S-MSE: ASSESSMENT META SEARCH ENGINE USING SEMANTIC SIMILARITY AND REPUTATION MEASURE

P. VIJAYA¹, G. RAJU², SANTOSH KUMAR RAY³

¹Research Scholar, Karpagam University, Coimbatore, TamilNadu, India
²Department of Information Technology, Kannur University, Kerala, India
³Department of Information Technology, Al Khawarizmi International College, Al Alin, UAE
E-mail: pvijaya0169@gmail.com

ABSTRACT

In order to increase web search effectiveness, Meta search engines are invented to combine results of multiple search engines as a result of larger coverage of indexed web. Meta search engine is a kind of system which is useful for internet users to take advantage of multiple search engines in searching information. Recently several approaches were developed using ontology and ranking measures. Accordingly, Meta search engine is developed here using ontology and semantic similarity measure. In order to bring semantic in keyword matching, a semantic similarity measure (SSM) is developed. Here, every concept sets are matched with the title sets using SSM that consider the hyponyms and hyponyms of the keywords presented in the title sets. Along with three different ranking measures relevant to contents, title sets and raking value given by the standard search engines are effectively combined to improve the effectiveness. Finally, the experimentation is carried out using different set of queries and the performance of the meta-search engine is evaluated using TREC-style average precision (TSAP) measure. The proposed semantic meta-search engine provides 80% TSAP which is high compared with existing search engine and meta-search engine.

Keywords: S-MSE, Search Engine, Semantic Similarity, Reputation Measure, TREC, TSAP

1. INTRODUCTION

With fluctuating capabilities, the ubiquity of the Internet and Web has directed to the emergence of numerous Web search engines. These search engines guide Web sites, images, Usenet news groups, content-based directories, and news sources with the objective of producing search effects that are most related to user queries. On the other hand, just a small number of web users really make out how to employ the true power of Web search engines. Search engines have started offering access to their services through different interfaces in order to address this problem [1]. For any specified query, search engine as a device to explore the Web must get the desired effects. Achievement of a search engine is directly reliant on the satisfaction level of the user. Within a short time of interval, users longing the data to be offered to them. They in addition anticipate that the most related and latest information are existing [2, 3] to them by the search engines. A large amount of the search engines can never totally please user's necessities and the search effects are frequently very imprecise and unrelated [4].

A lot of researchers who have accounted previously about different features of search engines are there in [5-17]. Search devices for the web can be categorized as Search Engines, Directory Services, Meta-Search Engines, and Hybrid Search Services. Google, Inktomi, etc., are distinctive search engines. Yahoo is a well-known directory service. Meta-Crawler, ProFusion, Dogpile etc., are meta-search engines. MSN search and Yahoo can be named as hybrid search services too, as they have a search engine with directory services integrated in them. Every search engine has three key efficient phases, that is, Web Data Acquisition (WDA), Web Data Indexing (WDI) and Web Data Rendering (WDR). They are alienated into general purpose and special purpose search engines [3]. A meta- search engine is the type of search engine to offer users with information services and it does not contain its own database of web pages. It sends search terms to the databases preserved by other search engines and presents users the effects that come from all the search engines queried [4].
Amongst the researchers, semantic search engine have obtained a considerable attention very lately. The processes of semantic search engine normally may comprise: 1) Extracting the relevant concepts from the sentence, from where the user question can be interpreted; 2) The meaningful concepts relevant to the query, is removed by means of Ontology. By employing semantically richer ontologies in semantic search engine, the subsequent advantages can be acquired. Initially, to explain the terminology of the application and the domain knowledge in more detail, ontologies can be applied, as example we can regard as, relations among categories in dissimilar sights can be termed. Secondly, for making semantically more precise annotations in terms of the domain knowledge, ontologies can be applied. Thirdly, with the assist of ontologies, the user can convey the queries more accurately and unambiguously which leads to improved precision and recall rates. Fourthly, class definitions and inference mechanisms, such as instance-level metadata, property inheritance, can be developed semantically through ontological [26].

