An Integrated Approach To Search Engine Evolution

Dharminder Singh[1], Ashwani Sethi[2]
Guru Gobind Singh Collage of Engineering & Technology
Guru Kashi University Talwandi Sabo, Bathinda, Punjab[1, 2]
dharminder.17@gmail.com[1], ashwani.gku@gmail.com[2]

Abstract: *Information retrieval systems are backbone of information technology. Everyday millions of users use the internet informational System like Search Engines to retrieve information over the WWW. Search Engine gives a high level abstraction to users over the web. On the basis of searching techniques Search Engine can divide into two broad categories Keyword Based Search and Semantic Search. This research proposes combine architecture of these two searching categories. The Proposed System combines these two techniques on the basis of query categories. This system's design works as a load balancing technique between these two techniques for more accurate and fast results. This system implements both techniques separately for different query category.*

Keywords: *WWW, URL, SERP, Inverted Index, Ranking, Web Crawler.*

I. INTRODUCTION

In this 21st century Internet has become an essential part of human being's life. Almost every education system, Govt. Organizations or any other part of human life, use directly or indirectly the Internet. Internet provides facilities to reach on the WWW, which can locate on anywhere in the world. Search Engine play an important role to explore the internet and also describe their uses for users. Every Search Engine works as a resource manager and produces their resources to users. Popularity of Search Engine depends on their resource used by their users. Search Engine is not only a tool which provides services for users, but also an effective tool to understand user's behavior [2]. Search Engines are primarily responsible for quality and updated results to their user. Search Engines have three primary tasks: Query Analysis, Searching of WebPages and Ranking of Web Pages. On the basis of searching every Search Engine falls into two broad categories: Keyword Based Search Engine and Semantic Search Engine. Both type of searching Keyword Based Search and Semantic Search have different searching algorithms to target different contents of web pages. Ranking of Web page always target the popularity of web page and quality content. Searching is more complicated task as compare to ranking. Once a quality Web Pages searching have done, Ranking is too easy with

quality of web pages. Keyword Based Search target only query keywords against web page contents. A web page with high occurrence of query keywords will get high rank in SERP. Keyword Based Search can easily be confounds by Black Hat techniques [1]. Spammer use a hidden density of popular keywords in web page to confounds the Search Engine [11]. These hidden keywords are only visible to Search Engines not for users. Semantic Search use knowledge base to refine and extend the query with pre described knowledge for particular area of search. Knowledgebase store enough terms related to query which describe query more clearly to Search Engine and search with more keywords on WWW. After analysis of Keyword Based Search and Semantic Search we have identified that Keyword Based Search is fast as compare to Semantic Search and Semantic Search produced more relevant results as compare to Keyword Based Search. Every Search Engine has three major tasks to produces the results to user query.

A. **Query Analysis:** Every search against users query will search on personal database of Search Engine instead on the Internet every time. Search Engine reliability more depends on user Query. Query works as a feed for Search Engine and describe the user's intent of search to Search Engine. If Search Engines can judge user's intention of search then they can produces more relevant results to users. User describes their requirement of search with some short keywords to Search Engine, now there can be several meanings for a particular query. The meaning of query is hidden in query itself [7]. Search Engine has primarily responsibility to understand the query meaning and produced quality results. Query judgment is a complicated task where two same queries may have different meanings for search. Query can be categorized as Navigational query, transactional Query and Informational Query [20]. These all categories point out user's intention of search in the Search Engine.

B. **Indexing:** Every Search Engine use Web Crawler to systematic search on WWW for web indexing. Web Crawler is also called Web Spider and works as robot [4], which automatically search updated results on different web pages and index them into Search Engines personal database called inverted index database. Inverted indexing provides indexing for contents of web pages with their location in particular web page [16]. Indexing of Web pages works same as indexing on last page of book.

C. **Searching & Ranking:** After the query analysis searching of web pages into Inverted Indexed database is second major task of Search Engine. Searching algorithms target the different content of web pages to judge the quality of web page [3]. Keyword Based Search hits the <meta> tag, <title> tag, <header> tag, <alt> tag; <body> tag, URL keywords, internal and external link text [8]. Ranking of web page also affected by density and best set of

keywords [5] [6]. The recommended keyword density is between 2% and 8%. [10]. Web Crawler Searching with Semantic Search algorithm produced more relevant results with the help of knowledgebase. Semantic Search extends the query and then put query to search. Knowledgebase content determines the enough knowledge for the query related areas. Search performs along with the knowledgebase keywords instead of only query keywords. After searching the query related results, ranking of these results is more important, because of human nature to hit top results in SERP. The average user views 2.35 pages of results where one page equals ten hits [24]. Ranking of web pages falls in: Popularity of Web Page and their quality contents [13].

