World Wide Web (WWW or Web), a repository of hyperlinked documents spread over the Internet, has become very popular in last few decades and has offered a powerful platform to disseminate and retrieve information. Nowadays, it has been known as the largest and most frequently accessed public information repository consisting of a variety of data types. The scale of the WWW is immense, consisting of at least ten billion publicly visible web documents distributed on millions of servers world-wide. In fact, it has expanded by about 2000% since its inception and is doubling in size every six to ten months. However, the lack of rigid structure and the information explosion over the WWW has posed tremendous challenges towards Information Retrieval requirements of today’s world. Users are often facing the problems of information overload and drowning while searching on the Web.

To automate the process of searching for relevant information from the WWW, the development of newer and powerful information retrieval tools such as Web search engines assume importance. People increasingly use search engines to fulfill a wide variety of their needs. But, they usually suffer from difficulties in finding desirable and accurate information due to low precision and recall caused by many present day search engines. For example, if a user searches for the desired information on a search engine such as Google, it will provide not only the web pages related to the query topic but also a large amount of irrelevant information, which user would unnecessarily sift for finding the exactly needed information. Thus, relevant information retrieval from WWW is the most challenging task due to many reasons: relevance is a subjective and time-varying concept with respect to users; Web is heterogeneous, highly dynamic and growing at a staggering rate; users have different expectations and goals such as informative, navigational or transactional, and moreover they often compose short and vague queries.

The informative and predictive web knowledge hidden in large repositories of data on the Web, if extracted through the application of web mining techniques, could enhance the performance of search engines by optimizing their search space. Web Mining aims to discover hidden web patterns from semi-structured or heterogeneous data representations such as textual, hyperlink structure and usage information, to improve the quality of services
offered by modern search engines. Page ranking, Query expansion, recommendation and personalization are the kinds of services provided with the aid of Web mining techniques.

Although modern search engines have been employing state of art web mining techniques to optimize their search process, one of the biggest problems which they face is their inability to find the exact information users are looking for and that too in an easy browsable manner. Several researches are available in the literature that work to resolve only one or few related problems of search engines, but not any unified technique has been reported that simultaneously can resolve most of the challenges faced by search engines.

In this thesis, a novel framework of search engine called “P&P (Pre-Mining and Post-Mining based) Framework” is being proposed. It can optimize multiple processes such as crawling, indexing and query processing of search engines to improve the quality of search results corresponding to a user query. Moreover, it provides relevant document retrieval and accomplishes the following two tasks:

1. Discovery of hidden predictive knowledge from WWW.
2. Utilizing the discovered knowledge for retrieval of relevant information.

Since the proposed framework involves many tasks, a suite of techniques distributed into two phases has been designed to separately handle the back-end and the front-end activities of search engines. The P & P framework basically works in following two optimization phases:

1. Pre-Mining Phase
2. Post-Mining Phase

A couple of techniques are being developed for each phase which are meant to mine the existing web resources to discover knowledge in the form of patterns that can be utilized in optimizing various tasks carried out by the search engine and thus, resulting in relevant page retrieval from the web.