In this research effort, to develop web documents retrieval we have offered a semantic meta search engine. Here, a new semantic similarity measure is designed to find the similarity between two keywords. This semantic resemblance measure makes use of mathematical set operation based on hyponyms and hyponyms. The main steps of the suggested meta search engine encloses, i) ranking based on web contents, ii) ranking based on title keywords, and iii) ranking based on reputation among different search engines. A hybrid measure is improved to re-rank the web document that is regained through dissimilar search engine by joining these three measures. At this point, the weighted average calculation formulae are applied with the dissimilar weights prearranged for three dissimilar ranking measures. Lastly, the presentation is assessed with the assist of TSAP measure. Remaining of the paper is arranged as follows: Section 2 offers the assessment of associated works and section 3 offers the semantic similarity measure designed for the semantic meta search engine. Section 4 offers the hybrid ranking measure for the suggested meta search engine. Section 5 converses the effects and section 6 ends the paper.

2. REVIEW OF RELATED WORKS

Using web search engines as corpus or physically gathered corpus, quite a lot of researches compute semantic similarity between two words. Paik, I et al. [24] have suggested a technique to find the word Ri between two words P and Q and removed a relation of the words with PLSI (Probabilistic Latent Semantic Indexing). The effects of the experiments demonstrated that by means of the PLSI with smaller latent class such was efficient in getting Ri which was more associated to P and Q, and by the PLSI with over 5 latent class was efficient in getting implied relation between P and Q. An information customization system that unites meta-search and unsupervised learning has been suggested by Mohamed Salah Hamdi et al. [23]. At the same time a meta-search engine searches for multiple search engines and comes back a single record of results. As the results, was regularly clutching the initial items from the relevancy-ranked list of hits returned by the individual search engines, the effects regained by this engine were highly related. Using the Kohonen Feature Map, a self-organizing semantic map such that the documents of related substances are situated close to one another, is erected.

A lightweight strategy for service discovery has been suggested by Giantsiou, Let al. [25]. Their strategy includes three main phases. Initially, the semantic service explanations were regained and accumulated locally during the crawling phase. After that the semantics of every explanation were plotted to a service meta-model and the effecting triples were accumulated in a RDF repository, in the homogenization phase. Lastly, the users were facilitated to query the fundamental repository and discover online services, at the search phase. Guang-ming and Wen-juan [18] have suggested the proficient meta search engine system, employed the CC4 neural network algorithm to compute web page relation degree, and attained the high degree of professional web pages. They worked out the problem of the breadth the people’s access to information, suggested a superior solution to explore in the ocean of information. It completely examined the web page information, used CC4 neural network algorithm to reviewer the related web pages and optimal rank and then joined the professional dictionary to strain the sort effects. Lastly, the experimental results demonstrated that the technique enhanced the search quality to the particular specialty.

V Raval and Padam [19] have suggested a meta-search engine called EGG that was meant to use power of the Google for more precise and combinatorial search. They attained through uncomplicated manipulation and automation of
Google functions that are available from EGG through the Google. The suggested method attained through uncomplicated manipulation and automation of the presented Google functions. The suggested meta-search engine sustained search based on “Combinatorial Keywords” and “Normal Search”. A detailed assessment expressed how one can exploit the competence of Google cluster architecture through its programmable Web services by making advanced search features at a third party user application level. Meow, a semantic meta search engine has been offered by Keong and Anthony [20] that was capable to change a query from a non-semantic search result into a semantic search result. DBpedia was employed as a knowledge base since it encloses a huge dataset that was capable to explain general ideas for different domains. Meow was able to enhance the search result from Google by using DBpedia. On the other hand, the information that was queried from DBpedia for definite topic may be huge and may contain ideas that are not related.

3. NEW SEMANTIC SIMILARITY MEASURE

Finding the similarity between two keywords is important for any information retrieval application. Generally, the string matching distance and vector space modelling-based approaches are widely used to find the similarity between keywords. Here, we have given new semantic similarity measure for finding the similarity between two keywords. In order to find the similarity between two keywords, the following mathematical model will be used. This semantic similarity measure is then used to find the ranking measure based on title keywords. At first, two keywords such as, \( t_1 \) and \( c_1 \) are given to Wordnet ontology (wordnet.princeton.edu) to extract the hypernyms and hyponyms of those two keywords. Once we extract them, three different sets are individually formed for those keywords. For example, first set of keyword \( t_1 \) contains a set of keywords that are related to the hypernyms \( \text{Hypr}_{t_1} \) and the second set contains the hyponyms \( \text{Hypo}_{t_1} \) and third set contains whole keywords presented in hypernyms and hypernyms \( \text{U}_{w_{t_1}} \). Then, the semantic similarity measure will be found out using the following formulae.