II. LITERATURE SURVEY

Phyo Thu Thu Khine proposed 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 submits a list of keywords then system search for the relevant records and ranking them on their occurrence basis. Indexing Relational Database 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: The 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. The 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 [9].

Junaidah Mohamed Kassim presents Semantic Search Engine design and use as well as traditional Search Engine. Every Search Engine has three essential parts: A database of web documents, a Search Engine operating on the database and a series of programs that determine how search results are displayed. This paper presents a close view of web generations. The first generation of web 'web1.0' form 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 off Semantic Web and Search Engine to improve the search resulted gained by current Search Engines [15].

Bernard J. Jansen presents 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 [18].

Duygu Tumer analyzes the Semantic Search and Keyword Based Search Engine performance of Search Engines. This paper takes 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 [14].

Shikha Goel proposed 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.

III. PROPOSED WORK

The proposed system is an integrated approach of Keyword Based Search and Semantic Search based on the Identified query category. The proposed system automatically identifies the user's intention of search through query and after identified the query category system pass this query to search. The System considers all the queries fall into three categories: Navigational, Transactional and Informational. To identify the query, system use knowledgebase, this contains enough knowledge about particular query, to identify the query category. Entire system can be divide into three modules; 1. Query Analysis, 2. Navigational and Informational query implemented with Keyword Based Search, 3. Informational Query implemented with Semantic Search.

Proposed System Algorithm:

Input: Query Keywords q1

Output: Ranked Results

Step 1: Apply query Category algorithm on query q1.

a) For Each Keyword k in Query q1.
   I. Match k with Query Category knowledgebase.
   II. Calculate the frequency of matched keyword.
   III. Store Keyword frequency in qkey1 table.
b) If qkey1 table has any stored keyword and query length is <=4
   I. Relate query q1 with specific category based on stored keywords.
c) Else relate query q1 with information query category

Step 2: Apply refined algorithm on q1.

a) Removal of stop words from q1.
b) Retrieve the keyword from q1.

Step 3: If Query q1 is Transactional or Navigational in Nature AND query length 4< go to Step 4 else go to Step 7

Step 4: For each keywords z in query

a) Match key z against web page keywords in database.
b) Calculate the frequency of matched keywords with in a webpage.
c) Store keyword frequency in keyword1 table.

Step 5: Sort the web pages stored in keyword1 table from high to low frequency.

Step 6: Show ranked results of sorted web pages from high to low frequency.

Step 7: For each Keyword y in query

a) Find equivalent keywords of key y in knowledgebase.
b) Store all equivalent in to an ekeys table.

Step 8: For each equivalent x in an ekeys table

a) Match x against web page keywords in database.
b) Calculate the frequency of matched x keywords with in a webpage.
c) Store keyword frequency in keyword2 table.

Step 9: Sort the web pages stored in keyword2 table from high to low frequency.

Step 10: Show ranked results of sorted web pages from high to low frequency.

As previously mentioned that after analysis of both searching techniques, this is clear that Keyword Based Search is fast and Semantic Search produced more relevant results. This system gives Navigational and Transactional queries to Keyword Based Search, where user wants quick results with a particular web page. For Navigational and Transactional queries system doesn't perform complex searching with knowledgebase. The System just shows the web pages related to query keywords directly. If a user has no particular page in mind then query will be consider as Informational query [21]. If query is identified as Informational and query length is greater than or equal to four, system will put this query to Semantic Search, where knowledgebase keywords extends the query and produced more accurate results and rank them according to density of quality contents. As Bernard describe that 80% of queries are informational in nature and 10% queries are Navigational or Transactional in nature [18] and the maximum query length was 25 terms and 75% of the queries were less three or less terms [19]. So this system only consider Navigational and Transactional

queries with knowledge base keywords and if query is not fall into these categories , The System by default assume that query is Informational in Nature.

