# Analysis of Meta-Search engines using the Meta-Meta-Search tool SSIR

Manoj M
Senior Research Fellow
Computational Modelling and Simulation Unit
National Institute for Interdisciplinary Science & Technology (NIIST) (CSIR), Trivandrum - 695019. India

Elizabeth Jacob
Scientist
Computational Modelling and Simulation Unit
National Institute for Interdisciplinary Science & Technology (NIIST) (CSIR), Trivandrum - 695019. India

## ABSTRACT
Numerous information retrieval tools like Search Engines, Web Directories and deep-web search portals exist. Meta-search engines (MSE) developed over the past years claim to automatically and simultaneously search many other such information retrieval tools and improvise the fused results.

This paper acquaints SSIR (http://www.ssir.in) a tier-three Meta-Search Engine or a Meta-meta search tool. Upon receiving a query, SSIR passes the modified query to various Meta-Search Engines in parallel, collects and processes the results and passes it to the user like any other Meta-Search Engine. Currently it is hosted as a live prototype and various improvements are being added.

Apart from being a public software product used for internet searching, it is being used as a tool for analysis of overlaps and uniqueness among search results of various MSEs. The study also analyses the percentage of contribution in the top results of popular search engines Google, Yahoo and Windows-Live in the top results covered by individual MSEs and by SSIR.

## Categories and Subject Descriptors
H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval – *Search process.*

## General Terms
Algorithms, Experimentation, Design.

## Keywords
Meta-Meta-Search, overlap, unique results.

## 1. INTRODUCTION
Most web users are aware of various popular search engines such as Google, Yahoo and Live-Search (Bing). Apart from these search engines many other tools such as web-directories, deep-web search portals and Meta-Search engines exist for Information Retrieval.

The concept of Meta-Search Engine (MSE) was introduced to rectify some of the limitations of other search tools. Studies have consistently shown that even with automatic indexing, the amount of web indexed by search engines is very small and there is little overlap between the different search indexes [1, 2, 3, 4, 5]. Other sources like web-directories [6] are small, but have very good precision, as they are indexed by humans.

It can also be seen that MSEs [7] use various techniques to improve the results and give more user interface and result processing options, like clustering.

This paper gives the architecture of the Meta-Search Engine SSIR (***S***SIR the ***S***uperior ***I***nformation ***R***etriever). The tool itself is then used to analyse how much overlapping and uniqueness is there between MSE results and their propensity when we consider more results from MSEs and also more number of MSEs. What percentage of top results of popular Search Engines (SEs) is present in MSEs is analysed with different constraints. The same is done for fused MSE results (i.e. SSIR) to examine whether coverage of SEs increases with tier-three fusion.

The results help us to reason if it is worth including more results or more MSEs in a tier-three MSE like SSIR. Including SEs to form a hybrid of tier-two and tier-three search could be a new direction to improve search.

## 2. GENERAL ARCHITECTURE OF SSIR
Logically SSIR is a typical Meta-Search Engine. Being a research prototype it is being constantly modified to provide better search results. It is publicly accessible through the domain http://www.ssir.in.

Like the functioning of any MSE, SSIR sends modified user requests to several search sources in real time and in parallel. It then aggregates the results, extracts required information from each search result page into a single list and displays them in a uniform manner, eliminating duplicates, and sorting them according to relevance of the result documents. The major difference here is that, SSIR uses MSEs as search sources, making it a tier-three meta–meta search tool. As MSEs use various methods to improve results [7], inclusion of such tier-two search helps to use many of the existing improvements, thereby not having to re-invent the wheel.

