CHAPTER 7

PERFORMANCE EVALUATION

7.1 INTRODUCTION

The validity of any solution can be proved only when the comparative performance of the existing solution and the proposed solution are analysed. Proposed solution may not be a complete solution to an existing problem but it may be a direction towards an alternate solution. Refinement may be required in few areas to improve it to be the best-suited solution for the problem. This chapter evaluates the comparative performance of TAOS with crawler-based updation and semantic search process with the existing search process. It analyses the positive and negative aspects of the proposals concluding with achievements and the areas that need further improvements.

7.2 TRAFFIC ADAPTIVE OPTIMUM UPDATING SCHEME

TAOS aims to improve the freshness of the search engine repository by eliminating the needless requests of the crawler-based updation. It also aims to update the repository adapting itself to the changes in the web traffic. TAOS ensures to optimally use the network resource by choosing an optimum time to notify the delta updates to search engine repository. This thesis considers the extensive work of Cho and Molina on page refresh policies for web crawlers to analyse the performance of TAOS. The following sections will introduce with the basic work of Cho and Molina and the results derived by them.
7.2.1 Repository freshness

In order to provide better performance, local copies of the remote data sources are maintained by the search engines. In this case, a part of the local copy may get out-of-date because changes at the sources are not immediately propagated to the local copy. Therefore the author Cho and Molina proposes few refresh policies that maximizes the freshness of the local copy. They propose synchronization policy taking into account how often a page changes and how important the changes are, and makes an appropriate synchronization decision. The drawbacks of the work were discussed in section 4.2. The performance of the policies is discussed in this section.

Several dimensions to the synchronization process are proposed in the paper in order to refresh a local copy of web page. The dimensions of the process include

*Synchronisation frequency* - The value decides on how frequently the local database needs to be synchronized.

*Resource allocation* – This deals with how often an individual element is synchronized. Uniform allocation policy and non-uniform allocation policy were proposed in the work. Uniform allocation policy synchronises all elements at the same rate regardless of how often they change. In non-uniform allocation policy, the elements are synchronized at different rates.

*Synchronisation order* – This decides on what order the elements in the database are synchronized. They propose fixed order, random order and purely random order. Fixed order synchronises all elements in the
database in the same order repeatedly. Random order synchronises all elements repeatedly with different synchronization order for each iteration. Purely random order selects a random element from the database and synchronises it.

*Synchronisation points* - This decides on the time window of the synchronization.

### 7.2.1.1 Discussion of results of synchronization order policies

The freshness of a local database $S$ at time $t$ is

$$F(S; t) = \frac{1}{N} \sum_{i=1}^{N} F(e_i; t)$$

which depends on the freshness of the individual local element $e_i$ at $t$. The age of the local database is

$$A(S; t) = \frac{1}{N} \sum_{i=1}^{N} A(e_i; t)$$

which depends on the age of the individual local element. The results were generated between the frequency ratio and their freshness.

Frequency ratio $r = \lambda / f$ where $\lambda$ is the frequency at which the real world element changes and $f = (1/l)$ is the frequency at which the local element is synchronized. The graphs of comparison are given in Figures 6.1 and 6.2.
Figure 7.1 Freshness graph over \( r = \lambda / f \)

Figure 7.2 Age graph over \( r \)
The authors conclude that the fixed order policy performs best by both metrics. They also conclude that as the variability in time between visits increases, the policy gets less effective. They also argue that the fixed order policy guarantees a bound on freshness and age values while purely random policy may lead to unlimited freshness and age values. This implies the TAOS freshness needs to be compared with the freshness and age of the fixed order policy.

TAOS freshness over frequency ratio is given in Figure 7.3. The TAOS freshness depends on the popularity of the element and it varies based on the policy used.
Figure 7.3 TAOS freshness over frequency ratio
7.2.2 RESOURCE UTILIZATION

The thesis argues that web crawlers utilize network and server resources by servicing the needless requests. The resource utilized by web crawler is depicted in Fig 7.6 and Fig.7.7. A graph, which records the failed and successful searches of updating during high and normal user activity periods in given in Fig 7.4. The results are recorded based on the number of routers crossed during the transmission.

![Successful and failed searches during normal user activity period](image1)

![Successful and failed search during high user activity period](image2)

Figure 7.4 Comparison of failed and successful searches
Figure 7.5 Time taken for updation

Time taken to search and update a lexicon
Bandwidth occupied and time taken to update a lexicon during normal user activity

Figure 7.6 Bandwidth consumed for updation during normal user activity period
Bandwidth occupied and Time taken to search and update a lexicon during high user activity period

Figure 7.7 Bandwidth occupied to update during high user activity period
The bandwidth occupied by crawlers to carry out 150 searches in the testbed environment with the routers crossed is recorded in Fig. 7.6. The bandwidth consumed by the failed search is more because the number of failed searches is far higher than the successful search. The bandwidth consumed by successful and failed searches is given in Fig. 5.7 and 5.8.

