Usearch: A Meta Search Engine based on a New Result Merging Strategy

Tarek Alloui[1], Imane Boussebough[2], Allaoua Chaoui[1], Ahmed Zakaria Nouar[3]
and Mohamed Chaouki Chettah[3]

[1]*MISC Laboratory, Department of Computer Science and its Applications, Faculty of NTIC,*
University Constantine 2 Abdelhamid Mehri, Constantine, Algeria

[2]*LIRE Laboratory, Department of Software Technology and Information Systems, Faculty of NTIC,*
University Constantine 2 Abdelhamid Mehri, Constantine, Algeria

[3]*Department of Computer Science and its Applications, Faculty of NTIC, University Constantine 2,*
Abdelhamid Mehri, Constantine, Algeria

Keywords: Meta Search Engine, Ranking, Merging, Score Function, Web Information Retrieval.

Abstract: Meta Search Engines are finding tools developed for improving the search performance by submitting user queries to multiple search engines and combining the different search results in a unified ranked list. The effectiveness of a Meta search engine is closely related to the result merging strategy it employs. But nowadays, the main issue in the conception of such systems is the merging strategy of the returned results. With only the user query as relevant information about his information needs, it's hard to use it to find the best ranking of the merged results. We present in this paper a new strategy of merging multiple search engine results using only the user query as a relevance criterion. We propose a new score function combining the similarity between user query and retrieved results and the users' satisfaction toward used search engines. The proposed Meta search engine can be used for merging search results of any set of search engines.

1 INTRODUCTION

Nowadays, the World Wide Web is considered as the largest information source in the World. But the challenge is to be able to find, from the huge amount of documents available on the Web and in a timely and cost-effective way, the documents that best match user information needs.

For a fairly rich and relevant search results, a search engine remains limited because it can't index all the web pages. This situation obliges the user to move, for the same query, from a search engine to another to find more relevant results to his needs.

To simplify the task to the user, it's interesting to offer him tools to invoke, with the same query, multiple search engines simultaneously. These tools are called Meta Search Engines (MSE). They provide a uniform query interface for Internet users to access multiple existing search engines (Meng, 2008). After the returned results from all used search engines have been collected, the Meta search engine merges them into a single ranked list. The major advantages of MSEs are their abilities to combine the coverage of multiple search engines and to reach the deep Web.

Result merging is combining the search results returned from multiple search engines into a single ranked list. A straightforward way to perform results merging is to download locally the retrieved documents and then compute their similarities with the user query using a global similarity function. Then, the results will be ranked using the computed scores (Renda and Straccia, 2003; Lu and al 2005). The main advantage of this approach is that it provides a uniform way to compute ranking scores. But the main problem of this technique is a longer response time due to documents' downloading.

Another result merging technique is to use scores, returned from each used search engine, in the computation of new normalized scores to make them more comparable. But not all search engines return local ranking scores, and even if they can be obtained, there is no information about how these scores are computed by each search engine.

Alloui, T., Boussebough, I., Chaoui, A., Nouar, A. and Chettah, M..
Usearch: A Meta Search Engine based on a New Result Merging Strategy.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 531-536
ISBN: 978-989-758-158-8

Also, when there are common results between different search engines, they are ranked differently since search engines use different ranking formulas and term weighting techniques.

In this paper, a new result merging strategy is introduced. In our approach, we combine two techniques: computing a similarity score for each retrieved result using title, description and snippets instead of the full document; and including users' satisfaction toward the used search engines in the computation of the final ranking scores.

The main contribution of our proposed approach is the score function that combines two important parameters: the similarity score between the user query and the retrieved document, and the users' satisfaction to each search engine. This function helps us to obtain our own ranking without neglecting the results ranking of each search engine.

2 RELATED WORK

Nowadays, very few search engines report their ranking scores, but many earlier approaches focused on taking into consideration and using these ranking scores (Aslam and Montague, 2001; Callan and al, 1995; Gauch and al 1996).

Another approach of merging results is to download the full documents, analyzes them and computes a similarity score between user query and each document. Then, these scores are used to rank the merged list (Renda and Straccia, 2003). The drawback of this approach is a high time cost before merged results can be displayed to the user.

