CHAPTER 5: RESULTS AND CONCLUSION

In this chapter the implementation of the proposed Semantics and Learning based Focused Crawler is given. The proposed crawler is compared with the Breadth First Crawler, VSM based Best First Crawler, Naive Bayes Breadth First Crawler, and Naive Bayes Best First Crawler on the basis of following three criterion:

a. Automated precision value.

b. Average time taken to retrieve 1000 pages from the Web.

c. Empirical validation results for precision value, recall value and F-measure.

5.1 Implementation

The proposed Semantics and Learning based Focused Crawler is implemented in Java using a Windows 7 setup on an Intel(R) Core(TM) i5 Processor 3.20 GHz and 4.00 GB of RAM, and using MySQL as the backend database. Crawler performance is generally measured in terms of precision rate, recall rate and F-measure.

Precision is the probability that a (randomly selected) retrieved document is relevant or is the average probability of relevant retrieval.

Recall is the probability that a (randomly selected) relevant document is retrieved in a search or is the average probability of complete retrieval.
The random selection refers to a uniform distribution over the appropriate pool of documents; i.e. by randomly selected retrieved document, we mean selecting a document from the set of retrieved documents in a random fashion.

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score. This is also known as the $F_1$ measure, because recall and precision are evenly weighted.

The pages which are having a VSM Weight Vector (VWV) or Semantic VSM Weight Vector (SVWV) score greater than a threshold value are considered as relevant and others as non relevant to the crawler domain.

An automated process for gathering seed pages is used. Open Directory Project (ODP) provides the categorical collection of URLs that are manually edited and not biased by any commercial user. From here we can find individual categories link. One can find maximum number of URLs related to the concerned topic by just browsing the http://www.dmoz.org from root to the topic of interest. For analysis and other purposes http://www.dmoz.org also provides dump for the whole database, which can be used as according to the needs of the individual. This dump was downloaded and its contents were extracted using to MySQL using PERL (Practical Extraction and Report Language) scripts.
The URLs belonging to the categories which are pointing to "Animation", "Cricket", "Science" and "Computer" at the last level were retrieved. A total of 323, 199, 250 and 563 such URLs were found through a module written in Java for each of the "Animation", "Cricket", "Science" and "Computer" domain respectively.

The set of these URLs acted as the initial seed pages for Breadth First Crawler, VSM based Best First Crawler, Naive Bayes Breadth First Crawler, Naive Bayes Best First Crawler, and the first phase of the proposed Semantics and Learning based Focused Crawler.

5.2 Automated Precision Results

Automated precision value designates all the pages obtaining similarity score greater than some threshold value as relevant to the domain and all other pages as irrelevant. The precision graphs by setting different similarity value thresholds for different topic domains for different crawlers are shown in Fig. 5.1 to Fig. 5.8.

From the results for automated precision it is evident to say that the proposed semantics and learning based focused crawler outperforms all other including breadth first crawler, Vector Space Model based best first crawler, Naïve Bayes classification based breadth first crawler and Naïve Bayes classification based best first crawler in terms of retrieving more number of pages having similarity score greater than the similarity threshold.
Results and Conclusion

Fig. 5.1: Automated precision graph for "Animation" domain and Similarity Score $\geq 1.0$

Fig. 5.2: Automated precision graph for "Animation" domain and Similarity Score $\geq 2.0$
Results and Conclusion

Fig. 5.3: Automated precision graph for "Computer" domain and Similarity Score >= 1.0

Fig. 5.4: Automated precision graph for "Computer" domain and Similarity Score >= 2.0
Results and Conclusion

Fig. 5.5: Automated precision graph for "Science" domain and Similarity Score>= 1.0

Fig. 5.6: Automated precision graph for "Science" domain and Similarity Score>= 2.0
Results and Conclusion

Naive Bayes Breadth First Naive Bayes Vector Space Proposed
Breadth First Crawler Best First Space Model based Crawler Crawler Crawler Crawler
Crawler Crawler Best First Phase-1 Phase-2 Phase-3 Phase-4

Fig. 5.7: Automated precision graph for "Cricket" domain and Similarity Score >= 1.0

Fig. 5.8: Automated precision graph for "Cricket" domain and Similarity Score >= 2.0
Fig. 5.1 to Fig. 5.8 shows that the consecutive runs of the proposed crawler outperforms all other crawlers for retrieving pages related to all the four domains. Also precision value shows an upward trend with increase in number of consecutive crawls of the proposed Semantic and Learning based Focused Crawler.

5.3 Time Performance Results

In all the crawlers the page is to be fetched from the Web and is to be pre-processed. The proposed approach consumes a constant amount of time for using WordNet and to decide upon the number of pre-runs for the final phase. But all that time consumed could be ignored while finding growth rate of the crawler’s performance.

![Average time performance graph](image)

Fig. 5.9: Average time performance graph

However the time comparison demands some perfect system conditions like network traffic, internet speed, server response time etc. The average time taken by each crawling technique to retrieve a group of
Results and Conclusion

1000 pages is determined (in seconds), and a graph is plotted as shown in Fig. 5.9.

