Design and Development of Meta Search Engine Framework using Horizontal Partitioning Relevancy Criteria and Result Integration Factor Algorithms for Efficient Data Retrieval

Sudeepthi Govathoti, Research Scholar, Andhra University, Visakhapatnam & Assistant Professor, Department of CSE, Anurag Group of Institutions, Hyderabad, India.
M.S.Prasad Babu, Professor, Department of CS&SE, Andhra University, Visakhapatnam, India.

Abstract — In recent years World Wide Web has tremendous growth in the volume of information that makes very difficult to locate information that is relevant to user’s interest. The search engine is the most important tool used to discover information in World Wide Web, but they are not appropriate nowadays as they do not crawl the deeper web [1]. Hence the concept of Meta search engine that is built on top of traditional search engines is adapted. This paper is aimed at the design and development of Meta search engine framework using horizontal partitioning relevancy criteria and result integration factor algorithm for efficient data retrieval. The input query given by the user gets forwarded to selected member search engines namely Google, Yahoo, and Bing. The data retrieved from member search engines are fused/merged and stored in the database. The proposed Meta Search engine framework has two core operations namely pruning and result integration. The data stored in fusion database is pruned by removing the null and irrelevant data by applying data validation and proposed “Horizontal partitioning relevancy criteria” algorithm. The pruned data is given as an input to result integration that performs re-ranking by applying result integration factor algorithm for effective display of the relevant content as per user’s interest.

Index Term — Meta search engine, Fusion database, Partitioning, Page ranking, Pruning.

I. OVERVIEW OF META SEARCH ENGINE
The search engine is a web software program that searches for information on World Wide Web. It extracts relevant information and displays the list of results based on respective page ranking mechanism. The search engine results page (SERP) is the page displayed by a web search engine in response to a keyword query given by the user [2]. The top Web-based search engines are Google, Bing, and Yahoo. The amount of information now on the World Wide Web is overwhelming as a result it makes very difficult to locate information relevant to user’s interest using these search engines. In order to provide a solution to the above problem the concept of Meta search engine also termed as “Aggregator” came into existence [3]. The Meta search engine is a search tool that accepts input from the user and simultaneously redirects it to multiple search engines and produces its own results. The retrieved results from multiple search engines are integrated and presented to the user. Meta search engine enhances the user’s experience in retrieving information with less effort.

The remaining paper is arranged as follows: In Section 2 Motivation for implementing Meta search engine, the related work and proposed work is presented in Section 3 and 4. The methodology is discussed in Section 5. The results are presented in Section 6 and the paper is concluded in Section 7.

II. MOTIVATION
The motivation towards implementing Meta search engine depicts that it is able to synthesize the advantages from each member search engine, and the result integration could improve the quality of searching results. The primary factors for implementing Meta search engine are listed as follows:

a) Provides a broader overview of a topic searched.
b) It increases the web coverage.
c) Improves convenience for users.
d) It provides an effective mechanism to reach the deeper web.
e) Helps user to reduce the overhead of searching multiple search engines.
f) It provides fast and easy access to the desired search

III. RELATED WORK
domain. Dr. Daya Gupta et.al [11] combined fuzzy AHP with a genetic algorithm to get more comprehensive and optimized results. Manika Dutta et.al [12] analysed five major search engines i.e. Google, Yahoo, AltaVista, Ask and Bing based on some features. Jai Manral et.al [13] developed an intelligent Meta search engine which aggregates results from various search engines. Abdelbaki Issam et.al [14] defined a fusion method based on user profile and improvements are suggested based on background clicks to meet the specific needs of users. M.M.E Gayar et.al [15] proposed enhanced crawling algorithm to make the retrieval process becomes faster. Hossein Jadidoleslamy et.al [16] investigated a variety of result merging methods. Lin Guoyuan et.al [17] reviewed the concept of Meta search and key technologies that greatly improve the user experience.

IV. PROPOSED WORK
The proposed work aims at design and development of Meta search engine framework using “Horizontal partitioning relevancy criteria” and “Result integration factor” Algorithms for efficient data retrieval. The proposed work is carried out in three steps namely Web harvesting, Pruning, and Result Integration that is depicted in figure 1:

![Proposed Framework of Meta search engine](image)

A Meta search engine user interface is designed by the authors through which user can post the input search query and specifies the category of the search keyword. Three top web search engines namely Google, Yahoo, and Bing are considered as member search engines to retrieve the information from heterogeneous network.

