# MetaCrawler: A Literature Review

Monali R Parthe
Computer Engineering Department
Savitribai Phule Pune University
SRTTC, Kamshet, Pune, India

Sarika Choudhari
Computer Engineering Department
Savitribai Phule Pune University
SRTTC, Kamshet, Pune, India

## ABSTRACT
As the world-wide web is increasing quickly, Now a day's searching information on internet, not only the information also find the truth & related data about topic, so It is complicated to find truth details and relevancy. Unluckily, there is no assurance for the exactness of information on the web. Likewise, different websites often provide inconsistent information on a subject, such as different terms for the same product. We design a general structure for the veracity (trueness) problem, and originate an algorithm called Truth Extractor, this operates the associations among web sites and their information, i.e., a website is truthness if it runs many bits of truth material, and a bit of material is possible to be true if it is provided by many truthness web sites. In this paper we use Truth Extractor to calculate true details among variance information, and identify truthness web sites better than the popular search engines.

## Keywords
Search Engine, Data quality, Truth Extraction Algorithm, Ranking, clustering

## 1. INTRODUCTION
The Internet is the set of a lot of information which is associated from all over the world. Currently the World Wide Web (WWW) turn out to be a major data storehouse. The web is highly active means everyday new websites are being added, some sites are removed & some are modified. So it becomes difficult to find out the truth details & related information from web because of information overhead. In world many people's using web as a main basis to find the information. The search Engine is a series (Set of Instruction) that searches & find out the objects in a database that corresponds to the characters, phrases & keyword which the users specified in search [3]. A Meta Search Engine queries data & combines the result of data from other popular search engines. Meta search engine not create their own database. It search database from other popular search engine. The Meta search engines provide the main advantage that it allows to search several search engines at once. We are introducing a new Meta Search engine, which queries data from other search engines & provide the corrective & proper data. We will rank each & every web page based on its relevance & popularity for that we are uses a modified page rank algorithm to rank the web pages and before ranking we are also use a Truth Extractor algorithm for finding truth details or data of particular search query.

## 2. PROBLEM STATEMENT
Now a day's searching something on search engine are very common thing but so many search engine shows incorrect output for your searched criteria. To improve user search result and experience in web searching we are producing a search engine which named as MetaCrawler. MetaCrawler extract top website for your searches from top search engine remove duplicate links and find truth facts from search result as show trueness and properly ranked output. Properly ranked correct output will give best experience to user.

## 3. PREVIOUS FINDING
A group of work has already been done on measuring search engine retrieval effectiveness. In previous work of Meta search engine, we found that each service returns dissimilar documents for the same query. Also there is an intrinsic time interval in each service results-one service may need a catalog of a text that is merely a day or two ancient, whereas another may have an index that is a month old. Thus you never know if the references returned are clean references, or hard references that were once appropriate [5]. Provide duplicate links. Also in previous Meta search engines provides user to select search engines from given or user want to search data on selected search engines that he wants. We have also observed that search engines shows irrelevant material and advertisement that not useful for user. [2] Earlier Experiments verified that by examining pages to ensure they present and have quality content, as well by providing the user with added meaningful power, over most of the orientations returned to the user could be robotically resolute to be unrelated and accordingly removed.

## 4. METHODOLOGY USED
**User Query:** User Query/Phrase which is given by user for searching information.

**Query Processing:** In Query Processing, Firstly user put query on metacrawler website. MetaCrawler is Meta search engines which extract data from popular search engines such as Google, Yahoo, and Bing etc. It collect scanned data from other search engines, index them and send to the next phase. We apply termination condition when all the URLs crawls then stop and go to next, otherwise continue crawling.

