Chapter 4

THE P&P FRAMEWORK FOR KNOWLEDGE DISCOVERY FROM WWW

4.1 INTRODUCTION

A Significant number of web users are often facing the problems of information overload and drowning due to the exponential growth of information on the WWW. Particularly, they usually suffer from the 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 search engines such as Google, they provide not only the web pages related to the query topic, but also a large amount of irrelevant 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 web knowledge hidden in large repositories of WWW, if extracted through different web mining techniques [11], could enhance the performance of search engines by optimizing their search space. Several researches are available in the literature that resolve only one or few related issues as described in Section 3.4, but not any unified technique has been reported that simultaneously can resolve most of these issues. In this chapter, a unified technique called “A novel P&P (Pre-Mining and Post-Mining based) Framework of Search Engine for Knowledge Discovery” [122, 123] is being proposed. It can optimize multiple processes such as crawling, indexing and query processing of search engines. 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.

The framework along with its component modules is explained in the next section.

**4.2 THE PROPOSED P & P FRAMEWORK OF SEARCH ENGINE**

The proposed framework of search engine is abbreviated as “P & P”. Here the two P’s are representing the proposed “Pre-Mining” and “Post-Mining” tasks performed by a search engine as shown in Fig. 4.1. These two unique phases differentiate this framework from other existing search engine architectures.

![Diagram of P & P framework](image)

**Fig. 4.1 The P & P Annotation**

The framework, shown in Fig. 4.2, is meant to optimize the information retrieval process of existing search engines by the application of web mining techniques at the front-end and back-end processes of search engines. Here, the task of applying mining techniques at the back-end is termed as Pre-Mining and at the front-end as Post-Mining. By applying this unified approach of amalgam of various mining techniques at different functional levels, a search engine is swayed to produce relevant pages in an organized manner as compared to other general purpose search engines.

It may be noted from Fig. 4.2, that at the back-end of the framework, following tasks are performed:

1. *Crawler* downloads the pages from the WWW, parses them and routes the parsed pages to the *indexer module*. In the general scenario, crawlers crawl up to a certain depth of the web sites or follow robot exclusion protocols to bring the web pages, but in this way, certain important pages remain un-crawled. However, in
the proposed work, the crawler is made intelligent so as to detect important and relevant pages from the web by means of mining the web graph and usage logs maintained by search engine, thereby, routing the relevant content to the indexer.

Fig. 4.2 The P & P Framework for Knowledge Discovery from WWW

2. The indexer creates the index on the basis of different terms present in the pages. Indexer module is also optimized by making it to consider semantics of different
terms and users’ usage patterns, which are extracted by mining the existing web resources (e.g. thesaurus, semantic or dictionary based sites) and usage logs of search engines respectively.

3. Besides serving general purpose queries, the indexer is made to consider domain specific queries also, which are generally meant for structured web content i.e. XML web pages. The XML index structures are also constructed at the back end and optimized by considering users’ usage patterns.

The index structures constructed thereof are incrementally refreshed so as to consider the new and updated content on the web. On the other side of the framework i.e. at the front end, the following activities happen:

1. User submits his query on the query interface of the search engine for retrieval of relevant pages. The query is sent to query processor, which in turn matches query terms with the index structures and retrieves a set of matched pages.

2. Before presenting query results to the user, a post-mining module which detects duplicate pages from the matched ones, filters out the results so that unique pages are presented to the user.

3. A rank and cluster generator module works on the basis of Web Structure, Content and Usage mining and organizes the results in the form of clusters, wherein each cluster presents the pages in an ordered list. This organization speeds up searching for the desired pages in the whole search space.

4. Rank optimization module works to improve the rank scores of pages based on history of their accessing patterns. This module reorganizes the result list so that most relevant pages move upward in the list i.e. in the user’s look around area.

5. Query Recommendation (or expansion) module gives prospect to the user to refine his query. It suggests similar queries on the basis of different measures such as keyword matching, clicked pages and click counts etc.

It can be observed that the proposed framework optimizes the working of a search engine from crawling till query processing with a view to provide relevant pages to the user in an efficient organization.
The following section describes the techniques proposed for the Pre-Mining and the Post-Mining phases of the P & P framework.

### 4.3 PHASES OF P & P FRAMEWORK

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. The P & P framework basically works in following two optimization phases:

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

Each phase consists of a number of techniques 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.

Fig. 4.3 represents taxonomy of the proposed techniques for the pre-mining and the post-mining phases. General search engines can employ a mixture of these techniques depending on their architectures.

![Fig. 4.3 Taxonomy of Proposed Pre and Post-Mining Techniques](image.png)
The following subsections describe various Pre-Mining and Post-Mining techniques.

4.3.1 PROPOSED PRE-MINING TECHNIQUES

The Pre-Mining Phase is an integration of different modular components generally meant for optimization of back-end processes carried out by search engine. The following Pre-Mining Techniques have been proposed for the pre-mining phase:

1. Optimization of Crawling and Indexing Process based on user feedbacks [122].
2. Construction of a Semantic Search Layer over keyword based Indexing using external web resources [123, 124].
3. Construction and Incremental Optimization of Index for XML pages based on Query workloads [125].

4.3.2 PROPOSED POST-MINING TECHNIQUES

The Post-Mining Phase constitutes different modular components for the optimization of front-end processes carried out by search engine. The following Post-Mining Techniques have been proposed for the pre-mining phase:

1. Search Result Optimization (or filtering) with elimination of duplicate or near-duplicate pages by mining the query logs [126, 127].
2. Search Result Organization by Clustering and Ranking the pages based on their content and hyperlink structure [128, 129, 130, 131, 132].
3. Search Result Organization by ranking the pages based on visits of links in pages [132, 133].
4. Rank Optimization (or improvement) based on sequential order of page accesses derived from query logs [24, 134, 135, 136].
5. Query Recommendation (or expansion) by mining the query logs [135].

4.4 KNOWLEDGE DISCOVERY BY P & P FRAMEWORK

Each Pre-Mining and Post-Mining technique mines hidden predictive patterns i.e. knowledge from web resources (web pages, web graphs, query logs, online databases...
etc.) by the application of various web mining and information extraction approaches and uses this knowledge to optimize different tasks performed by search engines as shown in Fig. 4.4.

![Fig. 4.4 Knowledge Discovery via P & P Framework](image)

It can be observed from the figure that each pre-mining technique maintains a local knowledgebase to store the discovered information and regularly consults this database to improve the efficiency of crawling and indexing process. Each post-mining technique does the same but optimizes the query processing carried out by a search engine. The optimized modules, in this way, manage large amount of web data in a concise and efficient manner so that when user submits his search query, the relevant pages are returned.

A detailed discussion on Pre-Mining and Post-mining techniques is given in the next two chapters.