In this step input query in the form of the keyword is then forwarded to the selected member search engines and fetches “URI”. In Web, URIs are used primarily for Web documents to access in a browser. It identifies the resource in response to HTTP requests sent to perform resource identification across Web. Every Web document has its own unique “Uniform Resource Identifier”. A URI (Uniform Resource Identifier) is a string of characters that is designed for unambiguous identification of the resource over a network. The data in the Web page includes three sections namely: Web page statistics bar, document section and descriptive section. The three sections of the search engine result web page is crawled and content of the Web page is fetched and extract the data from a web document using HTML parsing techniques.

V. METHODOLOGY
The users face difficulties in finding their desired information because of the explosive growth of information on the Web. The proposed work aims at providing a solution by developing a Meta Search engine framework using proposed “Horizontal partitioning relevancy criteria” and “Result integration factor” algorithm. Implementation of the Meta search engine framework proposed in section 4 is carried out in three steps namely Member Search engine selection & Query forwarding, Pruning, and Result Integration. The details are given below:

Step -1: Web Harvesting:

Web harvesting, also referred as screen scraping fetches and extract the data from a web document using HTML parsing techniques. Three top web search engines namely Google, Yahoo, and Bing are considered as member search engines to retrieve the information from heterogeneous global information repository. A query in the form of the keyword is given as an input to this step. Member search engine selection and specifying category of search keyword facility is provided to the user. Every member search engine has its own ranking mechanism. Google search engine result page is prioritized using page rank (PR) algorithm [18], Bing prioritizes pages that are related to recent events and click-through rate (CTR) factor. The title of the website is considered as one of the major factors in Yahoo's ranking algorithm.

In this step input query in the form of the keyword is then forwarded to the selected member search engines and fetches “URI”. In Web, URIs are used primarily for Web documents to access in a browser. It identifies the resource in response to HTTP requests sent to perform resource identification across Web. Every Web document has its own unique “Uniform Resource Identifier. A URI (Uniform Resource Identifier) is a string of characters that is designed for unambiguous identification of the resource over a network. The data in the Web page includes three sections namely: Web page statistics bar, document section and descriptive section. The three sections of the search engine result web page is crawled and content of the Web page is extracted. HTML parsing technique that is used for scraping data from the web documents is shown in Fig 2:
a) Fusion Database:

The concept of fusion can be defined as merging and storing the data obtained from multiple member search engines. The results retrieved from member search engines are given as an input to the fusion database. A sample record retrieved from document section of search engine result page (SERP) when “Data mining” is given as a input query keyword is shown in figure 3:

![Fig. 2. Sections of the Web page](image)

Fig. 3. Sample retrieved record of Document section of SERP

The fusion database maintains information in the form rows and column which consists of the following attributes namely <Title, URL, Description, Category of Keyword Specified and Input Query Keyword>. The retrieved results are fused/merged and stored in fusion database for further processing. A sample fused database is shown below in table 1:

![Fig. 4. Content of the Web page after Data validation.](image)

b) Horizontal Partitioning: Horizontal partitioning also termed as "sharding" technique is applied to remove null values in a tuple. The fused database is partitioned horizontally into two “shards” based on description attribute if it holds NULL values or not as shown in fig 5:

![Fig. 5. Horizontal partitioning applied on sample fused data.](image)

c) Content relevancy: The output obtained from horizontal partitioning is given as an input to content relevancy. It identifies the relevancy of attributes <Description, Title> with respect to an input query. The content relevancy is performed using a basic string comparison operation. The description and title attributes are compared with respect to the user input query, where if the data in either of these attributes found by proposed “Horizontal partitioning relevancy criteria algorithm” (HPRC Algorithm). However, it is experienced by the end user that results are not proper and thus, keeps end user navigating within the search results for a long time. Pruning involves three steps namely: data validation, horizontal partitioning and content relevancy. The fused database is given as an input to pruning step.

a) Data Validation: In the Data Validation Step, the data collected from step 1 is validated by holding document section of web page. In this step web status bar and descriptive sections are removed and holds the document section of web page. The validated data is stored in a database. Document section displays the results in the form of page title, URL, Snippet (description) for the given input query as shown in figure 4:

![Fig. 6. The concept of fusion can be defined as merging and storing the data obtained from multiple member search engines.](image)
matching, then the results are redirected to the next step. If the data is found irrelevant it is partitioned as a separate relation and irrelevant data is discarded. The final result obtained after applying pruning presents the relevant data by discarding null and irrelevant values.