Instead of downloading the full documents, most of the current metasearch engines (Lu and al 2005; Rasolofo and al, 2003; Jadidoleslamy, 2012), use representative information of the document such as the URL, title, summary (snippets) and search engine usefulness. Generally, a similarity between the user query and the document title and snippet is computed. The work reported in (Rasolofo and al, 2003) is the most likely similar to our work in the use of the available information such as document title, snippet, local rank and search engine usefulness. But instead of search engine usefulness, we chose to use users' satisfaction toward the used search engines.

3 THE PROPOSED APPROACH

In this section, we present a metasearch engine based on a new result merging strategy. The main objective of *Usearch* is to help any user getting more relevant results from the Web. This is achieved by querying multiple search engines simultaneously. So the proposed system avoids the user moving from a search engine to another by submitting his queries to the different search engines, retrieving the results, merging them using the proposed merging strategy and presenting them to the user in a single ranked list.

The proposed system functioning is distributed on four important modules: Interface Module, Querying and Retrieving Module, Pre-processing Module and Merging Module (Figure 1). The Merging Module implements the principle of our merging strategy. To better understand the proposed system, we will describe in the following the principle and the functioning of each module.

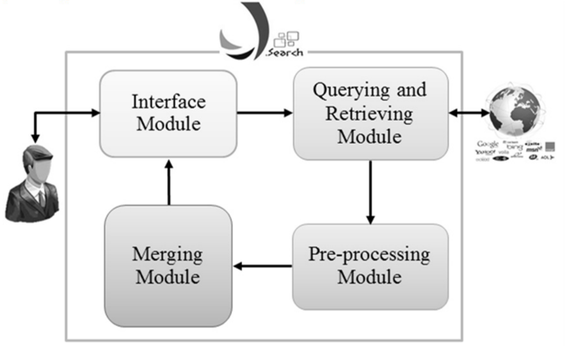

Figure 1: Architecture of the Metasearch engine *Usearch.*

3.1 Interface Module

This module allows users to express their queries in a simple and intuitive way using their own keywords without any specific search engine representation or constraint.

It receives the user query as an input and sends it to the *Querying and Retrieving Module*. After retrieving the search engines results and merging them in a single list by the *Merging Module*, they are sent to the *Interface Module* to be displayed to the user.

This module also allows the user to choose the search engines he wants to invoke, to set for each one his satisfaction, to activate and use advanced search of all used search engines. Indeed, the user can enter, like what is proposed in most search engines, more specific information on his information needs.

3.2 Querying and Retrieving Module

This module receives the user query, converts it into a query specific to each search engine and encapsulates it in the search engine URL. After executing simultaneously all search engines URLs, this module retrieves the returned results lists of all search engines and sends them to the *Pre-processing Module*.

3.3 Pre-Processing Module

This module has the following tasks:

- Analyzing the retrieved results for title, description (snippets) and URL extraction.
- Identifying common and non-common results between all search engines results.

First of all, each result list of each search engine is analyzed to extract for each result the following information:

- The result title,
- The result description,
- The result URL,
- The rank of the result in the corresponding search engine.

Then, the module analyzes all the results lists of all search engines, to identify the common and non-common results. In addition to the previous information, the module associates to each common result its rank in each search list it appears.

After identifying the common and non-common results, the module sends these two lists to the *Merging Module* to compute the scores of all results using the proposed score function and to merge them into a single ranked list.

3.4 Merging Module

This module implements the proposed result merging strategy. The two main tasks of this module are:

- Computing the scores of all results using the information extracted by the *Pre-processing Module*,
- Merging and ranking the results using the computed scores.

3.4.1 Score Function

The main contribution of this work is the proposed score function. It relies on the information extracted and sent by the Pre-processing Module to compute a similarity score for each retrieved result. To do this, we chose these following relevance criteria in the design of the score function:

- Occurrence of the user query "**as it is**" in the result description, noted "***$Query_{Occ}$***",
- Occurrence of query keywords in the result description, noted "***$Keywords_{Occ}$***",
- Occurrence of query keywords in the result title, noted "***$Title_{Occ}$***",
- Result rank in the corresponding search engine, noted "***Rank***".