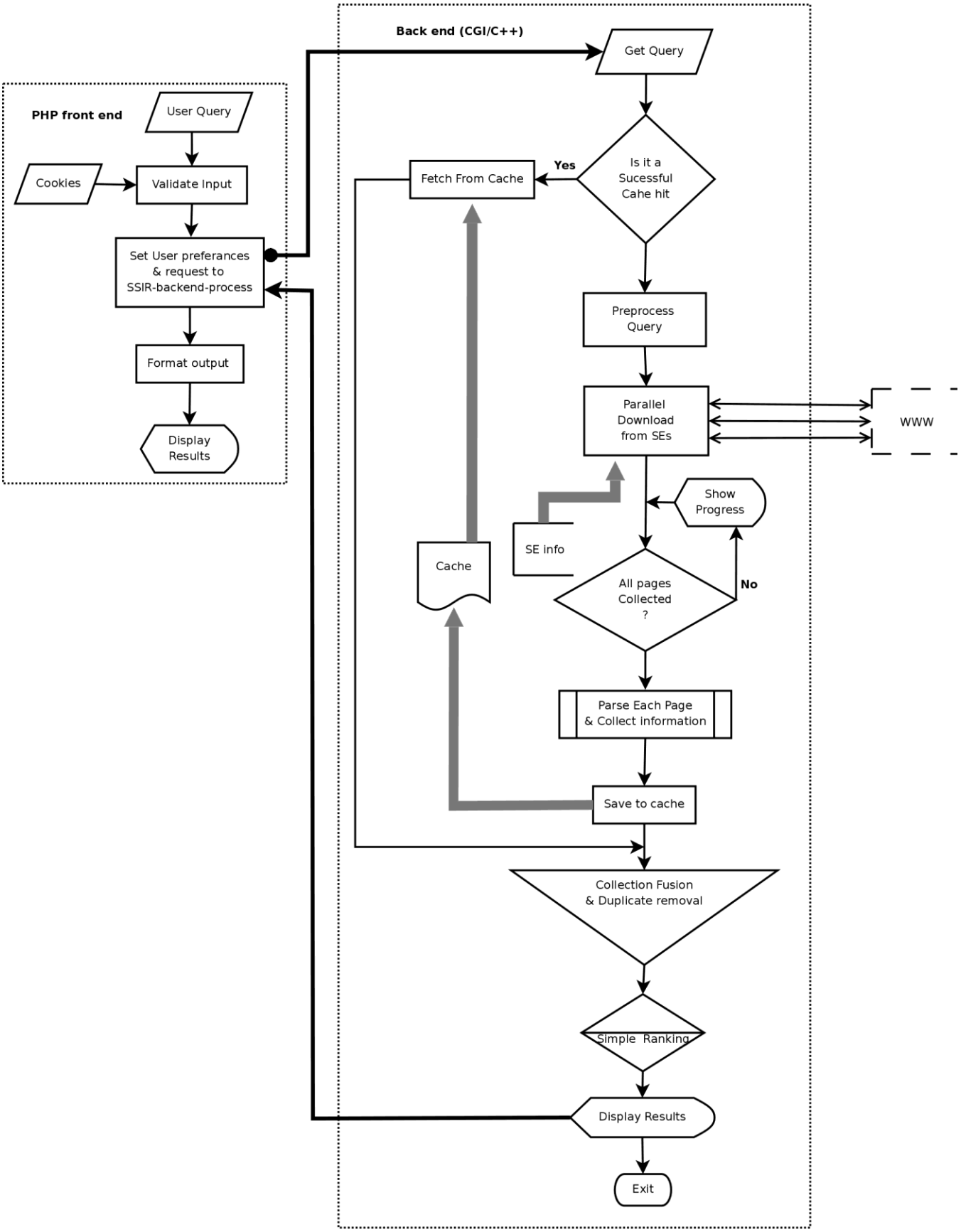



**Figure 1. Logical flowchart of the SSIR system.**

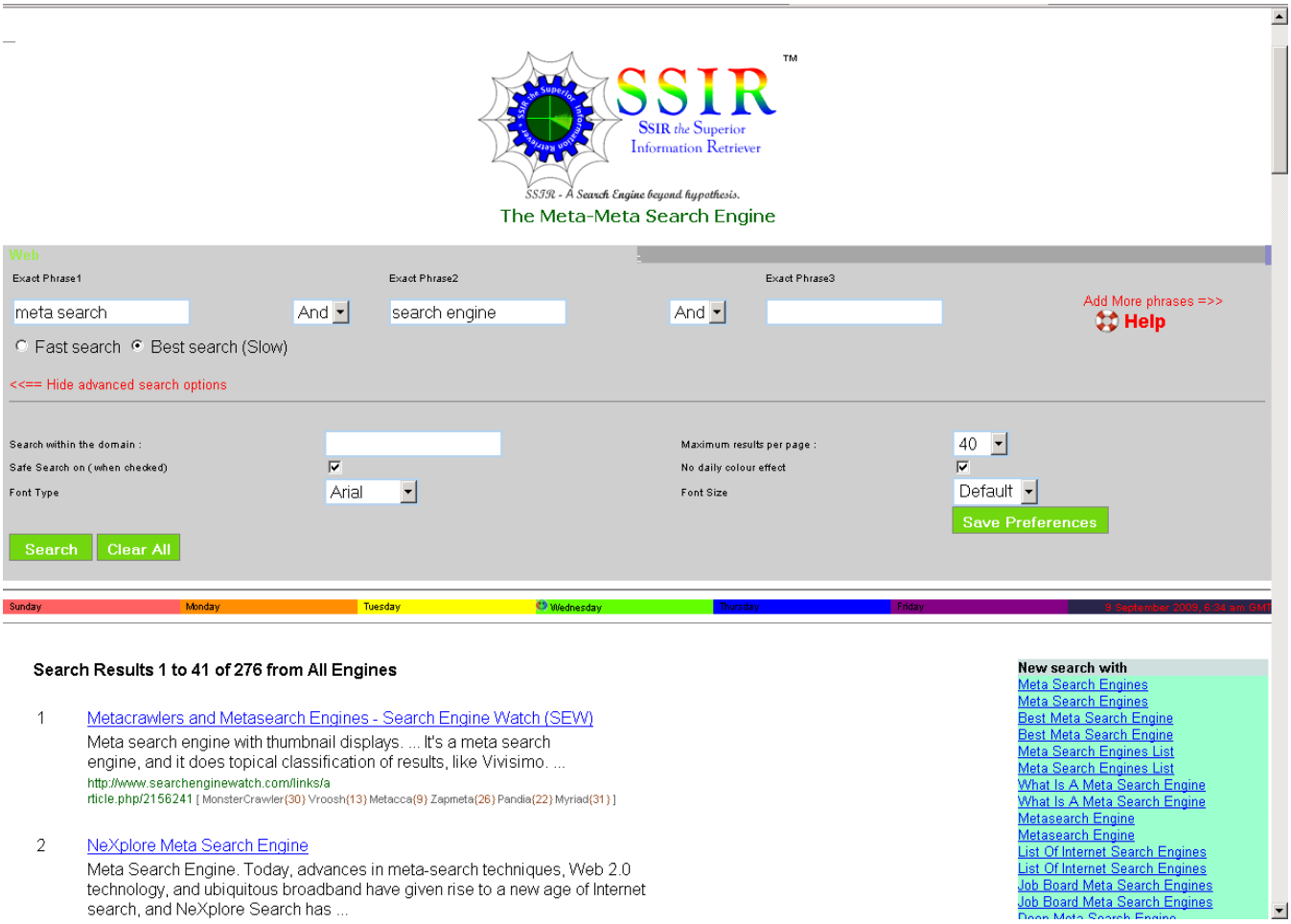



**Figure 2. The user interface web page with results of SSIR for the query "meta search" And "search engine"**

The SSIR system follows two-tier architecture (see Figure 1), a PHP-based user interface handler at the front-end and a CGI process at the back-end which handles the actual collection fusion (written in standard C++) and is not directly accessible to the end user.

## 2.1 User interface and search results

SSIR has a web interface (see Figure 2) similar to most SEs and MSEs. Instead of giving special query syntax as with many search tools, in SSIR users have better search interface.

User can give single keywords or exact phrases directly in each search box without any special syntax like quotes. Several such phrases or single-words can be combined using 'And', 'Or' or 'Not' by directly choosing from a dropdown list. User can use up to maximum of 10 such words or phrases. Options to select 'fast' or 'best' search is available as radio button. "Fast" option gets minimum results from each engine with a timeout of 10 seconds whereas "best" option is for users with patience who can get maximum and best available results. This was added based on volunteer feed back during testing.

The interface offers advanced options to search within a domain, to set the number of results per page, to select various font-types and sizes. Most of these aesthetic options are handled in PHP front-end using HTML tags and CSS-styles. These preferences can also be saved using cookies.

On receiving a search request, the query as well as user customization from cookies (if available) are read and validated. Based on it, a modified query is constructed and given to the back-end. Raw results from the back-end are formatted and sent to the user. During querying, search progress at the back-end is reported to the user at the front-end.