The illustration of “Horizontal partitioning relevancy criteria algorithm” (HPRC Algorithm) is as follows:

### 5.1 Algorithm: Horizontal partitioning relevancy criteria (HPRC)

**Input:**
- Fusion Database
- D = \{T_1, T_2...T_n\} \forall T_i \{i=1, 2...n\}

**Output:**
- Pruned Database
- PD = \{T_1, T_2...T_n\} \forall T_i \{i=1, 2...n\}
- Attributes \{A_1, A_2...A_n\}, \forall A_i \{i=1, 2...n\}

Represent set tuples and attributes in PRD.

**Parameters:**
- D = Database.
- PD = Pruned Database.
- T_i = Selected tuple.
- A_i = i_th Attribute.
- A(Q) = Set of attributes for input Query.
- T(Q) = Set of Tuples retrieved.
- Q = User input Query \forall k=1, 2...n.
- T2(Q) = Set of relevant Tuples.
- T1(Q) = Set of Tuples after discarding NULL Values.

1. For i=0; i<i+α; i≤ D do.
2. A_i = Query (D_i);
3. if (T_i != "\*" || "NULL" || 0) 
   a. T(Q) = \bigcup T_i \forall i=1, 2...n.
   b. A(Q) = \bigcup A_i \forall i=1, 2...n.
   c. D= \bigcup A(Q), T(Q).

   Else
   //Remove null / inconsistent/
   4. Match=Compare (Q_k,(A_i, U A_j))
      a. if (Match. equals(NULL))
         Partition T(Q) into T1(Q) discard null values
         b. Result= {Compare (Q_k, A_i) \parallel Compare (Q_k, A_j)}
         Partition T1(Q) into T2(Q) = \{T_1, T_2...T_n\}

   Final Database PD= \{T\} is with set of relevant data.

**Step 3: Result Integration**

The pruned data in step 2 is given as an input to result integration step. The pruned data is analyzed and reordered as per user’s interest by suggesting a mathematical approach that considers three parameters namely re-rank, weight and keyword relevancy score.

**a. Relinking:**

The Page rank is a query-independent concept proposed by the authors Brin and Page in their Research work on “The Anatomy of a Large-Scale Hyper Textual Web Search Engine” [19]. The concept of page rank depends upon the link structure of the web pages. The rank of each page is determined by the number of links (in links and out links) connected with each page, where “in link” refers to the links connecting to the current page and “out links” refers to the number of pages connecting from the current page. The reranking strategy applied for the Page Pi is expressed as:

\[
RR(P_i) = (1-d) + d \left( \frac{\sum_{j=1}^{P_n} PR(P_j)}{OL(P_i)} \right)
\]  

\(P_i\) represents i_th page
D= Damping factor
OL (P_i) = Out bound links of page P_i

The results retrieved from member search engines have their own ranking mechanism but the results presented needs to be further refined by considering the other parameters like weight and keyword relevancy score. The average score of these three parameters present the effective presentation of results with the highest precedence.

**b. Keyword relevancy score:**

The keyword relevancy score of each page is analyzed by comparing input query with the description attribute. The relevancy of a keyword is calculated by matching the input query with the content of description attribute. The expression for keyword relevancy score can be defined as:

\[
KRS(P_i) = \sum_{j=0}^{P_n} F(Q_k, A_j)
\]

\(P_i\) represents i_th page
F= Frequency of search keyword
Q_k = Input query keyword
Pi = represents i_th page
F= Frequency of search keyword
Q_k = Input query keyword
A_i = K_th Attribute of selected page.

The Result Integration Factor algorithm considers the
aggregate of reranking, weight and Keyword relevancy score can be given as follows:

$$\text{RIF}(P_i) = \sum_{i=1}^{PD} \frac{(RR(P_i) + W(P_i) + KRs(P_i))}{3} \quad (4)$$

W \( (Pi) \) = Weight of Page \( P_i \)
RR \( (P_i) \) = Rerank of \( P_i \)
KR\(s(P_i) \) = Keyword relevancy score of \( P_i \)

5.2 Algorithm: Result Integration Factor algorithm (RIF)

**Input:** Pruned Database \( PD= \{T_1, T_2...T_n\} \forall T_i \{i=1, 2…n\} \)

**Output:** RIF = Result integration factor for each tuple \( \{T_1, T_2...T_n\} \forall T_i \{i=1, 2…n\} \)

**Parameters:**

- \( PD= \) Pruned Database
- \( \text{RIF} = \) Result Integration Factor
- \( \text{PR}(P_i) = \) Page Rank of page \( P_i \)
- \( D= \) Damping Factor (0.85)
- \( \text{OL} = \) Number of Out bound links
- \( W (P_i) = \) Weight of page \( P_i \)
- \( F(A_k) = \) Frequency of an attribute \( A_k \)
- \( W (P_i) = \) Weight of Page \( P_i \)
- \( \text{KR}_{s}(P_i) = \) Keyword relevancy Score
- \( N = \) Number of Results retrieved.
- \( Q_k = \) User input Query

