Abstract—There is a need to quickly go through large amounts of textual information to find out documents related to our interests and this document space is growing on a daily basis at an overwhelming rate. It is common to store several million web-pages and hundreds of thousands of text files. Analyzing such huge quantities of data can be made easier if there is a subset of words which can provide the main features, concept, theme etc of the document. Appropriate keywords can serve as a highly concise summary of a document and help to easily organize documents and retrieve them based on their content. These keywords are provided by the SEO Suggester. The core methodology used in SEO is to upgrade both content and associated coding of the website to improve its visibility and importance in organic searches made by the search engines. The SEO optimized websites acquire better ranks and typically get a higher number of visitors. Page Rank is calculated based on the number of inlinks. Web crawler retrieves the link structure from the repository of web pages and obtains web graph from the website.

Keywords—Search Engine Optimization, Search Ranking, Web Crawling, Inlinks, Text Mining

I. INTRODUCTION

Many new challenges are created by the World Wide Web for information retrieval. It is very large and heterogeneous. Current estimates are that there are over 150 million web pages with a doubling life of less than one year. Most importantly, the web pages are extremely diverse. In addition to these major challenges, search engines on the Web must also contend with inexperienced users and pages engineered to manipulate search engine ranking functions. However, unlike document collections, the World Wide Web is hypertext and provides considerable auxiliary information on top of the text of the web pages, such as link structure and link text. The most relevant pages on the user's query should be ranked at the top. This can be done by taking advantage of the link structure of the Web to produce a ranking of every web page known as Page Rank. The search engine is a methodology which is based on different strategies using a computer program which is used to gather information from the Internet, and organizing and processing the searched information for displaying the search results to the users in a systematic way. Search engine optimization is a process that makes use of a search principle for search engine to acquire higher ranking for web-pages or websites to enhance the probability of website access. Since the surfacing of Google, Yahoo, Ask, Blekko, Volunia and Bing, the SEO principles and search engine innovations have been in the process of development. Page rank is an algorithm that is used to measure the importance of website pages using hyperlinks between pages. Some hyperlinks point to pages on the same site (inlink) and others point to pages in other website (outlink). Page rank is a vote, by all other pages on the web, about how important a page is. A link to a page counts as a vote of support.

II. MOTIVATION

A number of challenges have been faced for information retrieval. The quantity of information on the web is growing rapidly, as well as the number of new users inexperienced in the art of web research. People are prone to surf the web using its link graph, often starting with high quality human maintained indices such as Yahoo or with search engines. Human maintained lists cover up popular
topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics. Automated search engines that rely on keyword matching usually returns too many low quality matches. To make matters worse, some advertisers attempt to gain people's attention by taking measures meant to mislead automated search engines. Web page ranking using seo suggester addresses many of the problems of existing systems. It makes especially heavy use of the additional structure present in hypertext to provide much higher quality search results.

III. PROBLEM STATEMENT

The problem statement is “To Rank the most relevant pages of the user's query at the top of search engine”.

This project aims to demonstrate the process of ranking the most relevant pages of users query at the top. To show this working will be on page ranking based on the number of inlinks. To a web page, an inlink is the URL of another website which is directed to web page. As mentioned earlier, all traditional search engines do not rank the most relevant pages of users query. SEO optimized websites attract more number of visitors. The SEO suggester in this project will suggest suitable keywords to improve the ranking of the web page by text mining the contents of the website. Based on the calculation of page rank, this analysis includes two main problem statements:

- Ranking most relevant pages of users query at the top.
- SEO suggestions to improve the ranking of the webpages.

IV. LITERATURE SURVEY

A literature survey is a text of a scholarly paper, which includes the current knowledge including substantive findings, as well as theoretical and methodological contributions to a particular topic. Literature reviews are minor sources and do not report new or original experimental work. They present the study and analysis of existing systems. They allow using proposed scheme to develop new methods for removal of research gaps. The Table 2.1 shows literature survey of the proposed system.

Hema Dubey [1] in the research uses a normalization technique based on mean value of page ranks and calculates the rank of individual web page and Hypertext Induced Topic Search (HITS) depends upon the hubs and authority framework. A fast and efficient page ranking mechanism for web retrieval remains as a challenge. The proposed scheme reduces the time complexity of the traditional Page Rank algorithm by reducing the number of iterations to reach a convergence point. Madhurdeep Kaurand Chanranjit Singh [2] proposed a page rank mechanism called Hybrid Page Rank Algorithm which is based on both content and link structure of the web pages. This algorithm is used to find more relevant information according to users query. This paper also presents the comparison between SimRank Algorithm and the Hybrid Page Rank Algorithm. 

