com queries. Paper presents the usage of dogpile.com Search Engines and user’s intention towards the Search Engine. This paper identified that 56% users generally spend less than a minute with Web Search Engine and Meta Search Engine have higher user interaction as compare to Non- Meta Search Engine. This Search Engine used by 84 % user in USA and enters less than 3 terms per query. The maximum query length was 25 terms and 75% of the queries were less three or less terms. This Paper also highlights the session of different user with Search Engine, 81% of sessions were less than 15 minutes and nearly 72% of the sessions were less than five minutes [28].

Ricardo B. Yates *et al.* “The Intention Behind Web Queries” (2006), this paper presents Search Engines as a resource manager which provide their resource to users for popularity and profit. This paper present a framework for the identification of user’s interest in an automatic way, based on the analysis of query logs. This identification is made from two perspectives, the objectives or goals of a user and the categories in which these aims are situated. This paper categorize Web Search Engine query into three categories named Informational query, Not-Informational and Ambiguous Query Category. An Informational Query is that in which user wants to extract information from web for knowledge purpose. Ambiguous Queries are those that their goal cannot be identified from the query. This paper used Open Directory Project category to classify the categories [29].

Uichin Lee *et al.* “Automatic Identification of User Goals in Web Search” (2005), this paper highlights the goal of user’s web query, so that this goal can be used to improve the quality of Search Engine results. This paper firstly presents a result from human subject study and then proposes two types of features to identify the query, Past User-click behavior and anchor-link distribution. Navigational query can be identified with Past User- click behavior, if a query is Navigational user will hit the result which he/she has already in mind. If a user has no particular page in mind then query will be consider as Informational query. This paper mainly highlights the Navigational and Informational query. Paper presents that 60% of queries have predictable goals, where 40 % of queries have no predictable goal [30].

Daniel E. Rose *et al.* “Understanding User Goals in Web Search” (2004), this paper highlights the framework to manually classify queries from a web Search Engine. Paper identifies not only *what* user search but also *why* they are searching. Paper uses the Alta Vista Search Engine’s 500 queries to identify the results. Paper presents a manual classification of user queries. Paper highlights that 40 % of queries are non-informational, on the other side over 35% of all queries are for general research goal. From all query log 10% queries are sexually oriented [31].

Andrei Broder, “A Taxonomy of Web Search” (2002), this paper highlights the Search Engines query categories into three parts Navigational, transactional, Informational. This paper presents a survey in which different questions are answered by many users of Search Engine and finally results of user intention presented. Second view of this paper is analysis of Search Engine query log. Analysis shows the total intention of user query and categorizes the query. Combine results of both survey and Search Engine Query log identified that 45 % queries are informational. Navigational and transactional Queries are 20% and 30% respectively. This paper also highlights the generations of Search Engines, Internet and their consequence. Paper’s motive is to identify the third category of query name Transactional query [32].

Bernard J. Jansen *et al.* “Real life, real users, and real needs: a study and analysis of user queries on the web” (2000), this paper analyze transaction log of Excite Internet Search Service, containing 51,473 queries posted by 18113 users. This Paper identified the characteristics of Search Engine users. This Paper produced results on the basis of following three Session, Queries, and Terms. Paper presents that web queries are short. About 62% of queries contained one or two words. Fewer than 4% of queries had more than 6 terms. Paper presents that one in 18 users used Boolean capabilities in query and about one in 190 queries used nested logic. Paper also highlights the average user views 2.35 pages of results where one page equals ten hits. Over half the users did not access results beyond the first page [33].