The proposed score function is then as follows:

$$\begin{aligned} Score(i) = (C_1 * Query_{Occ}) \\ + (C_2 * Keywords_{Occ}) \\ + (Title_{Occ}/N) + (C_3/Rank) \end{aligned}$$

Where:

- ***i***: the result,
- ***C_1***: query occurrence coefficient,
- ***C_2***: query keywords occurrence coefficient,
- ***N***: the number of keywords in the user query,
- ***C_3***: engine ranking coefficient.

These coefficients are the weightings of the different performance criteria we chose in our score function. This score is computed for each result in the common and non-common lists. For the common results, if a result appears in ***m*** search engines results, the system will compute ***m*** scores because this result has a different "***Rank***" in the result list of each search engine.

So, the final score of the common results is computed using the following formula:

$$Score_{common}(i) = \sum_{j=1}^{m} CC_j * Score(j)$$

Where:

- ***i***: the result,
- ***m***: the number of search engines where the result "***i***" appears,
- ***CC_j***: contribution coefficient of the search engine "***j***" in the final results,
- ***Score(j)***: the score of the result "***i***" using its "***Rank***" in the search engine "***j***".

As said before, we combine two techniques: computing a similarity score for each retrieved result using title, description and snippets instead of the full document; and including users' satisfaction toward the used search engines in the computation of the final ranking scores. The users' satisfaction toward search engines is combined in the final score function of the common results. Indeed, each "***CC_j***" coefficient represents user satisfaction toward the corresponding search engine. And the sum of these coefficients satisfies the following condition:

$$\sum_{j=1}^{m} CC_j = 1$$

3.4.2 Results Merging

Once results scores computed, the *Merging Module* performs results merging and ranking according to their scores in a descendant order. In the ranking process, results with the same scores are treated according to the following conflict resolution policy:

- Priority 0 is assigned to results in the common list,
- If the result isn't in the common list, then the system will assign priority according to users' satisfaction toward used search engines as follows:
    - Priority 1 for the most satisfying search engine,
    - Priority 2 for the second search engine,
    - …
    - Priority ***m*** for the last search engine.

Once the results merged and ranked, the *Merging Module* sends the final list to the *Interface Module* to be displayed to the user.

4 EXPERIMENTAL RESULTS AND DISCUSSION

To test the feasibility and effectiveness of our result merging strategy, we implemented a prototype of the proposed system. Since our Meta search engine supports querying multiple search engines, we chose in a first step to fix the number of search engine to "3". And we opted for the three most famous and most used search engines: *Google*, *Bing* and *Yahoo*. Also, instead of testing *Usearch* on a document collection, we decided to test it directly on the Web. Indeed, what can work well on a small collection and gives very good results, can yield the worst results when deploying it on a changing environment such as the Web. So the large scale transition will not always give the same results as on a small collection like TREC collections, because there is a lot of heterogeneous documents; websites born, evolve and disappear; search engines always change their search algorithm and update frequently their indexes. Therefore it is more suitable to test a new approach for the Web directly on this one to better appreciate the contribution and the effectiveness of the work.

So in a first testing step, we devised the used queries into three sets as follows:

- A set of single keyword queries,
- A set of two keywords queries,
- A set of three keywords queries.

Each set contains 5 different queries for 5 different topics. Using these 3 queries sets, we varied the score function coefficients as follows:

- Coefficient $\boldsymbol{C_1}$ between 0.5 and 3,
- Coefficient $\boldsymbol{C_2}$ between 0.5 and 3,
- Coefficient $\boldsymbol{C_3}$ between 0.5 and 10.

According to the studies made every year by the *American Customer Satisfaction Index* (ACSI, 2014), in 2014, *Google* is the most satisfying search engine, followed by *Bing* and *Msn*, then *Yahoo* and finally *AOL*. So, based on these studies, we fixed search engines coefficients as follows:

- *Google*: coefficient $\boldsymbol{CC_1}$ = 0.5,
- *Bing*: coefficient $\boldsymbol{CC_2}$ = 0.3,
- *Yahoo*: coefficient $\boldsymbol{CC_3}$ = 0.2.

The following statistics are made on the 20 first results of the merged list using more than 15 coefficients variations for each query of the three queries sets. Indeed, some studies were made on the importance of the 20 first results and more precisely on the 3 first ones. According to these studies (Slingshot SEO, 2011; Goodwin, 2013), users consult mostly the 3 first results of the first results page, and rarely those of the second page.