\[
\text{Sim}(k_1, k_2) = \frac{n(T) - [(a \cdot H_w + \beta \cdot H_y_w + \gamma \cdot U_w)]}{n(T)}
\]

Where, \( n(T) \) is the number terms chosen for weighted summation. Here, the value for \( n(T) \) is three. \( H_w \) contributes to similarity based on the similarity in hypernyms of two keywords taken for similarity finding. \( H_y_w \) is used to bring the similarity of two keywords based on hyponyms set \( U_w \) and \( U_{w_{t_1}} \) bring the similarity value based on similarity of universal set. \( a, \beta, \gamma \) are the weightage constants that are computed based on the following formulae.

\( H_w \) is computed by taking the intersection and union operation of hypernyms sets formed from two different keyword. Similarly, \( H_y_w \) is computed by taking the difference in between the union and intersection similarity. \( U_w \) is found out using similarity of universal set.

\[
H_w = |(\text{Hypr}_{t_1} \cup \text{Hypr}_{c_1})| - |(\text{Hypr}_{t_1} \cap \text{Hypr}_{c_1})|
\]

Where, \(|(\text{Hypr}_{t_1} \cup \text{Hypr}_{c_1})|\) provides number of keywords presented after finding the union in between \( (\text{Hypr}_{t_1} \cap \text{Hypr}_{c_1}) \) and \( (\text{Hypr}_{t_1}) \).

\[
|\left(\text{Hypr}_{t_1} \cap \text{Hypr}_{c_1}\right)|\) provides number of keywords presented after finding the intersection in between \( (\text{Hypr}_{t_1}) \) and \( (\text{Hypr}_{c_1}) \).

\[
H_y_w = |(\text{Hypo}_{t_1} \cup \text{Hypo}_{c_1})| - |(\text{Hypo}_{t_1} \cap \text{Hypo}_{c_1})|
\]

Where, \(|(\text{Hypo}_{t_1} \cup \text{Hypo}_{c_1})|\) provides number of keywords presented after finding the union in between \( (\text{Hypo}_{t_1} \cap \text{Hypo}_{c_1}) \) and \( (\text{Hypo}_{t_1}) \).

\[
|\left(\text{Hypo}_{t_1} \cap \text{Hypo}_{c_1}\right)|\) provides number of keywords presented after finding the union in between \( (\text{Hypo}_{t_1}) \) and \( (\text{Hypo}_{c_1}) \).

\[
U_w = |(\text{U}_{w_{t_1}} \cup \text{U}_{w_{c_1}})| - |(\text{U}_{w_{t_1}} \cap \text{U}_{w_{c_1}})|
\]

Where, \(|(\text{U}_{w_{t_1}} \cup \text{U}_{w_{c_1}})|\) is found by taking modulus operation after taking the union operation in between \( (\text{U}_{w_{t_1}}) \) and \( (\text{U}_{w_{c_1}}) \). \(|(\text{U}_{w_{t_1}} \cap \text{U}_{w_{c_1}})|\) is found by taking modulus operation after taking the intersection operation in between \( (\text{U}_{w_{t_1}}) \) and \( (\text{U}_{w_{c_1}}) \).

The weightage constants are computed based on the following formulae. Here, the total number of keywords presented after taking the union operation is used as the weightage parameters of the three terms presented in the proposed semantic similarity formulae.

\[
a = \frac{1}{n((\text{Hypr}_{t_1} \cup \text{Hypr}_{c_1})} \]
where, $n((\text{hyp})_1 \cup (\text{hyp})_2)$ is number of keywords in union set of hypernyms. $n((\text{hyp})_3 \cup (\text{hyp})_4)$ is number of keywords in union set of hyponyms. $n(\cup_{1} \cup \cup_{1})$ are total keywords in the union set of universal set.