## 2.2 The backend processing

Once a query is received by the back-end, it is checked for *cache hit*. Un-fused results from each search engine are stored in cache for ~10 minutes. If cache hit result is found then results from cache are fetched, else query is modified accordingly for each engine and are sent to them in parallel. During the collection of results from the engines the progress is reported using JavaScript. Collected pages are processed and results like URL, title and page-description are extracted. These are saved to cache. The results are fused and duplicates are removed. The

current system employs a simple ranking based on the number of occurrences of the URL, [8] based on the fact that a URL coming from multiple algorithms may be worth more and less susceptible to search engine spamming.

SSIR can be configured through a web-based interface (only accessible through localhost for security purpose), directly embedded in the C++ back-end. Information regarding search sources, extracting data from result pages, URLs and other information for accessing various search result pages are added, manipulated and removed using this interface.

## 3. Analysis of Meta-Search Engine results using SSIR

For analysis purpose, the SSIR core MSE has been slightly modified. The web-based front-end is replaced by an automatic querying program. A PHP script based on yahoo-API is used to interface with Yahoo, while Scroogle[#1] is used instead of Google for technical reasons. The correctness of routines is verified by manually counting search pages with few results. The names Windows-live and Bing as well as SSIR results and fused results are used as aliases in this document.

Five MSEs Vroosh, Metacca, Zapmeta, Pandia and Myriad were selected for this study based on the maximum number of results in a single page and ease with which MSEs can be interfaced.

### 3.1 Configuration and Constraints of the experiment

Google Hot Trends[#2] display the top 100 hot search-terms of the past hour in the United States. Top Yahoo buzz searches[#3] displays around top 20 searches on a given day from the data collected from Yahoo search log files. This data colleted on two different days when combined formed around 250 queries. Single phrase user search is simulated with the collected 250 queries, using a C++ program, replacing the PHP front-end in the SSIR system (see Figure 1).

Only fully successful final results are used. The result data consists of successful results from around 134 unique queries. The sources consist of three popular search engines and five Meta-Search engines listed in the following table. The maximum results available from a single page are requested. The average response along with 95% Confidence Interval (C.I.) is shown in the table (see Table 1).

The use log-data and data saved in actual SSIR cache (disabling cache flushing) are used to fuse and analyse the raw results from SEs and MSEs. All results using different top results cut-off limits are based on identical data sets. A program to fuse and analyse data is created. The SE and MSE fused results with various combinations are generated. The combined statistical data from 134 results are manually calculated (using *OpenOffice calc*) and graphs are plotted (using *qtiplot*). Confidence intervals are calculated using *online*[#4] confidence interval calculator. Error bars on the top of bars of the graphs gives C.I. with 95% confidence interval.

Simple duplicate analysis based on URL comparison is currently used, heuristics based duplicate analysis is expected to create slight variations in the results.

**Table 1. Various engines used, results requested /available and average actual results returned**

| | Maximum Requested | Average Returned with 95% C.I. |
|---|---|---|
| Google | 100 | 99.44 (±1.08) |
| Yahoo | 100 | 99.68 (±0.63) |
| Bing | 50 | 49.81 (±0.27) |
| Vroosh | ALL | 45.31 (±0.63) |
| Metacca | 100 | 99.26 (±1.18) |
| Zapmeta | 40 | 40 (0) |
| Pandia | ALL | 61.11 (±3.29) |
| Myriad | ALL | 151 (0) |

### 3.2 Results based on sample queries

All available results and results with different cut-offs of top 50, 75 and 100 results are fused and the number of unique results in them are calculated. When 2, 3, 4 and all 5 MSEs are combined, their unique contribution in fused results is calculated.

Number of top results from Google, Yahoo and Windows-Live covered in three MSEs, Metacca, Myriad and Pandia and the same in fused MSE results of SSIR are also analysed.

#### *3.2.1 Amount of unique results and overlaps based on sample queries*

Top 50 results (or maximum available in the dataset) from the 5 MSEs are fused. Figure 3 shows the contributions of each MSE in the fusion. For example, from the Metacca and fusion/ SSIR bars, it can be seen that 35% of fusion consists of Metacca results of which 15% are unique to Metacca. The results are almost equally distributed among the 5 MSEs and the percentage of unique results is about 50% from each MSE. 65% of results in the fusion are unique whereas 35% consists of overlaps.