5.4 Empirical Validation Results

Empirical evidence (also empirical data, sense experience, empirical knowledge, posteriori) is a source of knowledge acquired by means of observation or experimentation. Empirical evidence is information that justifies a belief in the truth or falsity of an empirical claim. For validating the proposed system using expert evaluation i.e. empirical validation, 100 documents from the end results for the "Computer" domain for each crawler were selected randomly. The randomly selected set from each crawler was given to 10 experts for marking the documents which they feel as relevant or irrelevant. The experts' observations after finding the average from 10 experts and by considering the page as irrelevant if and only if it is marked as irrelevant by all the experts are given in Table 5.1. The precision value, recall value and F-measure are calculated using Equation 1.3, Equation 1.4 and Equation 1.5 respectively. From the table it is evident to say that the proposed crawler is having the highest precision, recall and F-measure values amongst the all. To avoid the effect of biased selection of pages for the purpose of expert evaluation the whole process is reiterated with five different sets of 100 random pages from each crawler's end results for the same domain. The average precision value, average recall value and average F-value were calculated and plotted as graphs shown in Fig. 5.10, Fig. 5.11, and Fig. 5.12 respectively.
## Results and Conclusion

Table 5.1: Precision, Recall and F-measure using expert evaluation

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Name of the crawling approach</th>
<th>Sample size</th>
<th>Can't say</th>
<th>Sample size left</th>
<th>No. of relevant pages</th>
<th>No. of irrelevant pages</th>
<th>No. of relevant pages</th>
<th>No. of irrelevant pages</th>
<th>No. of pages declared relevant by the crawler but irrelevant by the expert</th>
<th>No. of pages declared irrelevant by the crawler but relevant by the expert</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Naive Bayes Best First Search Crawler</td>
<td>100</td>
<td>20</td>
<td>80</td>
<td>41</td>
<td>39</td>
<td>41</td>
<td>39</td>
<td>6</td>
<td>6</td>
<td>.43</td>
<td>.85</td>
<td>0.57</td>
</tr>
<tr>
<td>2.</td>
<td>Breadth First Search Crawler</td>
<td>100</td>
<td>17</td>
<td>83</td>
<td>40</td>
<td>43</td>
<td>40</td>
<td>40</td>
<td>17</td>
<td>02</td>
<td>.47</td>
<td>.69</td>
<td>0.56</td>
</tr>
<tr>
<td>3.</td>
<td>VSM Best First Search Crawler</td>
<td>100</td>
<td>02</td>
<td>98</td>
<td>45</td>
<td>53</td>
<td>40</td>
<td>58</td>
<td>11</td>
<td>16</td>
<td>.50</td>
<td>.72</td>
<td>0.59</td>
</tr>
<tr>
<td>4.</td>
<td>Naive Bayes Best First Search Crawler</td>
<td>100</td>
<td>20</td>
<td>80</td>
<td>26</td>
<td>54</td>
<td>45</td>
<td>35</td>
<td>19</td>
<td>0</td>
<td>.32</td>
<td>.57</td>
<td>0.40</td>
</tr>
<tr>
<td>5.</td>
<td>Proposed Crawler</td>
<td>100</td>
<td>15</td>
<td>85</td>
<td>57</td>
<td>28</td>
<td>51</td>
<td>34</td>
<td>01</td>
<td>07</td>
<td>.59</td>
<td>.98</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 5.1: Precision, Recall and F-measure using expert evaluation
Results and Conclusion

Fig. 5.10: Empirical precision graph from expert evaluation

Fig. 5.11: Empirical recall graph from expert evaluation
Fig. 5.12: Empirical F-measure graph from expert evaluation

The empirical results show that the proposed crawler outperforms all other in terms of empirical precision value, empirical recall value and empirical F-measure.

5.5 Conclusion and Future Work

Generic crawlers and search engines are like public libraries; they try to cater everyone, and do not specialize in specific areas. Serious Web users are increasingly feeling a need for specialized and filtered libraries, where they can explore their interests in depth. Unlike public libraries, Web libraries have little excuse not to specialize, because it is just a matter of locating and linking to resources. Focused crawlers can be seen as the
tool for generation of these kinds of specialized libraries. Focused crawling has become an interesting alternative to the current Web search tools. It concerns the development of particular crawlers able to seek out and collect subsets of Web pages related to given topics.

The focused crawler also gives better user experience as it tries to give results which are more relevant to user’s information needs, thus leading to user satisfaction which is one of the parameters measuring success of information retrieval system. The quantity of data on Web is increasing exponentially so it is critical for information retrieval system to be very efficient while handling user queries which normally require high precision, thus focused crawler plays a crucial role over here. Many domains may be benefited from the focused crawlers, such as vertical portals and personalized search systems, which provide interesting information to communities of users.

In this work focused Web crawler based upon WordNet semantics, Vector Space Model (VSM) and hub score learning is proposed. Several variants of focused crawler including the proposed crawler were implemented. Crawling results for all the implemented crawlers were studied. The results show that the proposed crawler outperforms the others in terms of the precision rate, recall rate and F-measure, and also outperform all but Naive Bayes breadth first crawler, which produces the worst precision among all the competitors, in terms of average time taken for retrieving 1000 pages from the Web.
Results and Conclusion

In future the proposed crawler can be extended by considering the case of hidden Web. The classifiers like neural network, support vector machines etc. needs to be tested for possible enhancement to the proposed crawler’s performance.