4. HYBRID RANKING AND SEMANTIC SIMILARITY MEASURE FOR ONTOLOGY BASED META-SEARCH ENGINE

This section presents and discusses the proposed hybrid ranking measure. The hybrid ranking measure combines three ranking measure obtained through concept, title and reputation matrix. Here, the finding of similarity between two keywords in title matrix computation is done through the semantic similarity measure designed newly. The overall steps of the proposed hybrid meta search engines is explained through three different steps:

1. Pre-processing
2. Finding of ranking measure
3. Hybrid ranking measure

4.1 Pre-processing

At first, the query keywords are given to various search engines namely, Google, Bing and Yahoo. Then, top 'n' results from each of the three search engines are selected for rearranging the order of the results retrieved from the search engines. For the entire documents, the keywords are extracted from web document and stop words are removed from the extracted content of web document. Subsequently, sequential keywords are extracted based on the mutual relation between two consecutive keywords given in the document. The sequential keyword is extracted based on a mutual association value, which is defined by,

$$MA(k_i, k_j) = \log \frac{P(k_i, k_j)}{P(k_i)P(k_j)}$$

Where, $MA(k_i, k_j)$ is the mutual association between the keywords $k_i$ and $k_j$. $P(k_i, k_j)$ is the joint probability that both keywords to be present in the text window and the $P(k_i)$ is the probability that a keyword $k_i$ to appear in the text window. Once we find the mutual values for the entire combination of consecutive words, the final set of sequential keyword set is identified if the mutual association values are high. Then, the sequential keywords are stored to a set $SW$.

$$SW = \{sw_1, sw_2, ..., sw_n\}$$

Similar the process is repeated for all the documents and it is constructed as matrix where, row indicates sequential keywords and column indicates set of documents. Every value in the matrix is in mutual association that is obtained from the previous step.

4.2 Computation of ranking measure

This step is useful to find three ranking measure made for meta-search engine. Three different measures focus on different aspect like content, title and the reputation among three search engines.

i) Generating ranking measure for web documents based on concept (R1)

For the input query, we form the concept sets that are generated after matching the query word with the ontology. Then, the combination is generated for all the extracted keywords from the ontology with the query keyword. For all the combinations (concepts), we find the score value after matching the concepts with the sequential word matrix. It produces a concept matrix that contains set of concepts as rows and documents as columns. The concept matrix is represented as follows:

$$C = [c_1, c_2, ..., c_n]$$

where, $c_i$ is the concept vector that contains the mutual association value related to all the documents for the specific concept $c_i$.

Then, the trusting measure of concepts are computed for every value located in the concept matrix. The formulae for computing the trusting measure of concepts is given as,

$$Trc_i = \frac{c_i}{n}$$

Where, $n$ is the total documents considered.

After computing the trusting measure for all the values, we put the trusting measure values in the corresponding location so that we obtain new matrix, called trusting matrix of concept.

$$TRMC = [Trc_1, Trc_2, ..., Trc_n]$$

Then, the trusting matrix is used to provide the ranking measure of concept matrix $Rl(j)$ based on the following formulae.
Generating ranking measure for web documents based on title (R2)

To find the ranking measure based on title, we extracts the title keywords from the title tags of all the web pages returned by search engines and constructs the matrix regarding the titles. The title matrix is matrix that contains total title keywords as column and the concepts in the concepts matrix as rows. Then, the values are filled up by finding the semantic similarity score of the concept and title word. For finding the semantic similarity score, the new formulae designed in the previous section is used. The title matrix of the proposed approach is indicated as follows:

\[ R_1(j) = \text{Trc}_1^j + \frac{1}{2} \sum_{i=2}^{m} \text{Trc}_i^j; \quad 1 \leq j \leq n \]

Where, \( \text{Trc}_i^j \) is the title vector that contains the semantic similarity value related to all the documents. Then, the trusting measure of title matrix is computed for every value located in the title matrix. The formulae for computing the trusting measure is given as,

\[ \text{Trt}_i^j = \frac{\text{Trc}_i^j}{n} \]

Where, \( n \) is the total documents considered. After computing the trusting measure for all the values, we put the trusting measure values in the corresponding location so that we obtain new matrix, called trusting matrix of title.