Bandwidth consumed by partial upload of the updated document is based on factors of delta size, popularity and relevance of change. Percentage of bandwidth saved using TAOS given in Fig. 7.8 and 7.9 during high and normal activity period. The figures recorded that percentage of bandwidth saved due to delta transmission and that saved due to elimination of needless requests are almost equal. Bandwidth consumed during partial upload is given in Fig. 10 and Fig. 11. It depicts that the bandwidth consumed during partial upload is very less compared to the complete upload.
Figure 7.8 Percentage of bandwidth saved during normal user activity period
Figure 7.9 Percentage of bandwidth saved during high user activity period

Bandwidth used by failed searches □ Difference bandwidth

<table>
<thead>
<tr>
<th>Percentage of Bandwidth Saved</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>1</td>
</tr>
<tr>
<td>90%</td>
<td>2</td>
</tr>
<tr>
<td>80%</td>
<td>3</td>
</tr>
<tr>
<td>70%</td>
<td>4</td>
</tr>
<tr>
<td>60%</td>
<td>5</td>
</tr>
<tr>
<td>50%</td>
<td>6</td>
</tr>
<tr>
<td>40%</td>
<td>7</td>
</tr>
<tr>
<td>30%</td>
<td>8</td>
</tr>
<tr>
<td>20%</td>
<td>9</td>
</tr>
<tr>
<td>10%</td>
<td>10</td>
</tr>
<tr>
<td>0%</td>
<td>10</td>
</tr>
</tbody>
</table>
Comparison of complete and partial upload of updated document

Bandwidth for total transmission | Bandwidth for delta transmission | Number of routers

7.10 Comparison of complete and partial upload of web document during normal user activity period
Comparison of bandwidth occupied during complete and partial upload during high user activity period

Figure 7.11 Comparison of complete and partial upload during high user activity period
7.3 SEMANTIC SEARCH PROCESS

The semantic search process aims to introduce context and category throughout the search process. The context and category of the web document in the ranking process improves the precision of the search results and improves the ranking of web pages which lie deep inside the search results filtering those pages which are not relevant to the context of the user query. This thesis compares the work on personalization of search by Google and category-based search of ODP with the semantic search process results.

7.3.1 TRADITIONAL SEARCH PROCESS

The conventional search process does not include context of the search or the web document. The results produced for the query "Java" by Google and ODP is given in figure 7.12 and 7.13. The results contain web pages with varying context that may or may not be relevant to the user.
ODP categorises the web documents depending on their content. The categorization is done manually and the user can include category for their query through the user interface. The user interface provided by ODP is given in figure 7.14. Personalised search of Google maintains the search.
of the user the orders the results based on the preference of the user. The information extracted from history to order the results are not known. The impact on the relevance of the search results can be felt only after the entries of the search history is built up. In the semantic search process the relevance can be ordered in the first search results itself.

### 7.3.2 SEMANTIC SEARCH RESULTS

The semantic search process orders the results based on the context and category of the web document. The user interface screen of semantic search process and the result produced for a query is given in 7.15 and 7.16.

![Figure 7.15 User interface of semantic search process](image_url)

The context and category of the query needed for generating the user profile is received from the user or generated from the history. The search is...
narrowed based on the context and category of the query and the results are displayed.

Figure 7.16 Semantic search results for “JAVA”
The user relevancy reports of precision are given in section 6.5. Semantic search process narrows the search and improves the relevance of search results.

7.4 CONCLUSION

TAOS records effective and efficient utilization of resources. The results show that the bandwidth utilized in updating search engine repository using partial upload of the updated document has reduced by more than half of which is used during the crawler based updation. The value of reduction in bandwidth utilization differs with the degree of change. Degree of change is directly proportional to the bandwidth occupied during transfer, and hence an increase in delta size will decrease the amount of bandwidth saved. TAOS improves the resource utilization by eliminating the needless requests present in the crawler-based techniques. The results depict that the bandwidth consumed by failed searches are more than consumed by the successful searches. The autonomic computing design help in adapting to the dynamically changing environments. The notification of updates by TAOS based on various factors identifies the optimum page for notification adapting to the environments.

The precision of results demonstrated in the test results of semantic search process show improvement in user satisfaction. While there is no difference between the scores provided by novice and professional users in useful precision best precision records variations in the scores. The increasing document cutoff values recorded a comparatively less deteriorating trend in the precision as that compared in the normal search process.