www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 1 Jan. 2017, Page No. 20073-20089
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.40

Mining Web Graphs for Large Scale Meta Search Engine Results

Bandi Krishna[1] and Dr.V.B.Narasimha[2]
Research Scholar, CSE Dept, Osmania University[1]
Asst.Prof, CSE Dept, Osmania University[2]

Abstract:

Web is huge, To get efficient result s from the search engine is difficult task. Meta search engine (MSE) is search tool that sends user requests to several search engines or databases, aggregates the results, Merges or re-ranks them into a single list and displays them to users with the help of web graphs .MSE enable users to enter search query once and access several search engines simultaneously. This technique saves lot of time to the user from having to use multiple search engines separately by initiating the search at a single point. Today's Most of MSEs Employ only a small number of general purpose search engines. Building a large scale MSE using numerous specialized search engines is another area that deserves more attention. Arising challenges from building very large scale MSE includes automatic generation and Maintenance of high quality search engine representatives needed for efficient and effective search results, and highly automated Techniques to be added into MSE. In this paper, we are implementing our study on how to merge the search results returned from the multiple component search engines into a single ranked list through web graphs. Web graphs are essential for producing effective and efficient results.

Keywords:

Web graphs, Meta Search Engine, Large Scale, Implementation, Information Retrieval, and Algorithm.

I. Introduction

To the web user, a Meta Search Engine(MSE)appears much like a regular search engine(SE).MSE, unlike an SE does not have an index .Instead, it dynamically queries multiple search engines; extracts, fuses and re-ranks results and presents to users.

Meta search engines create what is know n as a virtual database. The do not compile a physical database or catalogue of the web. Instead, they take a user's request, pass it to several other heterogeneous database and then compile the result in a homogeneous manner based on specific algorithm. MetaCrawler , savvy Search and Mamma are some of the earliest Meta search engine.

Meta Search Engine (MSE) were proposed and built as data mining tools. Meta Search Engine (MSE) on internet has improved continually with application of new methodologies. Understanding and utilization of MSEs are valuable for computer scientists and researchers, for effective information retrieval.

II. Related Work

No Meta search engines are alike. Some search only the most popular search engines while others also search lesser-known engines, newsgroups, and other databases .They also differ in how the results are presented and the quantity of engines that are used. Some will list results according to database. Others return results according to relevance, often concealing which search engine returned which results. This benefits the user by eliminating duplicate hits and grouping the most relevant once at the top of the list.

Search engines frequently have different ways they expect requests submitted. For example, some search engines allow the usage of the word "AND" while others require "+" and others require only a space to combine words. The better Meta search engines try to synthesize requests appropriately when submitting them.

A Meta Search Engine acts as an agent for the participant search engine. It receives queries from users and redirects them to one or more of the participant search engines for processing the various algorithms are used for search engine selection and result merging that provides relevant information according to the user.

A. Search Engine Selector: If the number of component search engines in a metasearch engine is very small, say less than 10,it might be reasonable to send each user query submitted to the metasearch engine to all the component search engines in this case, the search engine selector is number of component search engines is large, as in the large scale meta search engine scenario, then sending each query to all component search engines will be an inefficient strategy because most component search engines will be useless with respect to any particular query. For example, suppose a user wants to find 50 best matching results for his\her query from a metasearch engine with 1000 component search engines. Since the 50best results will be contained in no more than 50 component Search engines, it is clear that at least 950 component search engines are useless for this particular query. Passing a query to useless search engines may cause serious problems for efficiency. Generally, sending a query to useless search engines will cause waste resources to the metasearch engine server, each of the involved search engine servers and the internet. Specifically, dispatching a query, including needed query reformatting, to a useless search engine and handling the returned results, including receiving the returned response pages, extracting the results record from these pages, and determining whether they should be included in the final merged results list and where they should be ranked in the merged results list of they are to be included, waste the resources of the metasearch engine server; receiving the query from the metasearch engine, evaluating the query , and returning the results bake to the metasearch engine whose results end up useless; and finally transmitting a query from the metasearch engine to useless search engines and transmitting useless retrieved results from those search engines to the meta search engine waste the network resources of the internet. Therefore, it is important to send each user query to only potentially to send search engines for processing .The problem of identifying potentially useful component search engines to invoke for given query is the search engine selection problem. Obviously, for metasearch engines, having an effective search engine selector is more important.

B. Search Engine Connectors: After a component search engine has been selected to participant in the processing of a user query, the search engine connector established a connection with the server of the search engine

and pass the query to it. Different search engines usually have different connection parameters. As a results,a separate connector is create for each search engine. In general, the connector for search engine S needs to know the HTTP (Hyper Text Transfer Protocol) connection parameters supported by S. There are three parameters,(a) the name and location of the search engine server,(b) the HTTP request method (usually it is either GET or POST)supported by S, and(c) the name of the string variable that is used to hold the actually query string. When implementing metasearch engine with a small number of component search engines, experienced developers can manually write the connector for each search engine. However, for large scale-scale metasearch engine, this can be very time consuming and expensive. Thus, it is important to develop the capability of generating connectors automatically. An intelligent metasearch engine may modify a query it receives from a user before passing it to be search engine connector if such a modification can potentially improve the search engine to add terms that are related to original user query to improve the chance for retrieving more relevant documents.

c. Web Service:

A Meta search engine is a search engine that collect results from other search engine. Web service offers such functionality and then present a summary of that information as the results of a search. Most search engines available on the web provide only a browser based interface; however, because Web services start to be successful, some of those search engines offer also an access to their information through Web services. Two types of search engines are observed, one that acts like a wrapper for the HTML pages returned by the search engine and other one is build upon the Web service offered by the search engine but this difference is visible only when looking at the internal processing of the service. Web service are built for, any process that can be integrated into external systems through valid XML documents over internet protocols. This definition outlines the general idea of Web service.Web service can be seen as software components with an interface to communicate with other software components. They have a certain functionality that is available through a special kind of Remote procedure Call. SOAP, the Simple Object Access Protocol [16] was developed to enable a communication between Web services.It was designed as a lightweight protocol for exchange of information in a decentralized, distributed as a decentralized, distributed, text-based framework for enabling communication between diverse parties that have no prior knowledge of each other.This is the requirement a transport protocol for web services has to fulfill. SOAP species a mechanism to perform remote procedure calls and therefore remove the requirements that to systems must run on the same platform or be return in the same programming language.