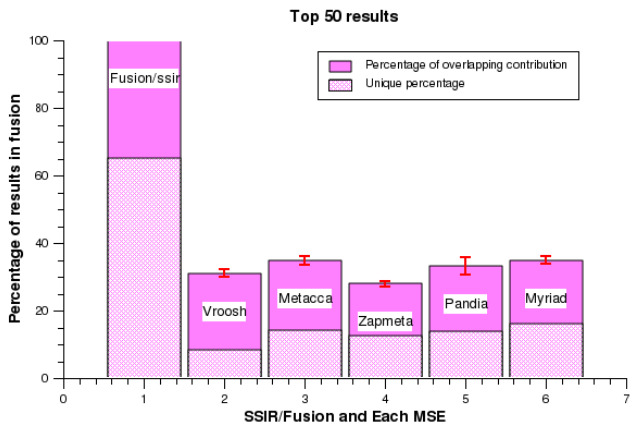
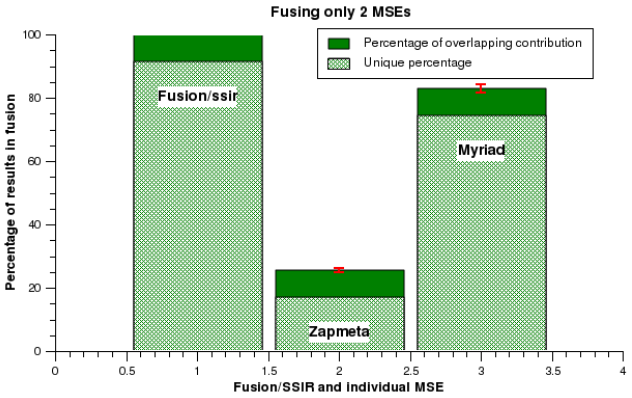



**Figure 3. Percentage of unique results and overlaps in each MSE and in fusion when top 50 MSE results are considered.**

**Figure 4. Percentage of unique results from 2 MSEs Zapmeta and Myriad and in fusion when all available results from both MSEs are considered.**



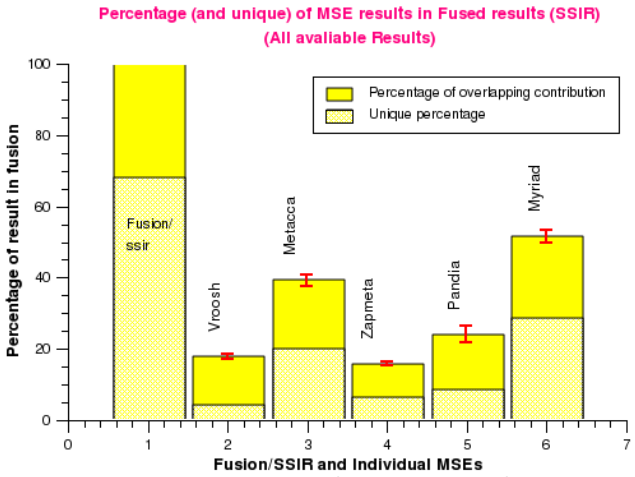




**Figure 5. Percentage of unique results from each MSE and in fusion when all available results from all five MSEs are considered.**

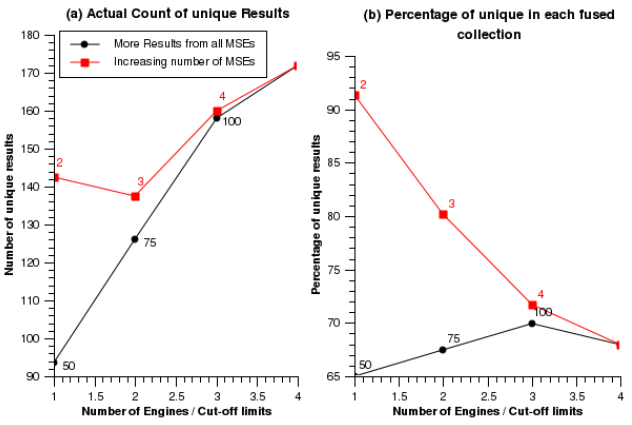



**Figure 6. (a) Number of results when more results from MSEs are considered and when number of MSEs are increased. (b) The percentage in corresponding total.**

Average total results in the fusion is around 144 (see Table 2). The results are same when all available results from the 5 MSEs are considered (see Figure 5).