1. For \( i=0; i<\alpha; i \leq PD \) do.
2. Arrange \( PD = \text{PR}(P_i) \)

   // Data stored as per retrieved PR/

3. Calculate Result Integration Factor(RIF)
   a) Calculate \( RR(P_i) = (1-d) + (d * \text{PR}(P_i)/\text{OL}(P_i)) \)
   b) Calculate \( W(P_i) = F(A_k) \)
   c) Calculate \( \text{KR}_{s}(P_i) = F(Q_k, A_k) \)
   d) \( \text{RIF}(P_i) = (RR(P_i) + W(P_i) + \text{KR}_{s}(P_i))/3. \)
4. Rerank the PD Database as per RIF’s value.
5. Return Reordered Pruned Database.
6. End for

VI. RESULTS AND DISCUSSION

The proposed Meta search engine framework aims for efficient data retrieval using “Horizontal partitioning relevancy criteria” and “Result Integration factor” Algorithms is implemented in in PHP version: 5 (Open Source scripting language) and MySQL version: 5 (open-source relational database management system) environments. The home page of Meta search engine interface designed by the authors for the proposed work is shown in Figure 6:

![Fig. 6. Meta search Engine User Interface](image)

The user interface designed provides a facility to select member search engines and mention category of the input search query. The results retrieved after selecting member search engines and specifying category of input query keyword is shown in figure 7:

![Fig. 7. Query Forwarding](image)

The input query in the form of the keyword is then forwarded to the selected member search engines and retrieved results are stored fused/merged into the database and displayed according to the order it is retrieved. The fused data is shown in figure 8:

![Fig. 8. The Fused Data from Member search engines](image)
The pruned data is given as an input to result integration step for reordering the results based on result integration factor calculated by considering the average of three parameters namely reranking, weight and keyword relevancy score is shown in figure 10:

The relevancy analysis of sample input query keyword “Maize” is shown in Figure 11:

The proposed framework is tested by giving sample input queries. The count of results retrieved, relevant and irrelevant with respect to input query is analyzed based on two parameters namely: Relevancy score and Irrelevancy score. The relevant results refers to exact match of data with respect to the user input query and the irrelevant results does not match for the given input search query that holds Null/ Irrelevant values like from YouTube, translate, images data as a part of search from google, yahoo, Bing member search engines. The efficiency of a search system is measured using two metrics namely: Relevancy score and Irrelevancy score. Relevancy Score is calculated as the ratio of the relevant results to the number of results that are retrieved. Irrelevancy Score is calculated as the ratio of the irrelevant results to the number of results that are retrieved. The formula is as shown below:

\[
\text{Relevancy Score} = \frac{\text{Count of Results}}{\text{Relevant Results}} \quad \text{(1)}
\]

\[
\text{Irrelevancy Score} = \frac{\text{Count of Results}}{\text{Irrelevant Results}} \quad \text{(2)}
\]

The results for sample input queries and calculation of relevancy and irrelevancy score is shown in table 2:
## TABLE II
RESULTS ANALYSIS FOR SAMPLE INPUT QUERIES

<table>
<thead>
<tr>
<th>Category</th>
<th>Input Query Keyword</th>
<th>Count of Results Retrieved</th>
<th>Count of Results Relevant</th>
<th>Count of Results Irrelevant</th>
<th>Relevancy Score</th>
<th>Irrelevancy Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>Maize</td>
<td>75</td>
<td>27</td>
<td>48</td>
<td>0.36</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Cotton</td>
<td>76</td>
<td>28</td>
<td>48</td>
<td>0.37</td>
<td>0.63</td>
</tr>
<tr>
<td>Shopping</td>
<td>Alexa</td>
<td>68</td>
<td>28</td>
<td>40</td>
<td>0.41</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Dell Laptops</td>
<td>69</td>
<td>29</td>
<td>40</td>
<td>0.42</td>
<td>0.58</td>
</tr>
<tr>
<td>Entertainment</td>
<td>Sekhar Kammula</td>
<td>74</td>
<td>29</td>
<td>45</td>
<td>0.39</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Sarabjit movie</td>
<td>64</td>
<td>30</td>
<td>34</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>Education</td>
<td>Python</td>
<td>79</td>
<td>31</td>
<td>48</td>
<td>0.39</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>R Programming</td>
<td>74</td>
<td>30</td>
<td>44</td>
<td>0.41</td>
<td>0.59</td>
</tr>
<tr>
<td>Jobs</td>
<td>SAP Developers</td>
<td>71</td>
<td>27</td>
<td>44</td>
<td>0.38</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>PHP Developers</td>
<td>67</td>
<td>29</td>
<td>38</td>
<td>0.43</td>
<td>0.57</td>
</tr>
<tr>
<td>Tourism</td>
<td>Singapore</td>
<td>75</td>
<td>25</td>
<td>50</td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Bali</td>
<td>69</td>
<td>23</td>
<td>46</td>
<td>0.33</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Graphical representation of Relevancy and irrelevancy Analysis for different queries searched with respect to the member search engines selected is shown in figure 12:

![Graphical representation of Relevancy and Irrelevancy Analysis](image)

The calculation of Result Integration Factor (RIF) for a sample of five keywords is done by considering three parameters namely Page rank, Weight and keyword relevancy score. The damping factor is defined as probability that the page is retrieved and the value of damping factor considered is \( d = 0.85 \). A sample calculation of Result integration factor is shown below:
A graphical representation of result integration factor calculation performed by considering the random page with their respective weights and keyword relevancy score is shown in figure 13:

![Graphical Representation of Result Integration Factor](image)

**Table III**

**SAMPLE CALCULATION OF RESULT INTEGRATION FACTOR**

<table>
<thead>
<tr>
<th>S.No</th>
<th>Keyword</th>
<th>PageRank</th>
<th>W(P_i)</th>
<th>KR_r(P_i)</th>
<th>RIF(P_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Maize</td>
<td>0.85</td>
<td>9</td>
<td>1</td>
<td>7.8</td>
</tr>
<tr>
<td>2</td>
<td>Samsung Phone</td>
<td>0.85</td>
<td>2</td>
<td>3</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>Web Mining</td>
<td>0.85</td>
<td>9</td>
<td>1</td>
<td>7.8</td>
</tr>
<tr>
<td>4</td>
<td>China</td>
<td>0.85</td>
<td>8</td>
<td>2</td>
<td>3.55</td>
</tr>
<tr>
<td>5</td>
<td>Ram Gopal Varma</td>
<td>0.85</td>
<td>2</td>
<td>6</td>
<td>0.43</td>
</tr>
</tbody>
</table>

VII. CONCLUSION

The Meta search engines that are built on top of traditional search engines enhance search capability of the user by integrating data from multiple member search engines. Meta search engine does not have a unique indexing mechanism to reorganize the data in its local copy of database. The inclusion of steps like Pruning and Result integration plays major role in the implementation of proposed Meta search engine framework. In proposed work pruning is performed to remove Null and irrelevant data by applying data validation and proposed “Horizontal Partitioning relevancy criteria algorithm”. The pruned data is reorganized by considering the three parameters namely Reordered rank, weight, and Keyword relevancy score using “Result Integration factor algorithm”. The relevant results are displayed as per user’s interest in descending order of RIF value. Semantic web capabilities can be applied to the data retrieved from Meta search engine framework for any kind of input query given by user. Application of security through cryptographic approach is carried out as future work.

REFERENCES


---

Prof. M.S. Prasad Babu was born on 12-08-1956 in Andhra Pradesh, India. He obtained his bachelors to doctoral degrees from Andhra University. He has 39 years of teaching and research experience. He guided 12 PhD's and 210 PG students for their thesis. He was the president of ICT section of ISCA in 2006-07. He attended about 50 National and International Conferences in India and abroad and presented keynotes. He contributed about 250 papers National and International journals and conferences. He developed 10 international reputed Web portals. He won ISCA Young Scientist Award in 1986, State Best teacher award for engineering in 2015 and Dr. Sarvepalli Radha Krishnan Best Academician Award of Andhra University for 2014. He was the Conference Steering Chair for eight IEEE ICSESS of Beijing, China from 2010 to 2017. Prof Babu is presently working as Vice Principal of AU Engg College, Chairman, Faculty of Engineering of Andhra University and senior Professor in Computer Science & Systems Engineering discipline.

G. Sudeepthi was born in, Rajahmundry, Andhra Pradesh, India in 1983. She received B. TECH from National Institute of Technology Warangal, India in 2005 and M. Tech in Computer Science & Engineering from KLC Vijayawada, India in 2007. She is working as part time research scholar in the Dept. of CS & SE, Andhra University and as a Assistant Professor at Anurag Group of Institutions Hyderabad, India. She has got more than ten years of teaching experience. She Qualified FET-2011 Conducted by JNTUH. She contributed seven research papers in International journals and Presented one research paper at International conference in Beijing, China and another in Warangal, India.