D. Result Extractors:

After a component search engine processes a query, the search engine will return one or more response pages. Atypical response page contain multiple (usually 10)search result records,each which corresponds to a retrieved Web page, and it typically contains the URL and the title of the page a short summary(called snippet) of the page content, and some other pieces of information such as page size.The upper portion of a response page from the Google search engine. Response pages are dynamically generated HTML documents and they often also contain content unrelated to the user query such as advertisements (sponsored links) and information about the host Web site. A program (i.e., result extractor) is needed to extract the correct search result record from different component search engines can be

merged into a single ranked list. This program is sometimes called an extraction wrapper. Since different search engines often format their results differently, a separate result extractor is usually needed for each component search engine. Although experienced programmers can write the extractors Manually , for large-scale matasearch engines, it is desirable to develop techniques that can generate the extractors automatically.

E. Result Merger:

After the results from the selected component search engines are returned to the metasearch engine, the results merger combines the results into a single ranked list. The ranked list of search result records is then presented to the user, possible 10 records on each page at a time, just like more search engines do. Many factors may influence how results merging will be performed and what the outcome will look like. The information that could be utilized includes the local rank of a result record from a component search engine, the title and the snippet of a result record, the full document of each result,the publication time of each retrieved document, the potential relevance of the search engine with respect to the query from where a result is retrieved, and more A good result merger should rank all returned results in descending order of their desirability. The existing architecture has many disadvantages in search engine selection,search engine connection and result extractors. We proposed a robust metasearch engine architecture using web services for heterogeneous and dynamic environment.

III. Proposed Work

MSE combine multiple search engines into a single unit. Here we are taking sample six search engines and combining into a one proposed work. We can also take more than six search engines .As part of search, we are giving same query to all search engines and analyzing results.

Here we are taking 50 users for consideration those 50 users will type the same query and record the traverse of users.

The search engines which we are using in MSE must be in our control. We combine all search engine results and then eliminate the duplicate links in various search engines.

Analyze the 50 users traversing data and display the results according to priority. Large scale in the sense, not only in terms of more number of search engines, but in terms of new methodologies and techniques introduced in MSE

MSEs employ only a small number of general purpose search engines. Building large-scale MSEs using numerous specialized search engines is another area that deserves more attention. Challenges arising from building very large-scale MSEs include automatic generation and maintenance of high quality search engine representatives needed for efficient and effective search engine selection, and highly automated techniques to add search engines into MSEs and to adapt to changes of search engines.

In this paper we are taking our study on how to merge the search results returned from multiple component search engines into a single ranked list; this is an important issue in MSE research. An effective and efficient result merging strategy is essential for developing effective large scale Meta Search systems.

Proposed framework for MSE is shown in Figure 1 that takes into account both ranking and clustering mechanisms for organizing and presenting web pages to the user. The whole process, from giving the user query, to getting the results are organized in the following modules.

III.I. MSE Interface for User: This module provides the way of interaction to the proposed framework. When a user gives a query to the MSE, then this query is further provided to the multiple SEs for searching the information on the web. The returned results of the SE are stored in the local database.

III.II. Similarity Score (SS): SS provides a relevancy score to each returned web page of a SE. This relevancy score is calculated for finding, to which extant the page is matching to the user query. Authors of the paper know that a relevant web page is more similar to other relevant WebPages than the irrelevant pages because it can perform better for both single term query search and for multiple term query search. Higher score indicates better matches.

III.III. Result generator (RG): RG is responsible for generating the required number of Links. Generation of results is based on the lower and upper relevancy score of the web pages provided by the RC module. It also decides the relevancy range of each cluster for which URLs to be assigned. The complete process of cluster generation is illustrated by the algorithm given in Figure 1. The generated results are purely based on the similarity rank of the retrieved web pages with web graphs.

III.IV. Web Page Adjuster (WPA): WPA is responsible for removal of duplicate web pages and assignment of ranked web pages to the corresponding cluster. Organizing the ranked results in the clusters is not meant to replace the traditional way of representing the search results with the new one. Higher the relevancy rank of the web page then higher is the possibility for the web page to be placed on the top of the cluster results. The complete process of assignment of web pages to the clusters is illustrated by the algorithm given in Figure 1.

III.V. Web Graphs: After getting the results from different multiple search engines, we have to
Apply web graphs on results which are extracted from search tool interface. Now we will get efficient results which are relevant to the user's request.

Algorithm: Efficiency Search Results through web graph (ESRW)

// Start of algorithm (Figure-1)

Step 1. Get the downloaded WP of each SE separately.

Step 2. For each WP of each SE, take top ten links that is first page of results

Step 3. If (pages are completed) // This step eliminates the duplicate links

{

Exit ()

}

Else

{

For each page Pi of downloaded WP with RS

If (Pi is already visited)

{

Eliminate the web page and go for next iteration

}

Else

{

Include the web page and go for next iteration

}

}

Step 4. Return the ordered results

Step 5. Apply web graphs on results

Step 6.Re-rank the web graphs results based on in which page accurately retrieved

Step 7. Show the Results

Step 8. Stop

Input: User query Q, downloaded web pages (WP), number of required clusters (NC).

Output: Labeled results with ranked links of web pages.

This algorithm will help to produce the efficient results. This Algorithm has been implemented and shown in the experiment analysis part

IV. Experiment Analysis

As a Part of implementation, we taken different search engines and typed the same query in all search engines.