**Eliminate Duplicate URLs:** This module check whether an extracted links in that one of the link is already in our URL list or has been recently fetched. If the URL is already present in the list, then it is not added into the list. [1]

**Truth Extractor:** Truth Extractor, utilize the relations between web sites and their information, i.e., a website is truthful if it runs many bits of accurate information, and a bits of material is possible to be true if it is provided by many reliable web sites. Our program show that Truth Extractor successfully finds true specifics among conflicting information, and identifies truthful web sites better than the popular search engines. A large amount of conflicting information around numerous objects, which is distributed by plentiful web sites (or other types of information providers), how to discover the correct detail about every object. We use the word "fact" to represent something that is claimed as a fact by certain web site, and such a fact can be each true or false. We apply condition over here, if the average of truth value of particular page is less than max or up to threshold,

then we have eliminate that URL, we have to proceed this until all URL finish. We first introduce the two most important definitions in this project, the self-assurance of facts and the honesty of web sites. [4]

Definition 1: (Confidence of facts.) The confidence of fact *f* (denoted by *s (f)*) is the chance of *f* being right, according to the finest of our knowledge.

Definition 2: (Trustworthiness of web sites.) The trustworthiness of a web site *w* (denoted by *t (w)*) is the expected confidence of the facts provided by *w*.

As in second definition, the trueness of websites is just as expected confidence of facts which it provides. For website the variable use is w, and for trueness of website *t (w)* by calculating the average of the confidence of facts provided by website w. [4]

$$t(w) = \frac{\sum_{f \epsilon F(w)} s(f)}{|F(w)|}$$

Where *F* $(w)$ is a set of facts provided by websites w.

In general, if fact f is the just fact about a thing, then its confidence *s (f)* can be computed as [4]

$$s(f) = 1 - \Pi \prod_{w \epsilon W(f)} \big(1 - t(w)\big).$$

Where *W (f)* is the set of websites provide that fact f.

**Ranking:**
Page Rank algorithm is a link analysis algorithm required for ranking web pages. This algorithm assigns a mathematical weighting to each factor of a hyperlinked set of documents. In the improved Page Rank algorithm, a web page is parted into some blocks by HTML document's structure and the maximum weight is given to connections in the block that is most significant to the given topic. The visited out links are regarded as advice to modify blocks' relevancy. The achievement of this novel algorithm helps in determining the problematic of topic drift. After applying ranking the ranked result is display to the user. Take any given element E is referred to as the *Page Rank* of E [1].

In general,

**PR (A) = (1 – d) + d (PR (T1) / C (T1) + …. + PR (Tn) / C (Tn))….. [1]**

Where,

PR (A) is the Page Rank of a page A

PR (T1) is the Page Rank of a page T1

C (T1) is the number of outgoing links from the page T1

D is the checking factor in the range of 0 < d < 1

## 5. PROPOSE SYSTEM

**MetaCrawler:**
In this paper, a Meta search engine using data from other search engines like Google, Yahoo, Bing, Ask etc. In previous Meta search engine the user has choice to select the search engines he wants to search, we are eliminating this information in this Meta search engines [2]. Also in previous work, they are not find the correctness of data, this problem is overcome by using Truth Extraction algorithm [4]. Instead of displays the separate tag's for same query like tags are image, web, and videos, in this search engine all this tag's are display on one single page. A metacrawler is similar to that of Meta search engine, which blended top web search results from popular search engines. Enables the user to retrieve information that is related to the topic. The user query is given through the chief Graphic User Interface (GUI).

The examiner query by the user is administered such a way that all the stop words are detached for improved handling of the query. The request is given concurrently in all the traditional search engines and the outcomes matching to the query are mined from the web record. The designed Meta search engine retrieves major results from different traditional search engines available [1].The retrieved results are extracted by using the crawler. Process Flow is given below:

1. Enter Query/Phrase
2. User Query on MetaCrawler websites.
3. Collect data from different search engines.
4. Eliminate the duplicates.
5. Apply Truth Extraction algorithm.
6. Properly ranked data using Ranking algorithm.
7. Combine data and shows result to the user [6].

