On Applying Metamorphic Testing: An Empirical Study On Academic Search Engines

Stevão Andrade, Ítalo Santos, Claudinei Brito, Misael Júnior, Simone R. S. de Souza, and Márcio E. Delamaro

Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (USP), São Carlos, Brasil
{stevao, srocio, delamaro}@icmc.usp.br
{misaeljr, italo.santos, claudineibjr}@usp.br

***Abstract*—Software testing can be a helpful practice to certify the quality of a product. However, there are programs which are hard, if not impossible, to determine the expected outputs. This problem is called the "oracle problem". Metamorphic testing (MT) is an approach that aims to alleviate this problem by setting a series of relations, called metamorphic relations (MRs). This paper applies MT through a user-oriented approach and the following MRs: MPublished, MPTitle, MPShuffleJD and Top1Absent as a strategy for evaluating, verifying, and validating four Academic Search Engines (ASEs): ACM, IEEE, ScienceDirect, and Springer. Therefore, we conducted an experimental study to analyze how MRs can contribute to verifying the correctness of the properties of ASEs. Results indicate that the ASEs have performed differently in their execution for each MR analyzed. This points out that the algorithms of the scientific search engines perform differently for the characteristics analyzed. This may not necessarily be caused by a programming fault; the design may contribute to the cause of the presented behavior. On the other hand, the approach is useful for search engine users that need to be sure that the ASEs behave as expected during a search, mainly to conduct secondary studies in which the results heavily depend on the correct behavior of the search engines used.**

***Index Terms*—software testing, oracle problem, metamorphic testing, academic search engines**

I. Introduction

Software testing is a mainstream approach to software quality assurance and verification. However, it faces two fundamental problems: the oracle problem and the reliable test set problem. In this paper, we focus on the first problem. The "oracle problem" refers to situations where it is extremely difficult, or impossible, to verify the test result of a given test case [2].

The present research extends the work from [12] where they use MT in a quantifiable approach for software quality assessment, which includes the verification and validation of software correctness. The authors applied their approach to alleviate the oracle problem and proposed the following MRs: (i) MPSite; (ii) MPTitle; (iii) MPReverseJD; (iv) SwapJD; and (v) Top1Absent. In this paper, we fully adapt the MPTitle and Top1Absent from the experiments presented in [12], and we create the MPublished based on MPSite, and put together MPReverseJD and SwapJD and transform into the MR called MPShuffleJD.

We divided the MRs into two groups. The first group aims to assess the ASE's ability to return scientific documents that contain an exact word or phrase. The MRs related in this group are MPublished and MPTitle. The second group aims to evaluate the ranking stability of the results in a query. This group comprises MPShuffleJD and Top1Absent.

Our work intends to verify whether it is possible to adapt existing MRs to apply them in similar contexts. To do so, we adapt five MRs which were initially proposed to assess search engines into four MRs designed to assess academic search engines (ASEs), namely: *ACM*[1], *IEEE*[2], *ScienceDirect*[3] and *Springer*[4].

ASE is a specialized product which only covers scientific information but is focused on a concrete user type: scientists. These users require search services, and ASE offers a range of instruments that not only allow them to locate precise and relevant information but can also evaluate the quality of the returned data. Besides, unlike other users, scientists are simultaneously consumers and creators of content, which makes them more critical to the functionalities offered by the ASEs, e.g., the way their papers are indexed. Based on this observation, we use a user-oriented testing approach proposed in [12] which defined MRs from the users´ perspective.

Therefore, the user does not need to understand the system overall to validate the ASE. Instead, they only need a testing technique that tells whether the few functions directly involved in the search can deliver what they want. When the test fails, it can either indicate a fault in the implemented software system or a deficiency in the algorithm (s) chosen by the search engine developer. For validation purposes, the user does not need to distinguish between these two cases. In this paper, we present experiments using MRs to verify if it is possible to adapt the MRs proposed in [12] for purposes of testing and quality assessment of ASEs based on user-oriented testing approach.

This work is organized as follows: Section II discusses related work. Section III describes the MRs used in our experiment. Section IV describes our experiment, including the defined research questions and the established hypothesis. In Section V, we analyze the empirical study results and describe a more in-depth discussion about our findings. Section VI shows the validity threats related to our experiment. Sec-

[1]https://dl.acm.org/
[2]https://ieeexplore.ieee.org/Xplore/home.jsp
[3]https://www.sciencedirect.com/
[4]https://link.springer.com/

tion VII summarizes this work, discusses some implications and future work opportunities.