If we type a Query in **GOOGLE** search engine,then it displays following links(top ten links)

GOOGLE		
QUERY	**Rank**	**URL**
INFOSYS	1	www.infosys.com/
	2	https://www.facebook.com/Infosys/
	3	http://www.moneycontrol.com/india/stockpricequote/computers-software/infosys/IT
	4	http://economictimes.indiatimes.com/infosys-technologies-ltd/stocks/companyid-10960.cms
	5	http://economictimes.indiatimes.com/tech/ites/working-with-ministrys-website-mca21-to-fix-glitches-infosys/articleshow/51691335.cms
	6	twitter.com/infosys
	7	http://in.reuters.com/finance/stocks/overview?symbol=INFY.NS
	8	www.infosys-science-foundation.com/
	9	https://www.linkedin.com/company/infosys
	10	money.rediff.com/companies/Infosys-Ltd/13020007

DOI: 10.18535/ijecs/v6i1.40

If we type a Query in **MSN** search engine,then it displays following links

MSN		
QUERY	**Rank**	**URL**
INFOSYS	1	www.infosys.com/
	2	https://en.wikipedia.org/wiki/Infosys
	3	http://economictimes.indiatimes.com/infosys-technologies-ltd/stocks/companyid-10960.cms
	4	https://www.infosys.com/about
	5	http://www.moneycontrol.com/india/stockpricequote/computers-software/infosys/IT
	6	https://www.linkedin.com/company/infosys
	7	money.rediff.com/companies/Infosys-Ltd/13020007
	8	www.infosys-science-foundation.com/
	9	https://www.linkedin.com/company/infosys
	10	http://www.bing.com/images/search?q=infosys&qpvt=infosys&qpvt=infosys&qpvt=infosys&FORM=IGRE

If we type a Query in **YAHOO** search engine,then it displays following links

YAHOO		
QUERY	**Rank**	**URL**
INFOSYS	1	www.infosys.com/
	2	www.infosys.com/careers/
	3	en.wikipedia.org/wiki/Infosys
	4	www.facebook.com/Infosys
	5	http://economictimes.indiatimes.com/tech/ites/working-with-ministrys-website-mca21-to-fix-glitches-infosys/articleshow/51691335.cms
	6	http://www.moneycontrol.com/india/stockpricequote/computers-software/infosys/IT
	7	http://www.livemint.com/Companies/NZOTuzB2LnxHHpk6oGgqvM/Are-Infosys-founders-not-happy-with-Vishal-Sikka.html
	8	twitter.com/infosys
	9	http://www.business-standard.com/article/markets/infosys-dips-after-block-deals-116040700201_1.html
	10	http://in.reuters.com/finance/stocks/overview?symbol=INFY.NS

If we type a Query in **BAIDU** search engine,then it displays following links

BAIDU		
QUERY	**Rank**	**URL**
INFOSYS	1	https://www.infosys.com/
	2	https://www.infosys.com/oracle/pages/index.aspx
	3	https://www.infosys.com/careers/
	4	https://en.wikipedia.org/wiki/Infosys
	5	https://www.linkedin.com/company/infosys
	6	https://www.youtube.com/user/Infosys
	7	https://twitter.com/Infosys
	8	https://www.infosysbpo.com/
	9	http://economictimes.indiatimes.com/infosys-technologies-ltd/stocks/companyid-10960.cms
	10	http://www.moneycontrol.com/india/stockpricequote/computers-software/infosys/IT

If we type a Query in **ASK.COM** search engine,then it displays following links

DOI: 10.18535/ijecs/v6i1.40

ASK.COM		
QUERY	Rank	URL
INFOSYS	1	en.wikipedia.org/wiki/Infosys
	2	www.infosys.com/
	3	www.infosys.com/about/
	4	economictimes.indiatimes.com/infosys-technologies-ltd/stocks/companyid-10960.cms
	5	www.linkedin.com/company/infosys
	6	www.moneycontrol.com/india/stockpricequote/computers-software/infosys/IT
	7	in.reuters.com/finance/stocks/overview?symbol=INFY.NS
	8	www.bloomberg.com/quote/INFO:IN
	9	www.infosys-science-foundation.com/
	10	www.infosysbpo.com/

If we type a Query in **BING** search engine,then it displays following links

BING		
QUERY	Rank	URL
INFOSYS	1	www.infosys.com/
	2	https://en.wikipedia.org/wiki/Infosys
	3	http://www.moneycontrol.com/india/stockpricequote/computers-software/infosys/IT
	4	http://economictimes.indiatimes.com/infosys-technologies-ltd/stocks/companyid-10960.cms
	5	https://www.linkedin.com/company/infosys
	6	http://money.rediff.com/companies/Infosys-Ltd/13020007
	7	https://www.infosys.com/about/
	8	https://www.infosysbpo.com/
	9	http://www.facebook.com/Infosys
	10	http://www.bing.com/images/search?q=infosys&qpvt=infosys&qpvt=infosys&qpvt=infosys&FORM=IGRE

If we type a same Query in Our Proposed **Meta Search Engine** then it displays following links