When we increase the input set of three MSEs to top 75 results (i.e. equivalent to considering top 75 results from each MSE if available; and at this point it is typically valid for Pandia, Myriad and Metacca) there is a slight increase in percentage of unique results in corresponding fused result. Also there is an increase in total results after duplicate elimination, so the total numbers of unique increases. This trend is continued when we take more results even from two or three MSEs results (see Table 2, Figure 6).

It can be seen that there is a decrease in unique results percentage when all available results are considered than when top hundred results are taken. This is because only the MSE (see Table 1) Myriad provides results greater than 100.

**Table 2. Percentage of unique results corresponding to different cut-offs.**

| | Percentage of unique results (p) | Average total results in fusion(t) | Average actual unique results ((p/100)*t) | No of results Processed (cut-off) |
|---|---|---|---|---|
| 1 | 65.03 | 144 | 93.6432 | 50 |
| 2 | 67.47 | 187 | 126.1689 | 75 |
| 3 | 69.96 | 226 | 158.1096 | 100 |
| 4 | 68 | 253 | 172.04 | All available |

**Table 3. Percentage of unique results corresponding to various MSEs.**

| | Percentage of unique results (p) | Average total results in fusion (t) | Average unique results ((p/100)*t) | Engines used |
|---|---|---|---|---|
| 1 | 91.34 | 156.1 | 142.58174 | Pandia, Myriad (2) |
| 2 | 80.15 | 171.53 | 137.481295 | Zapmeta, Pandia, Myriad (3) |
| 3 | 71.68 | 223.22 | 160.004096 | Metacca, Zapmeta, Pandia, Myriad (4) |
| 4 | 68 | 252.91 | 171.9788 | Vroosh, Metacca, Zapmeta, Pandia, Myriad (5) |

When we take two meta engines (see Figure 4) and fuse them the result contains around 90% of unique results. But when we increase the number of MSEs the percentage of unique results in the fusion decreases. This trend continues for more engines. But since the total number of results increases, there is an increase in the actual number of unique results (see Table 3, Figure 6).

**Table 4. Percentage of top Results from popular SEs in MSEs and fused result (5 MSEs) when considering 50, 100 and all available**

| All results | | | Top 100 results | | | Top 50 results | | |
|---|---|---|---|---|---|---|---|---|
| Top 40 of SE | Top 20 of SE | Top 10 of SE | Top 40 of SE | Top 20 of SE | Top 10 of SE | Top 40 of SE | Top 20 of SE | Top 10 of SE |
| 23.58(± | 30.41(±2.89 | 36.94(±3.74 | 23.58(±2.55 | 30.41(±2.89 | 36.94(±3.74 | 17.46(±1.69 | 23.36(±2.38 | 28.81(±3.51 |
| 48.94(± | 50.63(±3.31 | 52.46(±3.57 | 48.94(±3.08 | 50.63(±3.31 | 52.46(±3.57 | 44.83(±3.17 | 47.05(±3.3) | 48.13(±3.55 |
| 15.88(± | 20.49(±2.25 | 24.33(±2.81 | 15.88(±1.6) | 20.49(±2.25 | 24.33(±2.81 | 12.46(±1.36 | 16.53(±2) | 19.93( |
| 26.04(± | 33.66(±2.32 | 39.78 | 23.96(±1.56 | 31.98(±2.27 | 38.06(±3.31 | 18.36(±1.28 | 26.9(±2.01) | 34.18 |
| 58.96(± | 58.66(±2.9) | 57.91(±3.14 | 56.27(±2.39 | 58.62(±2.9) | 57.84(±3.14 | 32.28(±1.74 | 54.59(±2.92 | 57.01 |
| 38.92(± | 45.34(±2.97 | 46.19(±3.48 | 34.24(±2.38 | 41.87(±2.9) | 44.18(±3.35 | 22.46(±1.76 | 32.09(±2.62 | 38.13 |
| 18.43(± | 26.12(±2.49 | 34.4(±3.58) | 18.43(±1.68 | 26.12(±2.49 | 34.4(±3.58) | 16.46(±1.54 | 24.44(±2.39 | 32.69(±3.45 |
| 70.49(± | 74.93(±4.44 | 76.72(±4.54 | 70.49(±4.37 | 74.93(±4.44 | 76.72(±4.54 | 60.3(±3.96) | 74.1(±4.4) | 76.72 |
| 16.62(± | 23.17(±2.56 | 30.75(±3.56 | 16.62(±1.73 | 23.17(±2.56 | 30.75(±3.56 | 14.85(±1.57 | 21.83(±2.46 | 30.15 |
| 72.41(± | 79.96(±1.71 | 80.9(±2.24) | 72.07(±1.4) | 79.63(±1.73 | 80.52(±2.25 | 69.12(±1.29 | 77.65(±1.8) | 78.73(±2.33 |
| 84.83(± | 88.02(±2.1) | 89.18(±2.23 | 84.46(±2.05 | 88.02(±2.1) | 89.18(±2.23 | 78.66(±2.19 | 87.46(±2.11 | 88.96 |
| 57.18(± | 70.34(±2.71 | 78.58(±2.85 | 54.16(±2.33 | 68.25(±2.78 | 77.76(±2.85 | 45.84(±1.97 | 62.16(±2.71 | 74.55(±2.99 |

