CHAPTER 6

6 WEBSITE DEVELOPMENT AND METHODOLOGY USED

6.1 SYSTEM DEVELOPMENT

6.1.1 System Architecture

The proposed framework has been designed using ASP.NET for designing the front end. For designing the back end MYSQL has been used. In the current study an E-learning website has been set up for MCA students. Initially the website carries data for only MCA I Semester students of GGSIP University. In the future work further modifications to the website would be done. The modifications would include updating the study material, enhancing the structure of the website, providing log features etc. The test site contains modules for content extraction, gathering feedback, tutorials and other study material. Study Material Tab provides study material in the form of PowerPoint Presentations, Notes, and Previous Year Papers etc. for MCA I Semester students. The process of GUI development has been shown in Figure 6.1.
6.1.1.1 User Interface Component

The User Interface is an integral component in any of the software that is being developed. The user interface is the only way by which the user can interact with the system. It helps the user in developing the image of any business owning the website. It acts as a means for effective communication with the WWW. So, the User interface or UI is a platform that delivers variety of information to the users thereby providing comfort and ease. Hence it is said that the UI should be designed in a highly effective and efficient manner and must never be complicated. The UI designed has been shown in Figure 6.2.
6.1.1.2 Web page

A web page is nothing but a document available on the World Wide Web. It contains information related to various topics. It can be accessed via a Uniform Resource Locator (URL). They are device independent and can be accessed using either PC, laptop or mobile device. The web pages are written usually using HTML or DHTML. Web pages provide features of hypertext which included features like navigation bars and links to other pages in the website. Collection of pages is known as website. They can be interlinked with java script and other scripting languages to provide better functionality. Web pages are a part of client server model and can be retrieved from a local computer or from a remote web server. The web page has been shown in Figure 6.3
6.1.1.3 Add-in manager

This is a panel which has been inserted at the back end which helps in adding resources to our website. The panel consists of the cookie word and the path of the resource that has to be added. As soon as the entry is completed the data is updated in the database and hence is visible on the website. The user interface of the Add in manager is shown below in Figure 6.4.
Web Personalization and Recommendation module

Web Personalization is in great demand and is helping a lot website owner in attracting customers to their site. Users seem to be satisfied if they don’t get redundant content on the website. More satisfied users’ means more loyalty. This strategy is being implemented in almost all realms on the internet. In our work, the web personalization and recommendation module is also responsible for the same. This module comprises majorly of three sub modules: User Profiling, Content Extraction and Suggestions Panel. It extracts the details of the user via user profile module. The user profile module extracts data from the web access logs generated when one clicks on the website. The cookie data of the respective users is taken and combined with the content data. The content data is gathered from the content extraction panel. Finally the user specific result is shown to the user in the suggestion panel keeping in mind his current as well as background searches.

User Profiling

For personalizing the needs and demands of the customer, the personal data needs to be gathered. User Profiling could be of four types namely Local, Roaming, Mandatory.
and Temporary. Local profile is created when we first access the computer. Roaming profile is created with the help of system administrator. Mandatory profiling is a part of roaming for specifying particular settings for an individual. Temporary profiling is created whenever an error condition is being raised.

User profiles can also be accessed from the web access logs of a website. The log file contains a variety of information including the IP address, date and time of access, URL accessed etc. This type of information can be useful in accessing the list of frequent visitors to a website. In the proposed work the user profiling information has been carried out using cookies. Every user is issued a cookie id and this information along with the page accessed and date and time of last visit is recorded and stored in the database. A sample of it has been provided in Figure 6.5.

![Figure 6.5 Snapshot of Database](image)

**6.1.1.6 Content extraction**

Content extraction tab is a major component where data from various websites is gathered. The Content extraction panel requires a keyword be inserted which has been referred to as cookie word. After that the required URL needs to be entered from where the contents have to be restored. Once the URL has been entered, the next step is to
enter the keyword or the search string that is searched from the designated URL. A button named “extract content” has been provided. As soon as the user clicks on the button, relevant content from the website is scanned and extracted. All the URL’s that match the search string from the respective website which has been entered are extracted and are shown on the content extraction panel. A checkbox named save results as shown in Figure 6.6 has also been provided which gives the user an option of saving the data that has been extracted from the website. As soon as the checkbox is selected the list of URL’s obtained is added to the database of the cookie word. Later these URL’s are available as suggestions to the user on the main panel of the website. So the content extraction panel extracts contents from other website and redirects the user to other websites if good quality content is available on the other sites. The content extraction panel has been shown clearly in Figure 6.6.

