QUALITY ASSESSMENT OF SEARCH ENGINES USING METAMORPHIC TESTING

P S V Sridhar¹, P.V.N. Lokjith², B. Rajashekar³, K.V. Lavanya⁴
Department of Computer Science and Engineering
KL University, Vaddeswaram, Guntur, India
psvssridhar@kluniversity.in¹lokjithpigili@gmail.com²
booreddyrajashekar@gmail.com³ vlavanya.kolipaka@gmail.com⁴

ABSTRACT: Testing is extensively used approach for evaluating software satisfactory and supporting developers to remove faults within the software. The majority of the software program strategies count on the provision of an Oracle, a mechanism against which testers can verify the correctness of outcomes from test cases. There are numerous checking out techniques for software testing like Unit testing, Integration testing, System testing. In this paper we proposed a quantitative mechanism for testing the search engines. There are many tools for testing a software, however there may be no particular tool for assessing the search engines accurately. The internal structure of the search engine is not disclosed to the public due to the safety reasons. Primarily based on the results received from numerous queries from diverse search engines, the quality can be assessed.

Key words: Software testing, Metamorphic testing, search engines, Test Oracle.

1. Introduction

Test Oracle is a testing process for examining whether the program has passed or failed the test. Usually oracles are used for the small subset of inputs and outputs. In some scenarios, it becomes too expensive to use, which relates to oracle problem [2]. In this regard a new mechanism called Metamorphic Testing [1] is used to assess the quality of search engines [8]. However, it is not easy to assess the quality of a search engine since more volumes of data is present, and the entire structure of search engines is vast. Basically, the quality of a search engine mostly depends on the results that it gives for a user. Metamorphic testing is a user oriented approach, in this regard, the user is least bothered about the structure of a search engine, what the user needs is an accurate result for a particular search query [9].

Testing is important activity considered by any software industry as main goal involves acceptance of the software developed by the user as it’s being developed according to the users’ requirement. So, since the main part of the software is how surely it is being worked and for that testing is the main strategy needed. Since application development involves many phases and for each phase, there is a possibility of checking whether the developed part is working or not, considering the implementation phase, there will be various individual subsystems and each subsystem needs a check i.e., verification and validation [4].

2. Literature Review

The testing strategies involve giving some valid inputs and checking whether the designed software is performing required operations to get desired outputs. After the entire application is divided into subsystems, later into units, each phase needs testing strategies like unit testing and all the minor errors are to be removed and then the individual units are integrated to form a subsystem performing some operation which is needed to be tested and considered as integrated testing and then all the subsystems are needed to be combined as a whole giving us the entire system which is tested for its work.

Consider the search engines, they are designed for the purpose to serve the user required amount of accessible data accuracy. Since search engine developers use many algorithms and are not revealed due to their confidentiality, most users prioritize a particular search engine based on the accuracy of the results that it gives which is totally dependent on users’ requirement. Since the user is least bothered about how the search engine is designed and how it’s working, but what he needed is whether he is getting the accurate results or not and to test that there is no other possible way than giving a set of queries to observe the search engines’ behavior.

Construction of queries should be in such a way that it should almost satisfy users’ needs in the sense of how he searches and using Oracle is not possible since the internet is a vast data store and it’s not possible to have an exact answer to the state for any particular query and for that we can use Metamorphic testing strategy which considers the construction of testing queries which handles most of the queries that one usually searches.

The method involved in this approach is like giving a set of queries to the search engine and based on the number of results count available, we could compare the search engine’s ability to retrieve some data. Since most of the part lies in what format does the queries are constructed,
Construction of queries plays a major role in metamorphic testing. There will be two sets of queries we need to construct if we want to assess a search engine which is source query set and follow-up query set. Source query sets are mostly normal query sets which a user searches on the internet and based on the results obtained from the source query set, follow-up queries are needed to be constructed and for that metamorphic relations are used that defines how the source query is related to the construction of a follow-up query [1]. They are as follows:

2.1. Mpsite:

In this type of relationship, the follow-up query is needed to be constructed based on the initial result that the search engine delivers the follow-up query consists the source query term along with the site extension that it returns. The initial query could be given in quotes to ensure that the search engine returns the websites that got the sentence exactly which is termed as accurate.

For example, consider this, the source query is “greenhouse effect in nature” and consider this query term is given to Google and it returned results and in that obtained results, the top result is from Edu domain and so the follow-up query should be constructed as “greenhouse effect in nature” site:.Edu. For now the follow-up query is more specific to a particular domain, there is more possibility of getting the results only from that particular Edu domain and so the search engine data retrieval could be assessed.

The test engines we assessed was Google and AOL, for Google the follow-up query is constructed as .edu and for AOL it was constructed by adding the sites as ‘.com’ since its top result is from a commercial domain site (Fig.1).