**Table 5. Percentage of top Results from popular SEs in fused result when considering 2,3, 4 MSEs for fusion (i.e. ssir)**

| All results from 4 MSEs | | | All results from 3 MSEs | | | All results from 2 MSEs | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Top 40 of SE | Top 20 of SE | Top 10 of SE | Top 40 of SE | Top 20 of SE | Top 10 of SE | Top 40 of SE | Top 20 of SE | Top 10 of SE | | |
| 55.49(±1.85) | 72.09(±1.82) | 77.99(±2.39) | 40.11(±2.29) | 50.67(±2.58) | 60.67(±3.44) | 32.82(±2.46) | 40.04(±2.78) | 46.19(±3.79) | Google_in_Metacca | |
| 83.86(±2.13) | 86.75(±2.24) | 88.28(±2.27) | 82.39(±2.35) | 84.78(±2.52) | 85.75(±2.62) | 61.55(±2.39) | 60.97(±2.9) | 60.15(±3.13) | Yahoo_in_Metacca | |
| 56.29(±2.38) | 69.66(±2.73) | 77.84(±2.86) | 46.66(±2.46) | 56.16(±2.89) | 61.34(±3.42) | 40.09(±2.57) | 46.34(±3.03) | 47.01(±3.52) | Bing_in_Metacca | |
| | | | | | | | | | Google_in_Myriad | |
| | | | | | | | | | Yahoo_in_Myriad | |
| | | | | | | | | | Bing_in_Myriad | |
| | | | | | | | | | Google_in_Pandia | |
| | | | | | | | | | Yahoo_in_Pandia | |
| | | | | | | | | | Bing_in_Pandia | |
| | | | | | | | | | Google_in_Ssir | |
| | | | | | | | | | Yahoo_in_Ssir | |
| | | | | | | | | | Bing_in_Ssir | |

| | | | Google_in_Ssir | Yahoo_in_Ssir | Bing_in_Ssir |
|---|---|---|---|---|---|

### *3.2.2 Coverage of SEs in MSE top results*

The percentage of results in top 10, 20 and 40 results of three SEs google, yahoo and bing that also occur in Top 50, 100 and all requested results (see Table 1) of Meta-Search engines, Metacca, Myriad and Pandia taken separately and the same in fusion of all 5 engines under study has been tabulated (see Table 4). Also results of fusion with 2, 3 and 4 Meta-Search engines have been tabulated (see Table 5).

Around 30% (average of 28.8, 34.2, 32.7 in last column of Table 4) i.e. 3 results out of top 10 results of Google appear in top 50 results of the Meta-Search engines. Contribution of Yahoo is higher than other search engines. Around 60% of top 10 and top 20 results from yahoo are covered by Meta-Search engines, while the same from engines Google and Bing are much smaller. This increases to around 8 out of 10 (78.73%) results when 5 engines are fused in SSIR. It can be seen that there is much larger percentage of results from SEs in fused SSIR MSE results (last 3 rows of Table 4) confirming the fact that tier-three meta-search achieves better coverage of SEs.