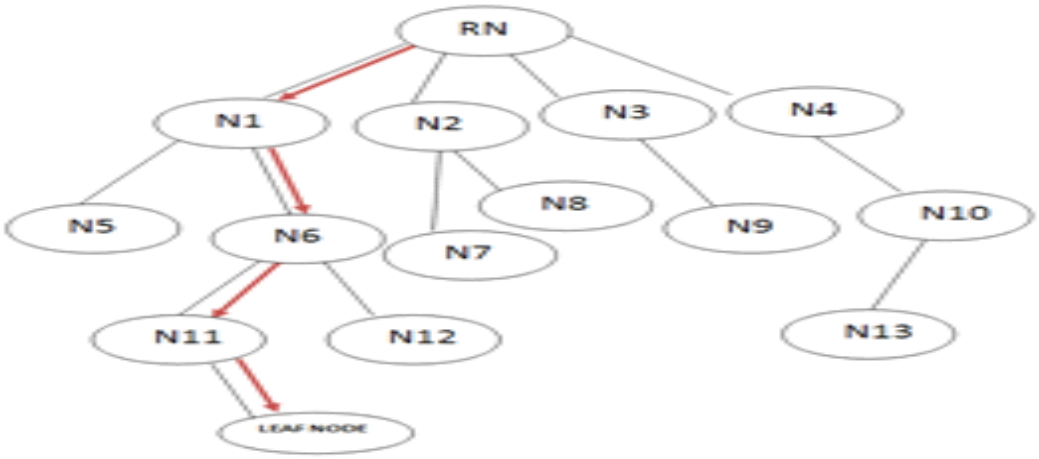
META SEARCH ENGINE		
QUERY	Rank	URL
INFOSYS	1	https://www.infosys.com/careers/
	2	en.wikipedia.org/wiki/Infosys
	3	www.infosys.com/
	4	www.infosys.com/about/
	5	www.linkedin.com/company/infosys
	6	www.moneycontrol.com/india/stockpricequote/computers-software/infosys/IT
	7	in.reuters.com/finance/stocks/overview?symbol=INFY.NS
	8	economictimes.indiatimes.com/infosys-technologies-ltd/stocks/companyid-10960.cms
	9	www.infosys-science-foundation.com/
	10	www.infosysbpo.com/

We got the above results with the help of web graphs.the INFOSYS website can be represented in the form of nodes that can be shown in the below figures

IV.I. Web Graph Representation of Top link in MSE

Every user who has entered any query for fetching required information can come with own requirements. After getting the search results ,the results may or may not be relevent to his needs. So, to navigate from Root Node(RN) to Leaf Node ,he/she has to travel from home page to his wanted page .

DOI: 10.18535/ijecs/v6i1.40



IV.II. Site Navigation of Above Graph

The user who searched "**infosys**" query to get into **jobopenings** page has to travel from **home page** to **jobopenings** page as shown below

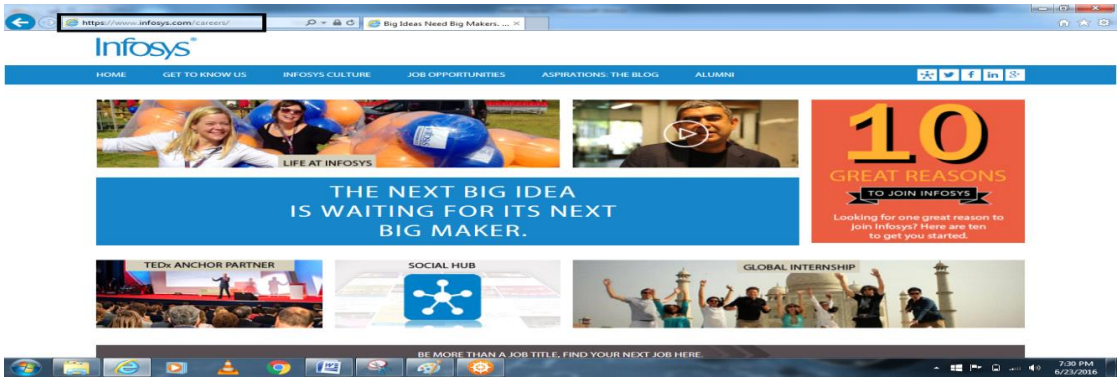
IV.III. Screen Shot Representation of Infosys(Top Link)

The following figures shows the sample site screen that depicts the actual flow of user navigation from home page to jobopenings page

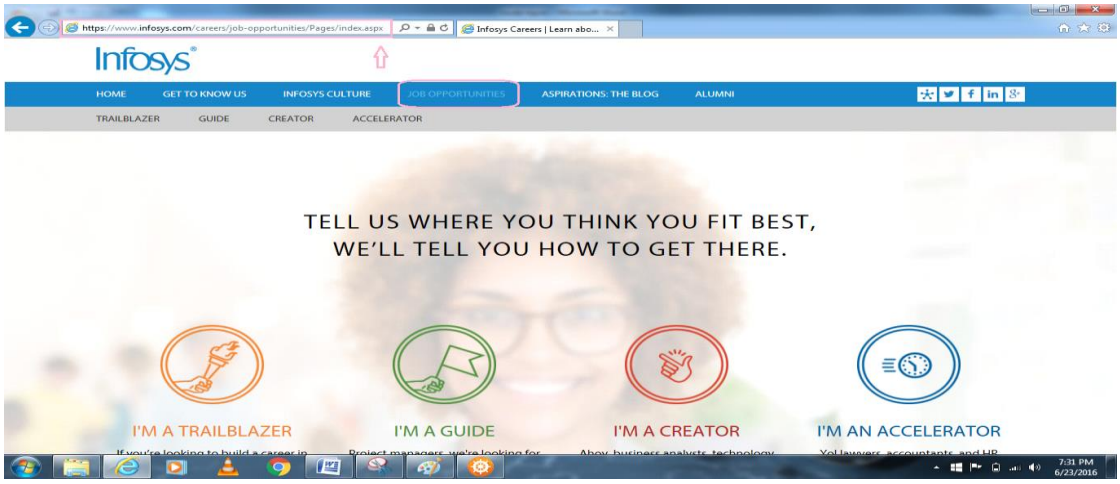
Home Page (Root Node)

DOI: 10.18535/ijecs/v6i1.40

Careers(Node 11)