\[ \text{TRMT} = [\text{Trt}_1, \text{Trt}_2, ..., \text{Trt}_n] \]

Then, the trusting matrix of title is used to provide the ranking measure of title matrix based on the following formulae:

\[ R_2(j) = \text{Trt}_1^j + \frac{1}{2} \sum_{i=2}^{m} \text{Trt}_i^j; \quad 1 \leq j \leq n \]

Generating ranking measure for web documents based on reputation matrix (R3)

Here, the ranking measure is designed based on the reputation of web pages among different search engines. The ranking given by three different search engine like, Google, Yahoo and Bing is added to obtain the reputation value of the web documents extracted.

\[ k(j) = G^r(j) + Y^r(j) + B^r(j) \]

Where, \( G^r(j) \) is ranking of web document ‘j’ obtained from Google, \( Y^r(j) \) is ranking of web document ‘j’ obtained from Yahoo and \( B^r(j) \) is ranking of web document ‘j’ obtained from Bing. Then, ranking measure for web documents based on reputation matrix (R3) is computed based on the following formulae, in which ‘n’ is the total number of document considered.

\[ R_3(j) = 1 - \frac{k(j)}{n} \]

4.3 Hybrid Ranking Measure

After finding three ranking measures, the hybrid measure for every web document is computed based on hybrid ranking measure. Here, the weighted average formula is used here to find the hybrid measure by putting appropriate weights for concept, title and reputation matrices.

\[ R(j) = \frac{w_1 v_1(j) + w_2 v_2(j) + w_3 v_3(j)}{w_1 + w_2 + w_3}; \quad 1 \leq j \leq n \]

Here, \( w_1=0.5, \ w_2=0.25, \ w_3=0.25 \). Once the hybrid ranking vector is calculated, the values are arranged in the descending order. The web documents are shown to the users based on the values of the sorted hybrid ranking vector vector.

5. RESULTS AND DISCUSSION

This section presents the experimental analysis of the proposed semantic meta-search engine and its discussion. The proposed system is implemented in Java under JDK 1.7 and all the experiments are conducted on a system with intel core i5 processor and 3 GB RAM. The proposed system is evaluated with 50 queries and only the sample evaluation of three queries is given in section.

5.1 Evaluation Factor

The meta-search engine provides combined results from various general purpose search engines. Therefore, the traditional parameters of search engine evaluations such as recall and precision cannot be used in case of meta-search engines. A popular measure for evaluating the effectiveness of search engines is the TREC-style average precision (TSAP) [21]. In this paper, TSAP at different ranking order N, denoted as TSAP is used to evaluate the effectiveness of the proposed algorithm. TSAP at ranking order N is defined as:
\[ TSAP = \frac{\sum_{i=1}^{N} r_i}{N} \]

Where \( r_i = 1/i \) if the \( i^{th} \) ranked result is relevant and \( r_i = 0 \) if the \( i^{th} \) result is not relevant. It is easy to see that TSAP takes into consideration both the number of relevant documents in the top \( N \) results and the ranks of the relevant documents. TSAP tends to yield a larger value when more relevant documents appear in the top \( N \) results and when the relevant documents are ranked higher.

5.2 Performance analysis of S-MSE based on ranking order

The performance of the S-MSE is evaluated with three different queries such as, clustering, cryptography and image mining. After obtaining the results from the hybrid search engine, the obtained results and given to the users to find the relevant pages of them. Based on the relevancy, TSAP measure is computed and it is plotted in a graph shown in figure 1. From the figure, we can identify that the highest TSAP value of 83% is achieved for “cryptography” when \( N=5 \). For the query keyword “clustering, the highest value is 90% that is obtained for \( N=15 \).

