# CHAPTER – 7

CONCLUSIONS AND FUTURE WORK

<table>
<thead>
<tr>
<th>S. No</th>
<th>Contents</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Present Work</td>
<td>122</td>
</tr>
<tr>
<td>7.2</td>
<td>Future Enhancements</td>
<td>124</td>
</tr>
</tbody>
</table>
CHAPTER - 7

CONCLUSIONS AND FUTURE WORK

7.1 Present Work

The DEEP WEB is usually defined as the content on the Web not accessible through a search on general search engines. This content is sometimes also referred to as the HIDDEN or INVISIBLE WEB. The Web is a complex entity that contains information from a variety of source types and includes an evolving mix of different file types and media. It is much more than static, self-contained Web pages. In fact, the part of the Web that is not static, and is served dynamically "on the fly," is far larger than the static documents that many associate with the Web.

The concept of the deep Web is becoming more complex as search engines have found ways to integrate deep Web content into their central search function. This includes everything from airline flights to news to stock quotations to addresses to maps to activities on Facebook accounts. In the screenshot below, notice the various deep Web sources offered by Google, including images, maps, news, video, shopping, scholarly content, blogs, and so on. However, even a search engine as far-reaching as Google provides access to only a very small part of the deep Web.

In this thesis a work is done on the following problems that occur in the deep web

- In the hidden web the traditional web crawlers are working only on http site it will not look at the secure Http site.
• Traditional algorithms are not effective
• New techniques should be improved for the best performance of the crawlers
• New design of specific search crawling is to be generated for getting the better results

The above points are referring to the problems of the deep web. For the above problems different types of solutions are presented in this thesis. A Survey is presented that will explain all the traditional web crawlers. The survey also portrays the need of newer methods of web crawler as the internet is never a same and it changes its architecture dynamically, a proven model of web crawler, which is capable of pulling the prominent required information from several hidden part of web.

Next A WEB CRAWLER MODEL is presented for crawling the secure pages i.e. HTTPS pages. Traditional crawlers will look only the HTTP pages but this model will help to search and crawl all the secure HTTP pages.

Next a framework is designed for deep web crawler. It is used for to crawl the all the pages. Improvised Genetic Crawler is used in this model. Precision; Recall metrics are giving better results by using this model.

Focus crawlers are used for searching specific string over the internet. In this improved specific crawler is designed. The improved specific Crawler gives better results than the earlier specific crawlers. This approach takes the minimum time. Improved Specific Crawler also gives the Non-Related pages also.
7.2 Future Enhancements

Though all the specified and designed methods are giving better results, it is needed to extend these approaches. As the crawlers are designed for secure web pages it will search HTTPS pages but this work can be extended for the e-commerce applications and need to look how this work will give better results.

All the works are presented in this thesis can be extended for further research.