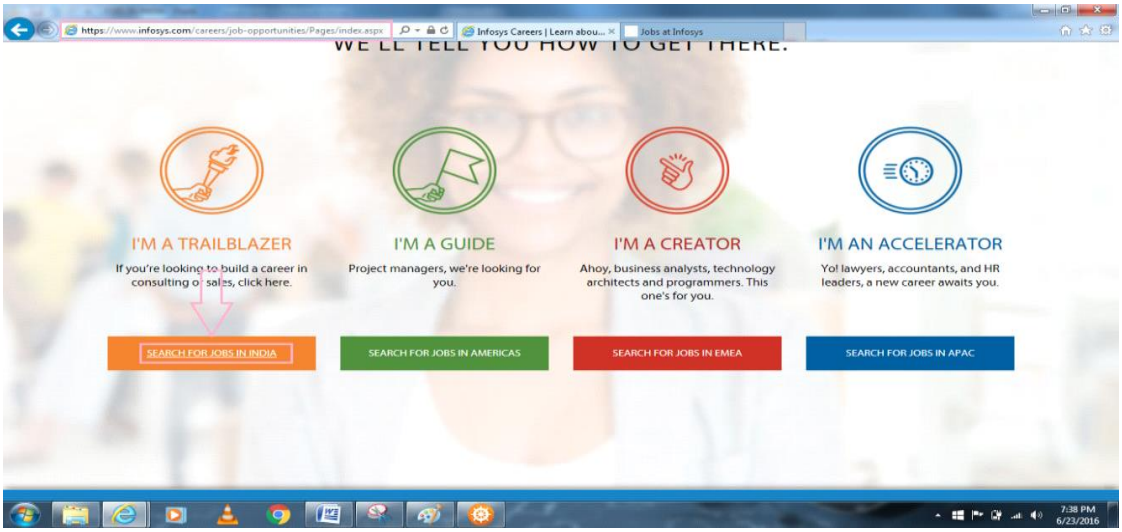


Job Opportunities(Node 6)



Select Job in India(node 11)

DOI: 10.18535/ijecs/v6i1.40



Searching job through keyword(leaf node)

So,For going into the **job openings** page unnecesarily he has to travel through 4 to 5 web pages approximately.To to get rid from this type of unwanted navigation ,can't we choose the most viewed path and place it in the top ,So that He can go to his required page instantly.Based on the number visitors for a particular page or the path that he/she travelled to reach needed page should come as the top search results.Below is the complete representation for choosing most viewed page.

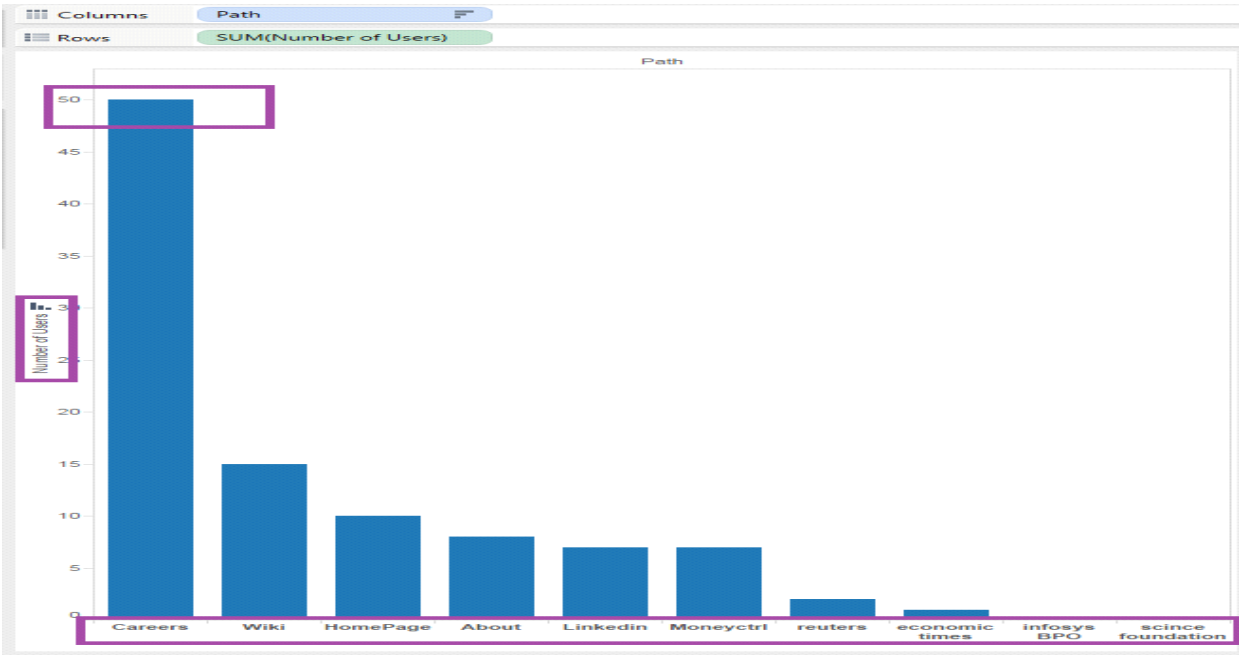
IV. Graphs Representation

Graph Representation for choosing most viewed path:

Bar Graph:

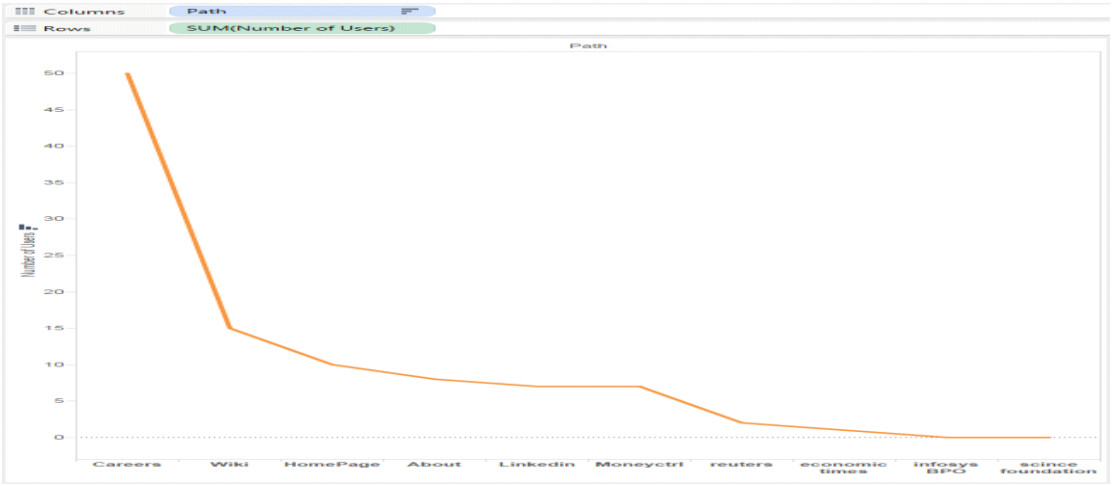
Here is comparision between the number of users and its corresponding webpage.The bar graph representation for the same is shown below