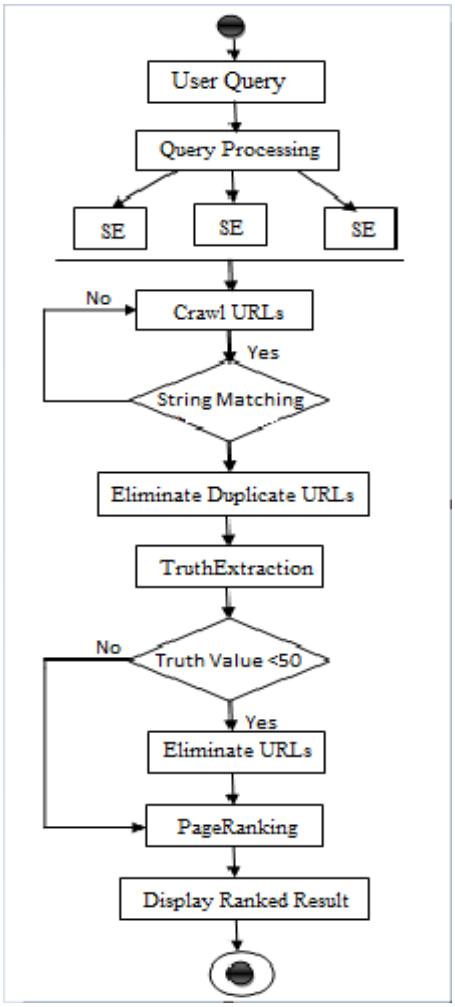


**Fig. Activity Diagram**

The system activity diagram shows the following activities taking place in the system: URL server prepares list of URL's and assigns a DOCs to each URL. If the URL is not visited then give it to the crawler. If the crawler is visiting the URL for the first time then the crawler downloads the file to see the access permission and saves this file for future reference. Now download the page and extract the information from this site. Now combine this page and store in the store sites. If the page has already been visited then download it from the store server.

# 6. PERFOMANCE EVALUATION PARAMETER

The performance of this search engine is evaluated with dissimilar no. of queries. First, the recovery effectiveness of the crawler is measured and then average precision value for dissimilar queries is calculated for all the search engines and is compared with that of the proposed crawler.

**Relevance Ratio**

The relevance ratio of a search engine is calculated using the as given below equation. [1]

$$Relevanced\ Ratio = \frac{\text{Number of Relevant URLs}}{\text{Total Number of URLs Retrieved}} * 100$$

This relevance ratio is used to estimate the performance of our search engine. It also gives improved performance than traditional search engines with improve the quality of web search results.

**Retrieval Effectiveness**

The RE value for dissimilar search engines for different queries is considered and is compared to that of the proposed crawler. [1]

$$RE = \frac{\sum_{i=1}^{N} ri}{N}$$

Where, *N* is the number of queries *r I is* the precision for query *I.* The no. of related web documents obtained by executing dissimilar queries on the traditional search engines and the proposed search engine along with average precision value and relevance ratio.

# 7. CONCLUSION

Identifying the Ranking techniques and used modified page rank on the basis of user analysis. This system will help the user to find the Truthfulness of information using Truth Extractor algorithm, about which our paper is based, calculating truth ratio giving correct information to the user about corresponding data. The result extends the conclusion of page rank and truth extraction by overcome the pervious drawback of traditional search engines.

This paper could be implemented on the web to find the truthfulness of data in various websites. It can also be used as an independent search engine. The existing semantic search engines may give the relevant result to the user query but may not be 100% accurate. Our algorithm computes trustworthiness of websites to rank the web pages. Simulation results show that our approach is efficient when compared with existing.

# 8. REFERENCES

[1] Design of A MetaCrawler For Web Document Retrieval [ISDA-978-1-4673-5119-5/12/ 2012 IEEE].

[2] An Analysis of Web Document Clustering Algorithms [ISSN-Volume 1, No.6, December 2011].

[3] Web Crawling Algorithms [International Journal of Computer Science and Artificial Intelligence Sept. 2014, Vol. 4 Issue. 3.].

[4] Truth Discovery with Multiple Conflicting Information Providers on the Web [IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 6, June 2008].

[5] The MetaCrawler Architecture For Resource Aggregation on the Web [November 8, 1996].

[6] An Intelligent Meta Search Engine for Efficient Web Document Retrieval [*(IOSR-JCE) e-ISSN: 2278-0661, Volume 17, Issue 2, Ver. V (Mar – Apr. 2015)*].

[7] Information retrieval on Internet using meta search engines: A Review [(Journal of Scientific & Industrial Research) Vol.67,October 2008,pp. 739-746].

[8] String Matching Algorithms and their Applicability in various Applications [(IJSCE) ISSN: 2231-2307, Volume-I, Issue-6, January 2012].

[9] www.metacrawler.com or www.zoo.com.