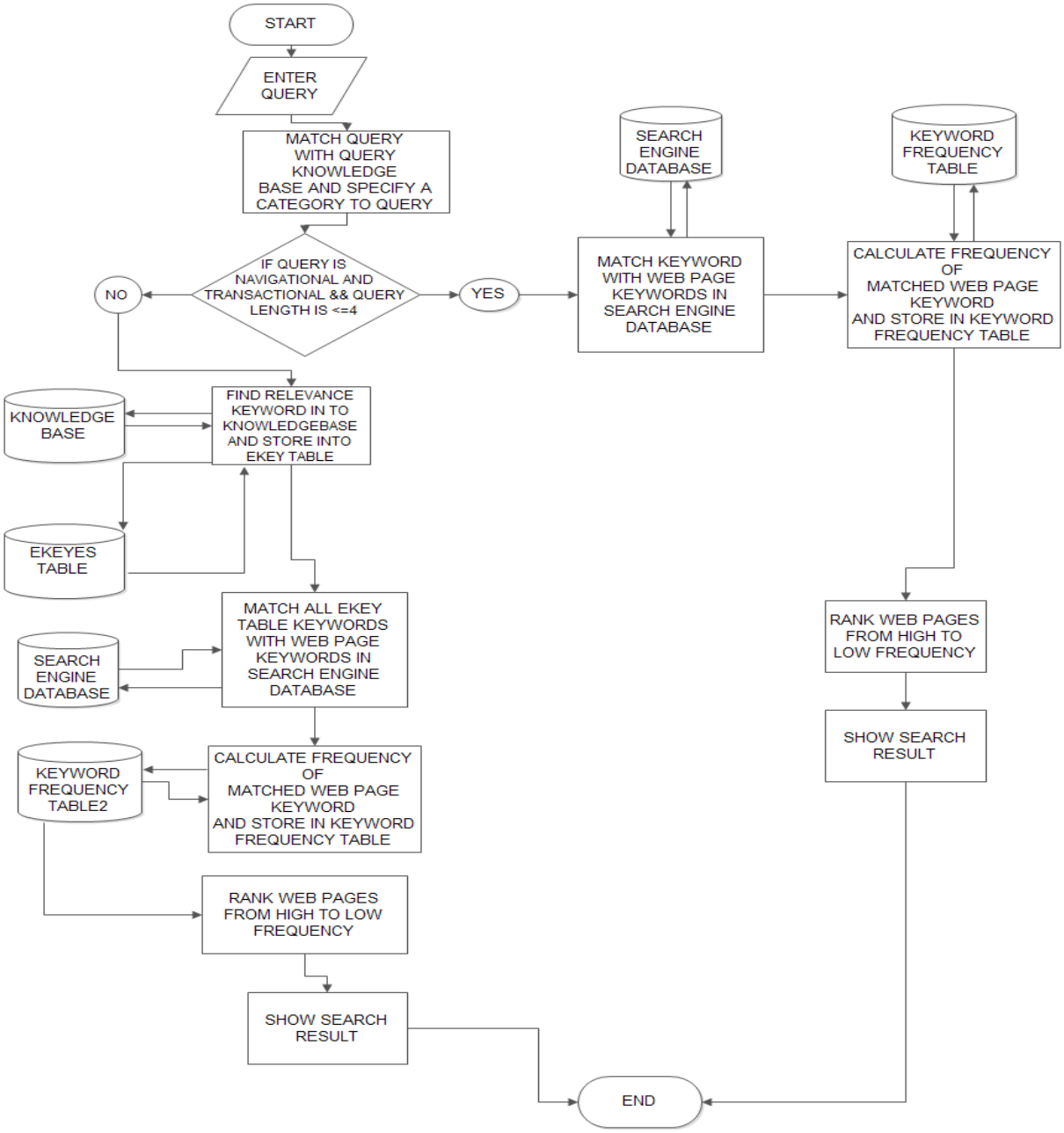

Figure 1: Flow Chart of Proposed System

IV. RESULTS & DISCUSSION

The Proposed System results include different parameters to test system quality. The result of proposed system defines different view of system results and results are divided into three categories of query and show each query category results separately. The System has tested on various different queries with different length. Each query category tested with different keywords and for accuracy of produced results. The Proposed System contains 1000 different web pages of different categories like Education, Sports, Literature and Business etc. Query categorizes system tested on different user's different queries for clearly identification of user intention. The System categorizes 80% of all queries. The Proposed System tested with more than 500 queries against 1000 stored web pages.

Table 1: Proposed system Results

Parameters/Proposed Techniques	**Keyword Based Search**		**Semantic Search**
	Navigational Query Results	Transactional Query Results	Informational Query Results
Time Efficiency	71%	74%	41%
Accuracy	75%	80%	87%
Categorize Queries	82%	77%	75%
User Query Specific Results	71%	75%	87%

V. CONCLUSSION & FUTURE SCOPE

The Proposed System presents combine architecture of two searching techniques Keywords Based Search and Semantic Search. This architecture based on the query category. The System puts a specific category to specific search technique. This proposed research uses the merits of two different searching techniques intelligently. The Proposed System can divide into three parts: Categorize to queries, Keyword Based Search for Navigational and Transactional Queries, Semantic Search for Informational queries. Every type of query has specific searching technique according to query requirement of search. Navigational and Transactional queries have less requirements of search resource so these types of queries handled by Keyword Based Search and gives quick results. Informational Queries have much required resources of search so Semantic Search handles this type of queries. Informational Query is less time efficient as compare to other query categories because of

this query category goes to Semantic Search. Approximate 20% of all queries are not recognizable to any query category. These non recognizable queries effects to The Proposed Systems Accuracy and User query Specific Results parameters. The System can be further improved by analysis more parameters. A new type of Query category can be developed to extend the system capabilities to handle the query.