DOI: 10.18535/ijecs/v6i1.40



Line Graph:

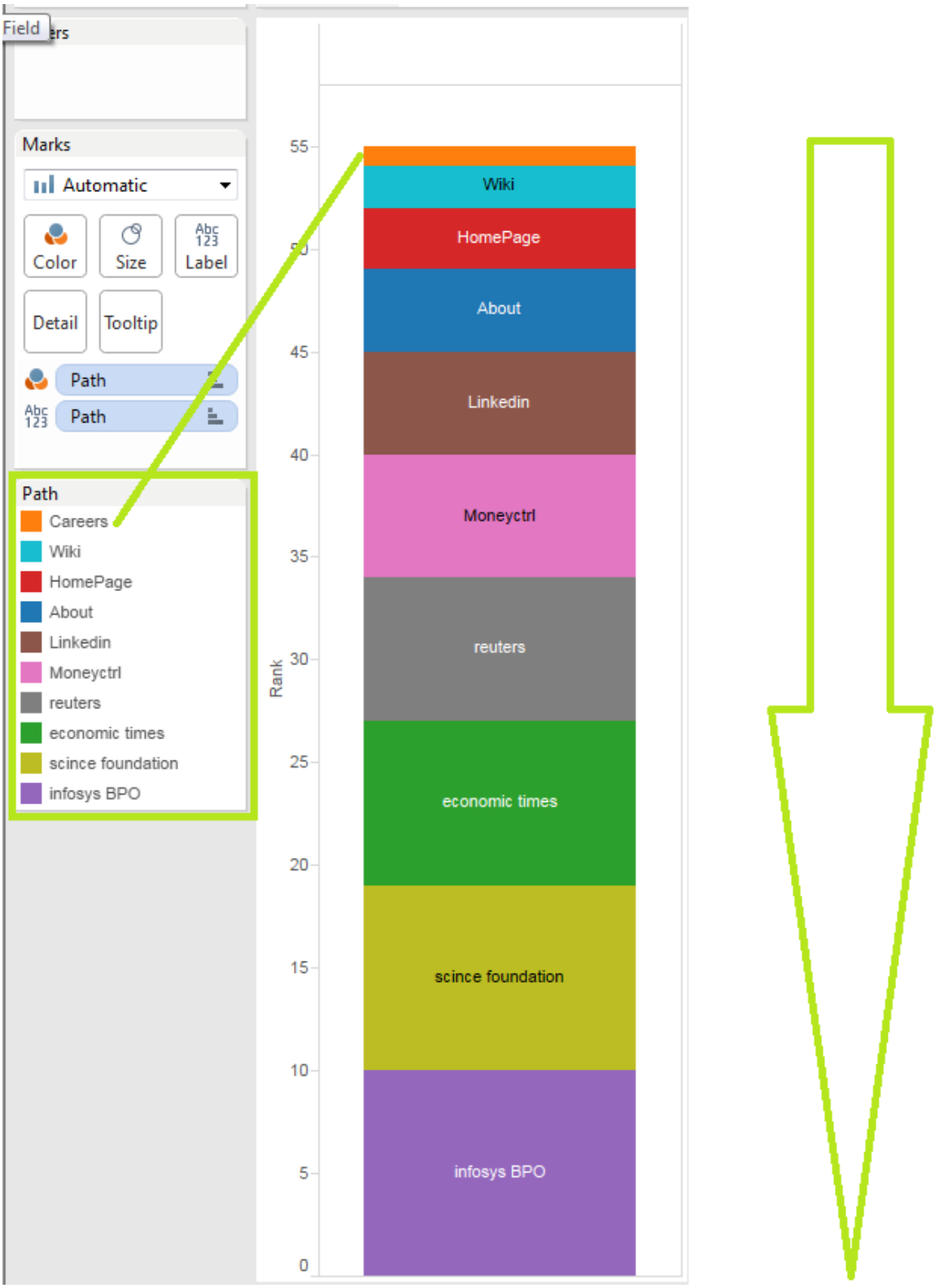
The line graph representation for the same is shown below



Ordering in sequence:

As resulted in the above two graphs ,one can decide the order of particular web links which are to be displayed when the user searched for same query now

DOI: 10.18535/ijecs/v6i1.40



Packed bubble View:

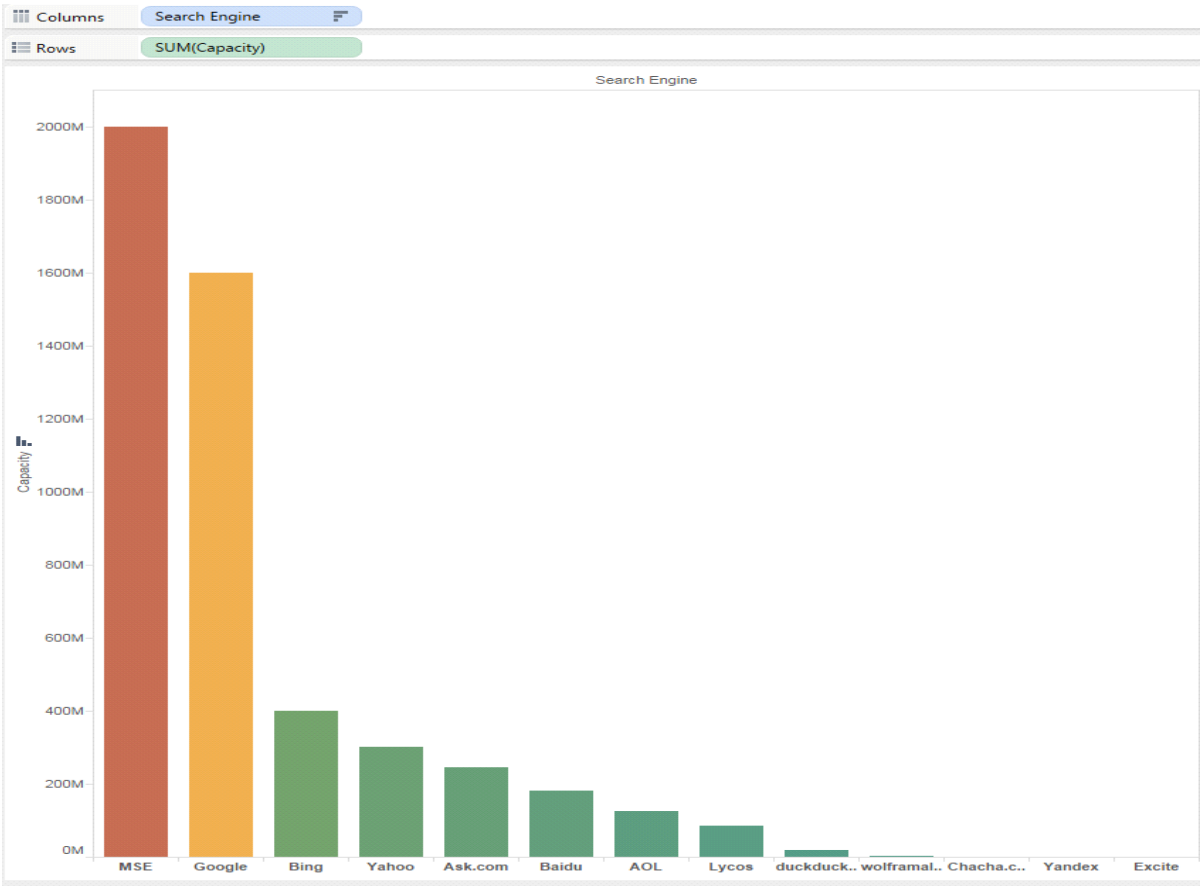
As per the number of visitors Careers page is selected as the most viewed path

DOI: 10.18535/ijecs/v6i1.40

Graph representation of Search Engine on its capacity:

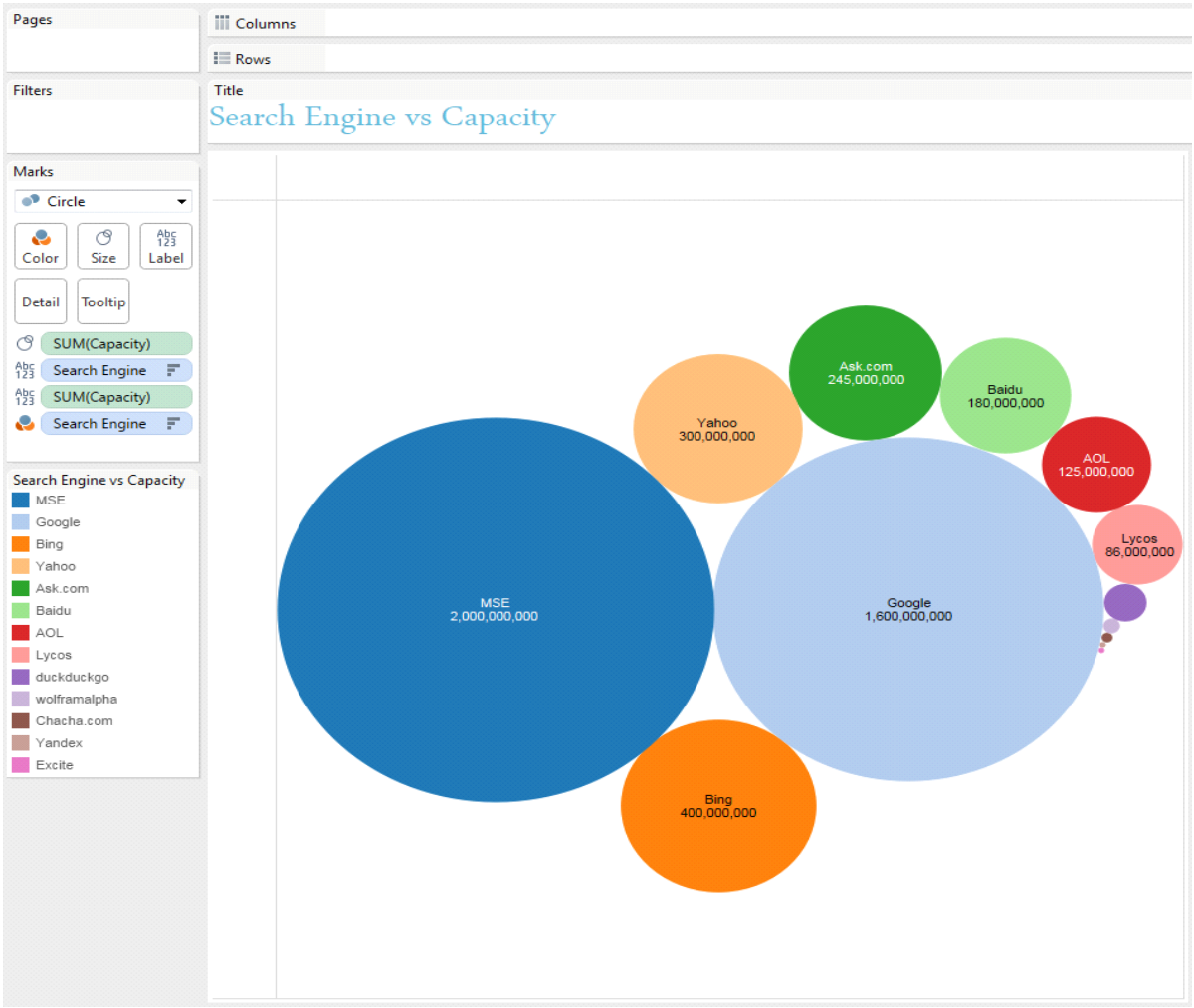
When we compare MSE with different search engines based on their individual capacities MSE will be the top search engine