II. Related Work

In previous works, researchers have studied the different contexts in which MT can be applied as a promising strategy to alleviate the oracle problem. Detecting real flaws in search engines [12], image processing applications [5] and finding faults in compilers [10] are some of the contexts in which MT has been successfully applied. However, there is an insufficient number of approaches that aim to deal with the evaluation of large and complex software, e.g., ASEs.

Zhou et al. [12] proposes an approach to alleviate the oracle problem for testing and quality assessment of (web) search engines. Thus, the authors developed a set of MRs that were identified and proposed from the users´ perspective. The results demonstrate that the approach can alleviate the oracle problem and mitigate non-specification challenges to verify, validate, and evaluate complex software systems.

The results presented in [12] demonstrated that different (web) search engines may show a variety of results concerning functional characteristics. They contribute directly to the user's perception and satisfaction. Our work is based on the premise that ASEs should respond precisely to the user's expectations. Then, different from traditional search engines, ASEs are used by researchers and the divergences or inconsistencies in the results may represent a threat to validity in research. Thus, we try to identify how each ASE responds when analyzed under the perspective of the MRs proposed by us in this paper.

III. Description of Metamorphic Relations

A set of MRs was used to evaluate if it is possible to adapt them to the ASE context for testing and quality assessment purposes. As mentioned early, we divided the MRs into two groups: (i) first group: MRs that evaluate the ability of the ASEs to return specific scientific documents (MPublished, MPTitle); and (ii) second group: the MRs that assess the ranking stability of the results (MPShuffleJD, Top1Absent). In this Section, we present a brief description of these MRs.

A. MPublished

MPublished aims to evaluate the retrieval functionality of scientific documents in the ASE. Hence, it focuses on the reliability of the ASE in retrieving scientific documents that contain exactly one word or phrase. Therefore, this MR makes part of the first group of MRs that evaluate the ability of the ASEs to return the scientific documents.

In MPublished, the *source query A* is constructed from keywords and then executed. The *follow-up query B* is built from the extraction of a parameter called "published" referring to the publication source (conference or journal) of the first item in the list of results of *source query A*. It is supposed that the first item in the list of results of the *source query A* is contained in the list of results of the *follow-up query B*. According to the template suggested in [8], we can describe this MR as follows:

- ***if*** the search terms of the *follow-up query B* are defined based on the parameters of an extracted published value that is contained in the results of the first paper returned in *source query A*;
- ***then*** the results of the *follow-up query B* must contain the first result extracted from *source query A*.

The primary goal of this MR is to evaluate the behavior of the search engine by specializing an original query. Table I presents an example of MPublished execution, the publication venue of the first result of the *source query* is extracted, and it is expected that by including this information in the *follow up query*, within the original keywords, the search engine returns at least the paper from which the information was extracted in the *source query*. If the ASE is not able to return this paper, this may indicate a malfunction or failure in the algorithm that is responsible for the query specialization mechanism.

B. MPTitle

This MR is interested in the ability of the ASE to understand and abstract scientific works through the title of the study results. Hence, it is part of the first group that evaluates the ability of the ASEs to return the scientific documents.

MPTitle is done by constructing a *source query A* for search, and then a *follow-up query B* is created which will contain the previously executed *source query A* and the title of the first paper returned in the *source query A*.

Thus, it is expected that at least one result is found when executing the *source query A*. Otherwise, it is observed that the MR was violated and there is a fault in the ASE. We can describe this MR as follows:

- ***if*** the search terms of the *follow-up query B* are defined based on the parameters of an article that is contained in the results list of the *source query A*;
- ***then*** at least one result of the *follow-up query B* should be equal to the first result extracted from *source query A*.

Table I shows an example of MPTitle execution. We can see that by constructing a follow up query based on *source query*, adding the name of the first paper returned, the specified paper will not be returned by the query. It may suggest a possible flaw in the search engine since the *follow up query* is presented as a specialization of the *source query* and should return at least the specified paper to the body of the search.

C. MPShuffleJD

MPShuffleJD states that a stable ASE should return similar results for *source query A* and *follow-up query B*, doing this by checking the stability ranking of an ASE. Moreover, the two sets of results must have a great convergence. It refers to this functioning as a type of quality attribute that is related to the stability of the ASE. Therefore, this MR is part of the second group that evaluates the ranking stability of the results.