For the first test set (with single keyword queries), Figure 2 represents, for each search engine, the average position of its 3 first results in the top 20 of the merged list.

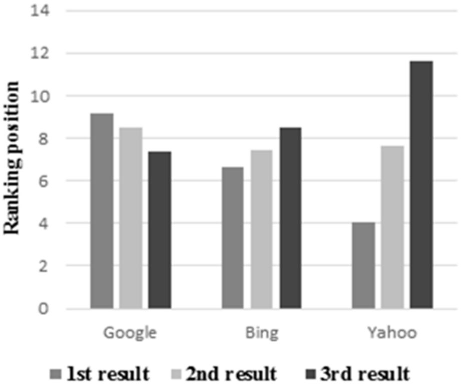

Figure 2: Average positions of the 3 first results of each search engine in the merged list for the single keyword queries set.

For the second test set (with two keywords queries), Figure 3 represents, for each search engine, the average position of its 3 first results in the top 20 of the merged list.

And for the third test set (with three keywords queries), Figure 4 represents, for each search engine, the average position of its 3 first results in the top 20 of the merged list.

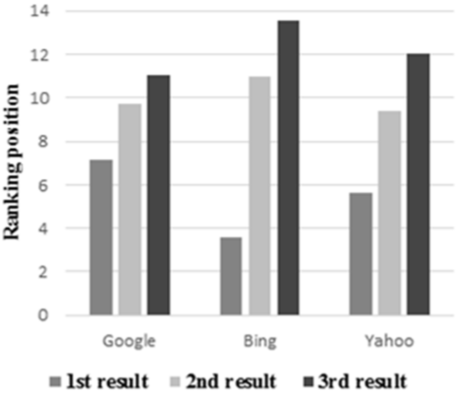

Figure 3: Average positions of the 3 first results of each search engine in the merged list for the two keywords queries set.

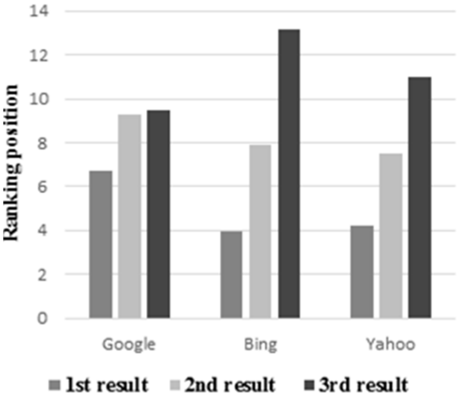

Figure 4: Average positions of the 3 first results of each search engine in the merged list for the three keywords queries set.

We can see from Figure 2, 3 and 4 that generally the three first results of each search engine rank in the top 20 of the merged list. We notice also, in the three types of queries sets, that the first result of both *Bing* and *Yahoo* is always better ranked than the first result of *Google*. Even if we have prioritized *Google* over *Bing* and *Yahoo* by giving him a coefficient of 0.5, we see that its first result is always less ranked than the first result of the others. This is due to *Google's* ranking strategy which generally privileged commercial links over other links that are probably more relevant to the user query.

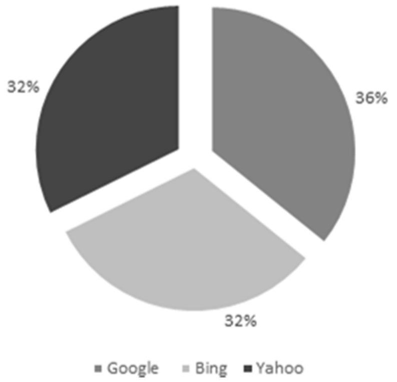

Figure 5: Average proportion of search engines participation in the top 3 of the merged list.

In Figure 5, we notice that generally each search engine has at least one result ranked in the top 3 of the merged list.

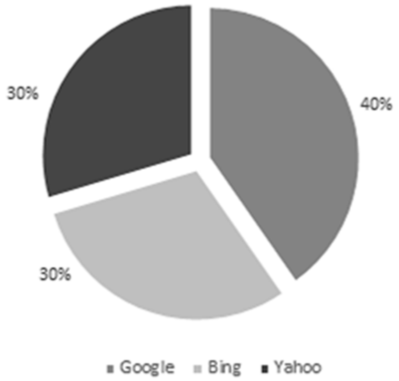

Figure 6: Average proportion of search engines participation in the top 20 of the merged list.