Bar Graph:



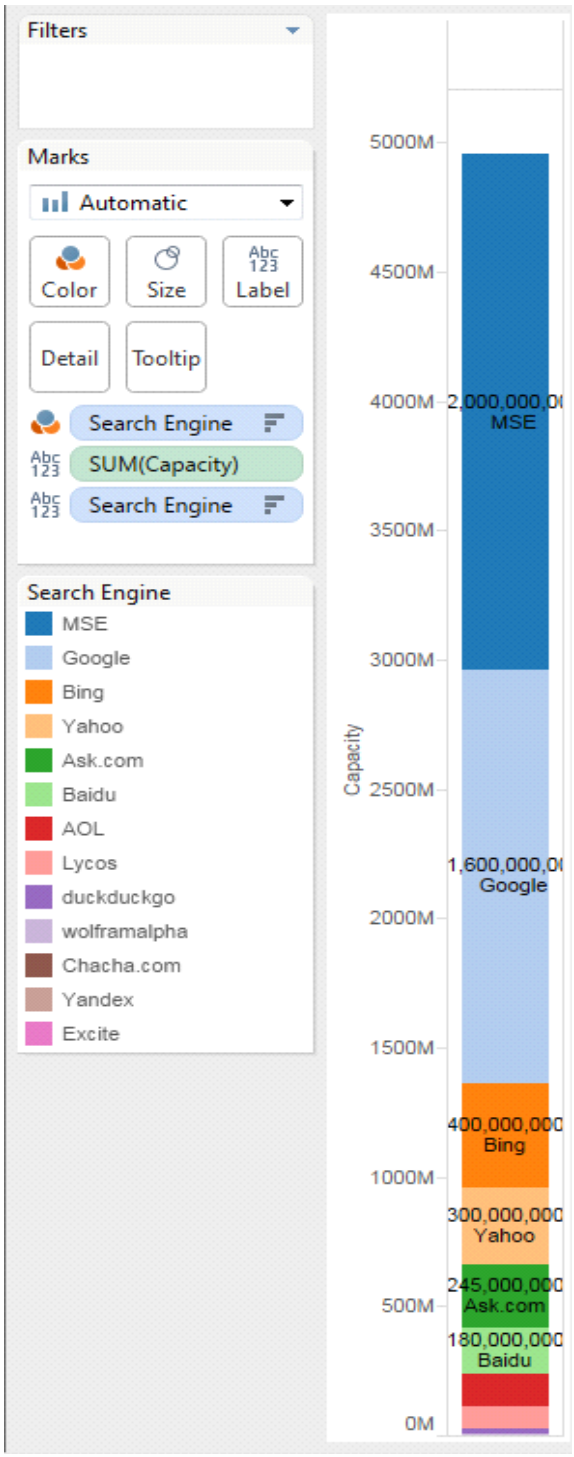
DOI: 10.18535/ijecs/v6i1.40

Packed Bubbles View:



Stacked Bars:

DOI: 10.18535/ijecs/v6i1.40



V. Conclusion

In this paper, we taken different search engines, typed the same query in all search engines and merged the results after that re-ranked the results with help of web graphs. we implemented this work and it produces efficient results. Finally This paper presents efficient search results through meta search engine by applying web graphs.

VI. References

[1]. Hao Ma, Irwin King, Senior Member, IEEE, "Mining Web Graphs for Recommendations".

[2]. Weiyi meng et a1.,"Building Efficient and Effective Meta Search Engines".

[3]. Joeran Beel et a1., "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co".

[4]. Larry Kerschberg et a1., "Intelligent Web Search via Personalizable Meta-search Agents".

[5]. Eric J. Glover et a1., "Architecture of a Meta search Engine that Supports User Information Needs".

[6]. Theodora Tsikrika et a1., "Merging Techniques for Performing Data Fusion on the Web".

[7]. Choon Hoong Ding et a1., "Guided Google: A Meta Search Engine and its Implementation using the Google Distributed Web Services"**.**

[8]. Hossein Jadidoleslamy et a1., "Search Result Merging and Ranking Strategies in Meta Search Engines: A Survey"

[9]. Felipe Bravo-Marquez et a1., "DOCODE-Lite: A Meta-Search Engine for Document Similarity Retrieval".

[10]. Yiyao Lu et a1., "Evaluation of Result Merging Strategies for Meta search Engines".

[11]. Khaled Abd-El-Fatah Mohamed et a1., "Merging Multiple Search Results Approach for Meta Search Engines".

[12]. Naresh kumar et a1., "A Meta Search Engine Approach for Organizing Web Search Results using Ranking and Clustering".

[13]. K.Srinivas et a1., "Web Service Architecture for a Meta Search Engine".

[14]. Andreas Meier et a1., "Meta-Search Engine Analysis".

[15]. King-Lup Liu et a1., "AllInOneNews: Development and Evaluation of a Large-Scale News Meta search Engine".

[16]. Zonghuan Wu et a1., "Creating Customized Meta search Engines on Demand Using SE-LEGO".

[17]. Adele E. Howe et a1., "SAVVY SEARCH A Meta search Engine That Learns Which Search Engines to Query".

[18]. John Mc Carthy et a1., "The Page Rank Citation Ranking: Bringing Order to the Web".

[19]. Steve Lawrence et a1., "Inquirus, the NECI meta search engine".

[20]. Thorsten Joachims et a1., "Optimizing Search Engines using Click through Data".

[21]. Zonghuan Wu1 et a1., "Towards a Highly-Scalable and Effective Met search Engine".