![Content Extraction Panel](image)

**Figure 6.6 Snapshot of the Content Extraction Panel**

**6.1.1.7 Suggestion Panel**

A Suggestion Panel has been added in the website which provides suggestions related to the cookie words available on the website. Initially when the user logs on to the website for the first time the major cookie words are available in the suggestions panel. But later depending upon the data surfed by the user, suggestions are updated. Either the updates could be in the form of data from our own website or it could be the
resultant of content extraction panel (i.e. content from other websites). Suggestion panel has been shown in Figure 6.7 as given below.

![Screenshot of Suggestion Panel](image.png)

**Figure 6.7 Snapshot of the suggestion Panel**

### 6.1.1.8 Feedback module

Feedback module is responsible for keeping a track of how far the users who visited the website were satisfied. It has been incorporated to check the satisfaction level of the users. The feedback form has been developed using Google docs (a utility of Google). A Google doc is a free web based web based utility offered through Google within Google drive. It helps in creating and editing of online documents along with collaboration with users in real environment. The form asks the user for his cookie ID. Cookie ID is generated as soon as the user visits the website for the first time. Along with the cookie ID the time of last visit is also displayed to the user. The cookie ID is recorded so that the satisfaction of each user visiting the website is recorded. Next the resource title accessed by the user is asked. The user is next asked as to how much satisfied one was from the set of responses that were generated or the resources which were provided by the website. In the end the user is asked on whether one would like to visit the website again or not. The feedback obtained from each user can be utilized
WEBSITE DEVELOPMENT AND METHODOLOGY USED

in enhancing the quality of the website thereby attracting more and more people towards our website. These responses were extracted in the form of excel sheets and mining algorithms were applied by utilizing MS Excel Mining utility (add in feature of MS Excel). A sample of the feedback form is shown in Figure 6.8

![Feedback Form](image)

**Figure 6.8 Snapshot of the Feedback form**

### 6.2 METHODOLOGY

It highlights the methodology adopted for system development. Detailed discussion is provided starting from the collection of data till the last step which aims at analyzing the user’s satisfaction level.

#### 6.2.1 Tools and technologies used

**6.2.1.1 Visual studio 2012**

An Integrated development environment (IDE) from Microsoft is used for development of programs, website, web applications and services. It utilizes development platforms
like Windows API, Windows Forms, Windows Presentation Foundation, Windows Store and Microsoft Silverlight. It supports a variety of programming languages. Built in languages offered by Visual studio include C, C++, VB.NET, C# and F#. Languages like XML/XSLT, HTML/XHTML, JavaScript and CSS are also supported by it. The application has been developed using Visual studio 2012 and programmed using ASP.net.

6.2.1.2 MySQL

It is the most widely used open source relational database management system (RDBMS). It has also been referred to as the second most widely used RDBMS. Applications such as TYPO3, MODx, Joomla, WordPress, phpBB, MyBB, Drupal are widely using MySQL as a solution. It is a relational data base management system and contains no GUI tools for administration or managing the data in the data bases. Command line tools may be included for creating and managing the databases. It is offered under two editions namely the open source MySQL Community Server and the proprietary Enterprise Server.

The database designing has been done using MySQL in the current work. Cookiemanager database has been created which comprises of two tables cookiedb and cookies. The cookiedb contains the following attributes S.No., ID, cookieword and LastVisit. A snapshot of the data recorded has been shown in Figure 6.9.
Another table which was created is the cookies table. The cookies table records cookie word, search URL, search word, path, the site rank and match count attributes. A snapshot of this table is provided in Figure 6.10.
6.2.1.3 Microsoft SQL Server 2008 Data Mining Add-Ins for Office 2007

The above utility is present as an Add in feature in MS Excel 2007. It offers set of wizards and tools which help in extracting useful information from the data extracted. It helps in deriving trends and patterns in data, visualize the patterns obtained thereby generating rich and colorful summaries for business analytics. It can be easily integrated by adding the data mining add in which is available freely and integrating it with SQL Server. It also provides users with powerful, easy-to-use tools for working with mining models in Analysis Services.

Analysis of data has been carried out using the above stated tool. In our current work clustering; data mining technique has been used for carrying out analysis which is thoroughly discussed in the next chapter.
6.2.2 Data Collection

Once a layout has been designed, collecting the data is the first step. In the current work data in both implicit and explicit form was taken. This is also referred to as primary and secondary data. Primary data can be gathered directly by interacting with users. On the other hand, secondary data is generated.

For gathering the explicit or primary data, users were asked to share their feedback in the form of e-questionnaires that were designed using Google Docs. These e-questionnaires were posted on the website and were filled by the user. The questionnaire helped in getting responses from the user in an effective manner. This utility provided the data in easy to use format. Every user visiting the website was asked to fill in this feedback form. After collecting this data analysis of the data was carried out. The results obtained helped in attracting more and more people towards the website.