The *Jaccard coefficient* [3] is used to verify the similarity between the results of both queries. We can describe this MR as follows:

- ***if*** the search terms of a *source query* are defined in an order;

TABLE I
METAMORPHIC RELATIONS EXAMPLES.

Metamorphic Relation	Source query	1st result	Follow-up query	Result
MPublished	(("Document Title":""Malware" OR "Monetization" OR "Online advertising" OR "WEB""))	IEEE Recommended Practice for the Internet - Web Site Engineering, Web Site Management and Web Site Life Cycle	(("Document Title":""Malware" OR "Monetization" OR "Online advertising" OR "WEB"")) AND ("Document Title": "IEEE Recommended Practice for the Internet - Web Site Engineering, Web Site Management and Web Site Life Cycle")	True
MPTitle	(("Document Title":Algorithm" OR "Linear programming" OR "Pulse-width modulation" OR Simulation))	Random swithing frequency pulse width modulation: Analysis, design and simulation	(("Document Title":Algorithm" OR "Linear programming" OR "Pulse-width modulation" OR Simulation)) AND ("Document Title": "Random swithing frequency pulse width modulation...)	False
MPShuffleJD	(("Document Title":""Business logic" OR "Coherence" OR "Domain theory" OR "Domain-specific language""))	Not Applied	(("Document Title":""Coherence " OR "Domain theory" OR "Domain-specific language" OR "Business logic""))	Jaccard Similarity Coefficient - 0.64
Top1Absent	(("Document Title":""Algorithm" OR "Approximation algorithm" OR "Biological Processes" OR "Cellular organizational structure""))	Detecting Phenotype-Specific Interactions between Biological Processes from Microarray Data and Annotations	(("Document Title":""Algorithm" OR "Approximation algorithm" OR "Biological Processes" OR "Cellular organizational structure"")) AND ("Document Title": "Detecting Phenotype-Specific Interactions between Biological Processes from Microarray Data and Annotations")	True

*Example extracted from the *IEEE* academic search engine

– ***then*** the result of a *follow-up query* with the same search terms in a shuffled order should be similar to the results of the *source query*.

Since there are no recommendations that state how the keywords should be organized in an ASE query, it is expected that the order of the terms should not interfere in the results. Therefore, the purpose of this MR is to evaluate how the ASE ranking mechanism is covered by shuffling the terms used in the query, verifying the degree of similarity between the *source* and the *follow up query*. We can see in Table I that different from the other MRs, this one, in particular, does not need the first result to create the *follow up query*.

D. Top1Absent

Top1Absent focuses on the quality ranking of the first result returned by the search. Therefore, it is part of the group of MRs that evaluate the ranking stability of the results. According to Signorini and Imielinski [9], this result can be considered the most important of all search results and more than 65% of search clicks are made in the first result.

The *source query A* is constructed using a list of terms, and then, the *follow-up query B* is built based on the *source query A*, using the name of the first paper returned in the *source query A*. We can observe in Table I that the *follow-up query B* of MPTitle and Top1Absent are the same. However, while MPTitle looks at a list of papers returned, Top1Absent focuses only on checking the first result returned in both queries. We can describe this MR as follows:

– ***if*** the search terms of the *follow-up query B* are the same as a *source query A* and are defined based on the title of the first paper returned on *source query A*;
– ***then*** the first result of the *follow-up query B* should be equal to the first results of the *source query A*.

Like MPShuffleJD, Top1Absent tries to measure the behavior of ASEs concerning their ability to rank results. Despite having a similar construction to MPTitle, its objective is to evaluate if the ASE is not only capable of returning the paper specialized in the *follow up query*, but also assessing if this paper is the first result in the list of returned papers.

IV. EXPERIMENT DESIGN AND ARCHITECTURE

A. Experiment Design

This experiment aims to analyze how the MRs presented in Section III can contribute to verifying the correctness of the properties of ASEs. An execution environment was defined to allow the application of MT and the selected MRs were evaluated. Specifically, we intend to investigate the following research questions:

- $\mathbf{RQ_1}$: Is it possible to point out possible faults in ASEs using MRs?
- $\mathbf{RQ_2}$: Is it possible to detect possible anomalies in ASEs algorithms using MRs?

$\mathbf{RQ_1}$ addresses the ability of the ASEs to return scientific documents containing exactly one particular word or phrase. Therefore, MPublished and MPTitle performance indicate the ability to use these MRs in revealing faults in ASEs. To evaluate the performance of MPublished and MPTitle, the metric ROCOF (rate of occurrence of failures) that is the probability that a failure (not necessarily the first) occurs in a given interval of executions.