Cui and Hu [3] explain the specific requirements for optimizing the search queries, and present design concepts and novel website building based on the empirical research pertaining to internal coding method and website contents. Further, the authors elaborate search engine optimization tools and strategies that highlight the e-commerce sites for the sake of effective website promotion. When a users searches a website using an optimized search engine, then the whole website can attain higher ranking position. This improves the traffic of website and enhances its sales capability. This makes it vital for employing specific tools, strategies and friendly methods for SEO search engines.

Hui-ye [4] in their research elaborated that the rank of the motel sites and its bandwidth increased after the execution of SEO techniques for Internet marketing. Several techniques of SEO to increase the bandwidth and ranking of results of search including text title, label text, picture note, HTML
modification, map offsite, open website of catalogue registered in DOMS, web ping, discussion of internet boards and signature lines for keywords were used by the authors.

Chung and Hui [5] state that SEO tools can help organizations like banks, governments and other institutions to improve their web services in order to boost their business in the current world competition. The authors highlight specifically using image searches, proximity organic searches and top-keyword for optimizing the web server: The SEO tools and techniques which are used for web server development include: search indexed getting web pages, optimizing the web server, choosing the correct set of keywords, on-site web analytics, to attract links and off-site web analytics. In addition, the supplementary web intelligence techniques are query ranking factors and page raking factors. The quality of services can be ensured by focusing on Trust, Performance, Reliability, Enterprise Application Integration, Security and Reputation.

<table>
<thead>
<tr>
<th>Sr.No</th>
<th>Paper Name</th>
<th>Proposed System</th>
<th>Research Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>An Improved Page Rank Algorithm Based On Optimized Normalization Technique Ilena Dubey 2011</td>
<td>New Page Rank algorithm which uses a normalization technique based on mean value of page value of page rank</td>
<td>Page Rank Algorithm based on Damping Factors favours older pages and doesn't link with new pages</td>
</tr>
<tr>
<td>2</td>
<td>A Hybrid Page Rank Algorithm An Efficient approach Madhurdeep Kaur, Chamranjit Singh August 2014</td>
<td>A comparison is made between the SimRank and Hybrid page algorithm on the basis of accuracy and time of retrieval</td>
<td>A limited Environment for ranking of web pages which are more relevant to users query</td>
</tr>
<tr>
<td>3</td>
<td>Search Engine Optimization for website promotion Cui and Hui 2011</td>
<td>Website Optimization through coding and enhancing website content</td>
<td>It entails the development of new search engines as competition becomes tougher in future</td>
</tr>
<tr>
<td>4</td>
<td>The application of search engine optimization for internet marketing An example of the motel web sites H. L. Hsing, H. J. Chen, L. M. Huang and H. H. Yi, 2011</td>
<td>Promoting Internet marketing by employing search engine optimization technique</td>
<td>Only for Internet Marketing environment</td>
</tr>
<tr>
<td>5</td>
<td>A web server design using search engine optimization techniques for web intelligence for small organizations P. T. Chung, S. H. Chung and C. K. Hui 2012</td>
<td>Use of Web Intelligence and SEO Techniques</td>
<td>This Technique can only be used for small organization</td>
</tr>
<tr>
<td>6</td>
<td>A Study on Tactics for Corporate Website Development Aiming at Search Engine Optimization M. Yunfeng, 2010</td>
<td>Exploitation of SEO Techniques</td>
<td>getting higher rank in natural search results</td>
</tr>
</tbody>
</table>

Yunfeng [6] made use of SEO algorithms and techniques for the development of websites. Web design methods and Internet marketing are covered by the research canvas. An essential component of our everyday life now includes the search engine. The basic elements of websites are created by using are trieval principle of SEO. Therefore, the rank of websites is greater in search engine in user search results. The research emphasizes using the specific SEO algorithms and methods such as page rank algorithm and Hilltop algorithm. The author stresses upon using certain tactics for web development which include contents of website tactics, domain and hostname tactics, and links and keyword tactics. The significant approach of tactics is based upon the development of websites by aiming SEO and getting higher rank in natural search results.