Implicit or the secondary data was gathered via the web access logs that were generated automatically by the Web and Application servers. Web access logs contain all the information related to a particular user like his I.P. address, date and time of access, URL hit by the user etc. Every time a user clicks a website his data gets recorded in the web access logs. During the research work the information related to a particular user was taken from the web access logs. The web access logs contain huge amount of information. However this data being unstructured cannot be readily used for analysis. It needs to be passed through a preprocessing stage before making it suitable for analysis.

But in the current work the cookie details of every user recorded in the web access logs was used. Every new user has been issued a cookie ID. The cookie ID contains a unique ID number and date and time of his last access on website. The content extracted from various other websites can also be considered as implicit data.

The sample data collected is shown in Figure 6.11 and 6.12.
6.3 SEQUENTIAL PATTERN MINING

Sequential pattern refers to an ordered set of pages which is said to satisfy the support given and is also maximal at the same time. In context of web mining, support refers to the percentage of customers carrying the pattern. Sequential pattern mining can be classified into Apriori and Pattern Growth based. Apriori is further categorized into generalized sequential pattern mining, sequential pattern mining with regular expression constraints, sequential pattern discovery using equivalence classes and...
sequential pattern mining. Pattern growth based has been classified into prefix projected; frequent pattern projected and weighted sequential pattern mining. Sequential pattern mining is considered to be a part of Association Rule Mining (ARM). ARM helps in deciphering interesting associations or correlation amongst the data items present in the data set. It is considered to be an integral and important component of data mining and is being extensively used by data mining community. Market Basket Analysis (MBA) is stated as a classical application of Association rule mining. It finds out the relationships between items purchased in supermarket. Association rules are of the form X->Y. Here X and Y refer to the data items present in the data set. It helps in finding out attributes value condition that occurs frequently together in a dataset. Association rule mining is not only applicable for supermarket chain. But rather it has its application in areas such as telecommunication networks, market risk management, inventory control, agriculture sector. The major aim of this technique is only to detect relationships amongst variables in huge data sets. It helps a lot in boosting the performance of businesses. But unfortunately it does not take into account the order of transactions which can be very important in some applications like Web usage mining wherein navigational patterns of the users have to be deciphered. Here the role of Sequential pattern mining comes. Given a set of input data sequences, sequential pattern mining finds out all the sequences that have a user-specified minimum support.

In context of Sequential pattern mining constraints are considered very essential. In [94] it has been stated that constraints play a vital role in majority of data mining applications for determining effectiveness and efficiency considerations. constraint-based sequential pattern mining finds complete set of sequential patterns satisfying a given constraint C. Seven categories of constraints were discussed [94] namely Item,
Sequential pattern mining can be applied in several areas including Healthcare, Education, Web usage mining, Text mining, Bio informatics, Tele communications, Intrusion detection etc.

In the proposed work Sequential pattern mining has been used using regular expressions as a constraint. The proposed algorithm has been presented in the next section.

6.4 PROPOSED ALGORITHM

Sequential Pattern Mining using Regular Expressions Constraints has been utilized for carrying out the proposed work. Pattern Mining is based on association rules, which helps in the software to suggest contents to different users depending on their performance behavior. The suggestions panel mines the data depending on association rules using regular expression and suggests contents within a subtopic to a user. The code snippet for the same is given below in Figure 6.13.

```csharp
List<LinkItem> list = new List<LinkItem>();

// 1.
// Find all matches in file.
MatchCollection m1 = Regex.Matches(file, @"(<a.*?>(.*?)</a>)", RegexOptions.Singleline);

// 2.
// Loop over each match.
foreach (Match m in m1)
{
    string value = m.Groups[0].Value;
    LinkItem i = new LinkItem();

    // 3.
    // Get href attribute.
    Match m2 = Regex.Match(value, @"href=""(.*)""\","", RegexOptions.Singleline);

    if (m2.Success)
    {
```
Figure 6.13 Code Snippet for Sequential Pattern Mining using regular expression constraints

The algorithm is used with MatchCollection Data Structure in ASP.NET where in we have used the matches method of Regex Class and provided the regular expression [@"(<a.*?>\(.*?)</a>)" ] for matching required search terms from the collection of web page data.

Regex class is used for representing a read only regular expression. It is contained in System.Text.RegularExpressions namespace [95]. The Regex class comprises of a lot of methods which have been defined precisely in the MSDN library. In the current work the Matches method of this class was used. The Matches method searches the specified input string for all occurrences of a specified regular expression. The static Matches methods are equivalent to constructing a Regex object with the specified regular expression pattern and calling the instance method Matches. The Matches method uses lazy evaluation to populate the returned MatchCollection object. The Matches(String, String) method is similar to the Match(String, String) method, except that it returns information about all the matches found in the input string, instead of a single match. It accepts parameter as an input which consists of regular expression language elements that symbolically describe the string to match [95].