<table>
<thead>
<tr>
<th>Queries</th>
<th>TSAP values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=5</td>
</tr>
<tr>
<td>Clustering</td>
<td>80</td>
</tr>
<tr>
<td>Cryptography</td>
<td>83</td>
</tr>
<tr>
<td>Image mining</td>
<td>68</td>
</tr>
</tbody>
</table>

5.3 Comparative analysis with existing search engine and Meta search engine

i) Analysis with existing search engine

Figure 2 shows the comparative evaluation of the proposed meta-search engine with the standard search engine like, google, yahoo and bing. For the three queries, the proposed meta-search engine outperformed all the search engine in TSAP value. The second highest value is achieved for yahoo and third ranking for google.
ii) Comparative analysis with existing meta search engine

Analysis with 2002 web track data

In this section, we plot a comparative study of the proposed approach with some other search engines. We use the TREC 2002 web track and 2002 TREC Web Track has 50 topics indexed from 551 to 600. In this paper, for each topic, only the title part is used as a query to send to the search engines, because the titles are short, similar to most Internet queries submitted by real users. The average length of the titles of these 50 topics is 3.06. The description and the narrative describe what documents should be considered relevant to the corresponding topic. This information is served as the standard criteria for us to judge the relevancy of the collected result documents. Each query is submitted to every component search engine.

For each query and each search engine, the top 10 results on the first result page are collected. The results are plotted based on the precision as evaluation criteria.

\[
\text{Precision} = \frac{\text{Number of relevant documents}}{N}
\]

Here, we consider N as total number of top results and the relevant document from the top result are compared manually. The relevancy of each document is manually checked, based on the criteria specified in the description and the narrative part of the corresponding TREC query. The collected data and the documents, together with the relevancy assessment result, form our test-bed. The test-bed is stored locally so it will not be affected by any subsequent changes from any component search engine.

<table>
<thead>
<tr>
<th>index</th>
<th>Title</th>
<th>S-MSE</th>
<th>Google</th>
<th>Yahoo</th>
<th>Bing</th>
</tr>
</thead>
<tbody>
<tr>
<td>551</td>
<td>intellectual property</td>
<td>0.92</td>
<td>0.88</td>
<td>0.87</td>
<td>0.79</td>
</tr>
<tr>
<td>552</td>
<td>Foods for cancer patients</td>
<td>0.91</td>
<td>0.84</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>553</td>
<td>federal funding mental illness</td>
<td>0.93</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>554</td>
<td>Home buying</td>
<td>0.95</td>
<td>0.89</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td>555</td>
<td>criteria obtain U.S.</td>
<td>0.92</td>
<td>0.81</td>
<td>0.80</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 3. Precision Analysis
The table 3 represents the precision based analysis of the proposed approach with three major search engines Google, yahoo and Bing. The analysis showed that the S-MSE algorithm has better precision over the other search engines on data 2002 TREC web track. We processed all the 50 indices from 551 to 600, the same result obtained for all indices as above. In this section, we also made a comparison of the semantic meta-search engine, which has been done with an existing meta-search algorithm. The existing meta-search is used for evaluation of result merging strategies for meta-search engines [22]. The above stated approach implemented three algorithms for the evaluating the meta-search process and the algorithms are derived based on the similarity of the documents and the similarity measures used are SRRsim, SRRrank and SRRsimMF[22]. The comparative evaluation is given in figure 3. From the figure, the proposed meta-search engine obtained the maximum TSAP value for all the ranking order. For N=10, the proposed meta-search engine obtained 80% TSAP while the existing measure achieved about 25%.

<table>
<thead>
<tr>
<th>Queries</th>
<th>TSAP values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=5</td>
</tr>
<tr>
<td>SRRsim</td>
<td>0.381</td>
</tr>
<tr>
<td>SRRrank</td>
<td>0.370</td>
</tr>
<tr>
<td>SRRsimMF</td>
<td>0.381</td>
</tr>
<tr>
<td>S-MSE</td>
<td>0.770</td>
</tr>
</tbody>
</table>

6. CONCLUSION

We have presented a semantic meta-search engine to improve web documents retrieval. Here, a new semantic similarity measure is designed by the use of mathematical set operation based on hyponyms and hyponyms. This measure is used to find the similarity between two keywords. Along with, three different ranking measures relevant to contents, title sets and ranking value given by the standard search engines are effectively combined using weighted average formulae. Finally, the evaluation of the proposed meta-search engine is carried out using different set of queries and the comparative analysis is performed using TSAP measure. The hybrid semantic meta-search engine provides 80% TSAP which is high compared with existing search engine and meta search engine.

REFERENCES