As we can see in Figure 6, *Google* has mostly 40% results in the top 20 of the merged list. However, Bing and Yahoo have equal proportions (30%) of participation in the top 20 of the merged list. Even if we have given *Google* the highest coefficient "***CCj***" (i.e. 0.5), it obtained less than 50% of the results in the merged list, and *Bing* and *Yahoo* have obtained equal parts of results (30%) in the merged list.

After testing different coefficients combinations, we identified the combination that seems the most suitable and that ranks the first result of each search engine in the top 3 of the merged list (1st : *Google*, 2nd : *Bing* and 3rd : *Yahoo*):

- Coefficient C_1 = 1.5,
- Coefficient C_2 = 0.5,
- Coefficient C_3 = 9.

This is due to the coefficient C_1 that gives more importance to the exact user query than the keywords it contains and most search engines employs this kind of strategy in ranking their results, i.e., they usually look for the exact query occurrence in the document before looking for the query keywords occurrence.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented a novel approach of result merging strategy that combines two techniques: computing a similarity score for each retrieved result using title, description and local rank instead of the full document; and including users' satisfaction toward the used search engines in the computation of the final ranking scores.

According to the experimental results, we can see that the system produces a well merged list where the participation of the three used search engines reflects well the users' satisfaction we introduced into the score function.

Also through the experimentations, we noticed that the top 3 of each search engine are always ranked in the top 20 of the merged list.

Even if the preliminary results we obtained are satisfying, this work is a first proposition in multiple search engines querying and needs some improvements and further experimentations. Indeed, as future work, we prospect to integrate in the score computation more information about the user information needs to have a ranking that best matches his needs. This information can be taken from a user profile or user interests for example.

We plan also to test our Metasearch engine on a user community to obtain more exhaustive results.

This Meta Search engine will be part of a personalized information retrieval system which main goal is to get the most relevant documents to the user information needs. First of all, the system will build the user profile and then will use it to reformulate user's queries in order to get the most relevant results to his needs. So, using the metasearch engine, the system will be able to cover a large proportion of documents from the Web and thus will return more relevant documents to the user.

REFERENCES

Meng, W., 2008. Metasearch Engines, Department of Computer Science, State University of New York at Binghamton.

Renda, M. E., Straccia, U., 2003. Web Metasearch: Rank vs. Score based Rank Aggregation Methods.

Lu, Y., Meng, W., Shu, L., Yu, C., Liu, K., 2005. Evaluation of Result Merging Strategies for Metasearch Engines In *6th International Conference on Web Information Systems Engineering (WISE Conference)*, New York.

Aslam, J., Montague, M., 2001. Models for Metasearch In *ACM SIGIR Conference*, pp.276-284.

Callan, J., Lu, Z., Croft, 1995. W. Searching Distributed Collections with Inference Networks In *ACM SIGIR Conference*, pp. 21-28.

Gauch, S., Wang, G., Gomez, M., 1996. ProFusion: Intelligent Fusion from Multiple, Distributed Search Engines In *Journal of Universal Computer Science*, 2(9), pp.637-649.

Rasolofo, Hawking, Y. D., Savoy, J., 2003. Result Merging Strategies for a Current News Metasearcher In *Inf. Process. Manage*, 39(4), pp.581-609.

Jadidoleslamy, H., 2012. Search Result Merging and Ranking Strategies in Meta-Search Engines: A Survey In *International Journal of Computer Science Issues*, Vol. 9, Issue 4, No 3, p. 239-251.

The American Customer Satisfaction Index (ACSI), http:// http://www.theacsi.org/the-american-customer-satisfaction-index, (Accessed 15 June 2015).

Slingshot SEO, 2011. A Tale of Two Studies: Establishing Google & Bing Click-Through Rates, *Slingshot SEO*, http://www.slingshotseo.com/wp-content/uploads/2011/10/Google-vs-Bing-CTR-Study-2011.pdf, (Accessed 8 July 2014).

Goodwin, D., 2011. Top Google Results Gets 36.4% of Clicks [Study], http://searchenginewatch.com/article/2049695/Top-Google-Result-Gets-36.4-of-Clicks-Study, (Accessed 24 October 2013).