REFERENCES

[1] Patil Swati P., Pawar B.V. and Patil Ajay S., Search Engine Optimization: A Study, Research Journal of Computer and Information Technology Sciences, Vol. 1(1), P.P 10-13, February 2013.

[2] Khalil ur Rehman and Muhammad Naeem Ahmed Khan, The Foremost Guidelines for Achieving Higher Ranking in Search Results through Search Engine Optimization, International Journal of Advanced Science and Technology, Vol. 52,P.P 101-110, March 2013.

[3] John B. Killoran, How to use Search Engine optimization techniques to Increase website visibility, Transactions Professional Communications IEEE, ISBN: 0361-1434, 2013.

[4] S.G.Choudhary, S.R.Kalmegh and Dr. S. N. Deshmukh, Semantic Search Algorithms based on Page Rank and Ontology: A Review, 3rd International Conference on Intelligent Computational Systems IEEE, 2013.

[5] Zhou Hui., Qin Shigang., Liu Jinhua and Chen Jianli, Study on Website Search Engine Optimization, International Conference on Computer Science and Service System IEEE, 2012.

[6] Ping-Tsai Chung, Sarah H. Chung and Chun-Keung Hui, A web server design using Search Engine optimization techniques for web intelligence for small organizations, Systems, Applications and Technology Conference IEEE, ISBN: 978-1-4577-1342-2, 2012.

[7] Robin Sharma, Ankita Kandpa and Priyanka Bhakuni, Web Page Indexing through Page Ranking for Effective Semantic Search, Proceedings of7'h International Conference on Intelligent Systems and Control IEEE, ISBN: 978-1-4673-4603-0, 2012.

[8] Shikha Goel, Sunita Yadav and Raj Kumar Goel , Search Engine Evaluation Based on Page Level Keywords, IEEE , ISBN: 978-1-4673-4529-3, 2012.

[9] Phyo Thu Thu Khine, Htwe Pa Pa Win and Khin Nwe Ni Tun, Keyword Searching and Browsing System over Relational Databases, IEEE, ISBN: 978-1-4577-1539-6, 2011.

[10] Fuxue Wang., Yi Li and Yiwen Zhang, An Empirical Study on the Search Engine Optimization Technique and Its Outcomes, IEEE, ISSN: 978-1-4577-0536-6, 2011.

[11] Santosh Kumar Ganta and Satya P Kumar Somayajula, Search Engine Optimization through Web Page Rank Algorithm, International Journal of Computer Science and Technology, Vol. 2, ISSN: 0976-8491, P.P 427 -431, September 2011.

[12] Joeran Beel., Bela Gipp and Erik Wilde, Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co., Journal of Scholarly Publishing, P.P 176–190, Vol. 41 (2), January 2010.

[13] George S. Spais, Search Engine Optimization as a dynamic online promotion technique: the implication of activity theory for promotion manager, Innovative Marketing, Vol. 6, P.P 7-24, 2010.

[14] Duygu Tümer, Mohammad Ahmed Shah and Yılmaz Bitirim, An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia, Fourth International Conference on Internet Monitoring and Protection IEEE, ISBN: 978-0-7695-3612-5, 2009.

[15] Junaidah Mohamed Kassim and Mahathir Rahmany, Introduction to Semantic Search Engine, International Conference on Electrical Engineering and Informatics IEEE, ISBN: 978-1-4244-4913-2, 2009.

[16] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, An Introduction to Information Retrieval, Cambridge University Press, 2009.

[17] Yi Jin, Zhuying Lin and Hongwei Lin, The Research of Search Engine Based on Semantic Web, International Symposium on Intelligent Information Technology Application Workshops IEEE, 978-0-7695-3505-0, 2008.

[18] Bernard J. Jansen, Danielle L. Booth, Amanda Spink "Determining the user Intent of Web Search Engine Query", WWW2007 Poster Paper, 2007.

[19] Bernard J. Jansen, Amanda Spink," Web Searcher Interaction with the Dogpile.com MetaSearch Engine", Journal of the American Society for Information science and Technology, 2007.

[20] Ricardo Baeza-Yates, Liliana Calderon-Benavides," The intention Behind Web Query", SPIRE, 2006.

[21] Uichin Lee, Zhenyu Liu and Junghoo Cho, "Automatic Identification of User Goals in Web Search", WWW2005, 2005.

[22] Daniel E. Rose, Danny Levinson," Understanding User Goal in Web Search ", WWW2004, 2004

[23] Andrei Broder," A taxonomy of web search", SIGIR Forum, Fall2002, Vol.36 no.2, 2002.

[24] Bernard J. Jansen, Amanda Spink," Real life, real users and real needs: a study and analysis of user queries on the web", Information Processing and Management, page 207-227, 2000.