2.2. MPTitle:

This is another form of metamorphic relation. In this, the follow-up queries are needed to be constructed based on the first result’s title that it returned. Consider there is a source query as “q” and for this query, consider that the search engine returned a top result with abcd as the title and then the follow-up query should be constructed as “q”abcd.

For example, consider this, the source query is given as "honest compliment" and the follow-up query is obtained by swapping the words as ‘compliment honest’. These queries are given to Google and AOL for assessment (Fig.3).

2.3. MPReverse:

In this type of relationship, the follow-up query is constructed based on reversing the given set of source queries, consider the source query is given as “A and B and C and D”, then the follow-up query should be constructed by reversing the order of terms in the query like “D and C and B and A”.

For example, consider this source query "honest and good and bad and wrong", and for that, the follow-up query is constructed as “wrong and bad and good and honest”; these are needed to be given to both search engines for assessment(Fig.2).

2.4. MPSwap:

In this type of relation, there are only two words are needed to be considered and the follow-up query is constructed by swapping those two words. Consider the source query term as w1 w2, then the follow-up query is constructed as w2 w1.

For example, consider the source query is given as ‘honest compliment’ and the follow-up query is obtained by swapping the words as ‘compliment honest’. These queries are given to Google and AOL for assessment (Fig.3).

2.5. Top1Absent:

In this type of relationship, the follow-up query is constructed based on the first website that it fetches and so that the next search will only be in that particular site only and consider the source query as “q” and for that the website that it returned is web, then the follow-up query is constructed as “q”web.

For example, consider the source query as "dragon", for this, the top most result of google is obtained from “Wiki”, so the follow-up query is constructed as “dragon” wiki.
The AOL has returned the top result as “Moviefone” and thus the query is constructed as “dragon” Moviefone (Fig.2).

3. Implementation

3.1. Search Engine Testing

There are many search engines currently available on the internet like Google, AOL, Yahoo, Ask, etc. There are many testing mechanisms to test the quality of a particular search engine by comparing it with other search engines. The quality cannot be estimated accurately as the internal structure and the working mechanism is not disclosed to the public to avoid security breaches. The test cases passed are tested through a metamorphic testing mechanism by which we compare the number of search results between Google and AOL.

A set of search queries is given to both the search engines and the number of search results obtained in each website is taken as a parameter for analysing the quality of search engines. The quality of the search engine can also be estimated by comparing the number of results obtained from each search engine and the search engine which gives maximum results can be considered as the best search engine among them.

3.2. Working Methodology

This project is implemented by using the files concept in Java. Google Chrome is opened and a query is given as input and the obtained webpage of the search results is saved as an html file in a folder. The same search query is given in AOL and the obtained html page is saved. The html file obtained from each search query for a particular search engine is taken as input to the program. The output of the program is saved as a text document.

After taking the html file as input, all the text from the web page is retrieved and the first line of the web page which includes the number of results stored in the text file which is obtained as the output of the program. Each set of results obtained from a particular search engine is stored in its respective text output file. All the output files are stored in one particular folder.

3.3. Analysing Search Results

The number of search results obtained for MPSite, MPTitle, MPreverse, MPSwap, Top1Absent [1] type of relations are evaluated. The analysis can be represented graphically by taking both search engines as histogram on the X-axis and the corresponding number of results is taken on Y-axis. This pictorial representation is used to easily estimate the quality of the search engine.

4. Algorithm

1. Start
2. Enter the search query constructed using metamorphic relations in search engines and obtain the results
3. Save the results contained webpage
4. Read the results file into the Java program.
5. Search for the string obtained between “About” and “results” and store it to some variable.
6. Display the obtained result number and store in a text file using file writer class
7. Analyse the results and conclude the best among the assessed search engines
8. Stop.

5. Text Outputs

The search results obtained for MPSite and MPTitle are graphically represented in Fig.1., the results for MPreverse and Top1Absent are represented in Fig.2., and the results obtained for MPSwap are represented in Fig.3. By comparing the three graphs, we can conclude that Google is the best search engine compared to AOL.

Results for MPTitle and MPSite
Results for MP\text{reverse} and Top1\text{Absent}.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{HistMPsiteMPtitle}
\caption{Histograms of MPsite and MPtitle}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{HistMPreverseTop1Absent}
\caption{Histograms of MP\text{reverse} and Top1\text{Absent}}
\end{figure}

6. Conclusion

In this project a number of results obtained is taken as a parameter for analysing the quality of a search engine and found that google has comparatively more number of search results than AOL and it can also be analysed based on the time factor. The quality cannot be tested precisely using the time factor, because the time taken to load a web page may vary due to the change in the speed of the internet. The top results for any query in a search engine is based on the number of visits to that particular site, but these top results may not be reliable because we cannot assure it as reliable based on number of visits to that page. This can be avoided by providing the top results based on the rating and certification of the websites.

The reliability defines the quality of the search engine. When a search query is given to the search engine it must retrieve the results from the trusted websites and show them in the top results and all the non-reliable sources can be avoided so that it may not cause any security breaches of the user.
References