When 2 MSEs are fused about 50% (5 out of top 10) of Google results are covered in the fusion. This figure jumps to 60% for fusion with 3 engines and 78% when 4 engines are combined. The same trend is observed with yahoo and bing (see Table 5). Though there is increase in the coverage of SEs in MSEs when number of results is increased from 50 to 100 (see Table 4), adding more MSEs (see Table 5) to the fusion yields far better results.

***Note***: The top 100 and all available results for Metacca and Pandia are same since the average available results are less than or equal to 100 (see Table 1) for both these engines.

## 4. CONCLUSION

Using SSIR, a tier-three meta–meta search tool, some interesting general conclusions about fusion can be made. Fusing meta-search engines has an additive effect as far as uniqueness of results is concerned making tier-three fusion worthwhile. Taking more results from multiple MSEs as well as increasing number of MSEs in fusion are both beneficial. The percentage of unique results in fused result is found to be decreasing with addition of more engines but the actual number of unique results shows an increase. So addition of MSEs in fusion must be done judiciously as it requires more resources.

The coverage of SEs in fusion increases with the number of MSEs combined. This trend suggests that it may be better to use more MSEs than to increase the number of results per MSE for increased SE coverage.

It can be concluded that to increase coverage of web search, for better utilization of resources and better performance, a hybrid fusion of tier-two and tier-three MSEs with fewer results from each Search/Meta-Search engine may yield useful search results with more optimal utilisation of resources.

## 5. FUTURE WORK

With the two-tier SSIR architecture, reverse-proxy like configuration can be easily incorporated. Then the system becomes scalable, and facilities like load balancing can be easily added.

Relevance of documents is also a very important factor in online information retrieval. Further research has to be focussed on improving the ranking of results in fused data. Many researches have reported that combining retrieved results has improved results through effects like chorus effect, skimming effect and dark horse effect [8]. Our primary observation of fused results also draws to similar conclusion, though further analysis is required for a more conclusive assertion. Our simple ranking scheme is based on total frequency of occurrence of result URLs. Based on combination effect an improved ranking algorithm for MSEs can be developed. Apart from collection fusion, this has a much wider application such as in SE ranking, to improve relevance and reduce SE spamming using data fusion.

## 6. ACKNOWLEDGMENTS

This research is a part of Ph.D. programme supported by the CSIR (Council of Scientific and Industrial Research), India by means of Senior Research Fellowship. The authors acknowledge the use of facilities at National Institute for Interdisciplinary Science and Technology, Trivandrum, a constituent laboratory of CSIR where this research has been carried out.

## 7. REFERENCES

[1] Lawrence, S. and Giles, C. L. 1998. Searching the World Wide Web. Science. 280, 5360(1998), 98-100.

[2] Lawrence, S. and Giles, C. L. 1999. Accessibility of information on the web. Nature. 400, 6740(1999), 107-109.

[3] Bharat, K. and Broder, A. 1998. A technique for measuring the relative size and overlap of public Web search engines. In Proceedings of the seventh international conference on World Wide Web 7 (Brisbane, Australia, April 1998), 379-388.

[4] Sander-Beuermann, W. and Schomburg, M. 1998. Internet Information Retrieval: The Further Development of Meta-Searchengine Technology. In Proceedings of the 1998 Internet Summit (Internet Society, July 22-24, 1998).

[5] Gulli A. and Signorini A. 2005. The indexable Web is more than 11.5 billion pages. In Poster proceedings of the 14th international conference on World Wide Web (Chiba, Japan, 2005). ACM Press, 902–903.

[6] Geniac. 2004. ODP and Yahoo Size Charts. Geniac.net. (Jan 2004). http://www.geniac.net/odp/.

[7] Manoj M. and Jacob E. 2008. Information retrieval on Internet using meta-search engines: A review, J. Sci. Ind. Res. 67, 10(2008), 739-746.

[8] Nassar M. O. and Kanaan G. 2009. fCombMNZ: An Improved Data Fusion Algorithm. In Proceedings of

International Conference on Information Management and Engineering (2009), icime. 461-464.

## 8. URLS

[#1] http://www.scroogle.org

[#2] http://www.google.com/trends/hottrends.

[#3] http://buzzlog.buzz.yahoo.com/overall/

[#4] http://dimensionresearch.com/resources/calculators/conf_means.html