$\mathbf{RQ_2}$ aims to verify the ranking stability of the results. The performance of MPShuffleJD and Top1Absent indicates the ability of these MRs to ensure that the ASEs have stability and similarity in their functioning. To evaluate the MPShuffleJD, we used the *Jaccard similarity coefficient* [3] as a metric. Finally, to assess the performance of Top1Absent, we used the metric ROCOA (rate of occurrence of anomalies) which is the ratio of the total number of anomalies and the duration of the observation.

It is important to highlight that all metrics described are applied from a sample of 33 observations. Each observation corresponds to the average of the results of 30 executions for each ASE.

1) Hypothesis formulation: To answer the research questions, RQ_1 and RQ_2 were converted into four hypotheses so that the statistical tests could be carried out. Therefore, we propose two hypotheses related to RQ_1 and RQ_2.

Each hypothesis is related to the execution of a specific MR and measures the effectiveness concerning detecting possible flaws on the evaluated ASEs. Table II presents the null (H_0) and alternative (H_1) hypothesis for each research question.

TABLE II
NULL AND ALTERNATIVE HYPOTHESIS.

	RQ_1
Mpublished	$H_0 : \mu_{ieee} = \mu_{acm} = \mu_{scid} = \mu_{springer}$ $H_1 : \mu_{ieee} \neq \mu_{acm} \neq \mu_{scid} \neq \mu_{springer}$
MPTitle	$H_0 : \mu_{ieee} = \mu_{acm} = \mu_{scid} = \mu_{springer}$ $H_1 : \mu_{ieee} \neq \mu_{acm} \neq \mu_{scid} \neq \mu_{springer}$
	RQ_2
MPShuffleJD	$H_0 : \mu_{ieee} = \mu_{acm} = \mu_{scid} = \mu_{springer}$ $H_1 : \mu_{ieee} \neq \mu_{acm} \neq \mu_{scid} \neq \mu_{springer}$
Top1Absent	$H_0 : \mu_{ieee} = \mu_{acm} = \mu_{scid} = \mu_{springer}$ $H_1 : \mu_{ieee} \neq \mu_{acm} \neq \mu_{scid} \neq \mu_{springer}$

For example, the first hypothesis test for $\mathbf{RQ_1}$, checks whether (i) **Null hypothesis $\mathbf{H_0}$:** There is no difference in MPublished performance in detecting flaws in the ASEs or (ii) **Alternative hypothesis $\mathbf{H_1}$:** There is at least an ASE in which MPublished has performed differently in identifying failures. The same procedure is adopted for the other hypothesis tests which evaluate the effectiveness of the other MRs.

To verify each hypothesis, the Kruskal-Wallis hypothesis test [6] was performed. It is a non-parametric statistical test alternative to ANOVA [3] in the case of one factor with more than two treatments. The Kruskal-Wallis was conducted regarding the dependent variables described in Table III, aiming to show if there was a statistically significant difference between the investigated scenarios.

B. Experiment Architecture

In this subsection, we will describe the experiment steps. Figure 1 presents the primary structure of the experiment. First, we use the data set provided in [1], which makes a collection of the publications of scientific papers of the computation area available in DBLP[5]. It offers open bibliographic information on major computer science journals and proceedings. We randomly extract some keywords from the *Computer Science data set* to generate the search strings that will perform the search in the ASEs used in our experiment: ACM, IEEE, ScienceDirect, and Springer. The results of each execution are collected, and the process continues until it reaches the desired number of performances.

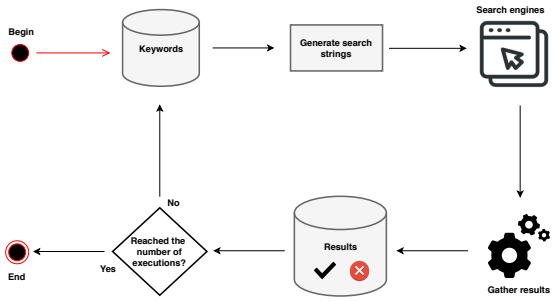

Fig. 1. Experiment Architecture

V. EXPERIMENT RESULTS

This experiment analyzes how each MR behaves when applied in the context of ASEs. The results are discussed in two ways for each MR investigated: (I) descriptive statistics and (II) hypothesis testing. To answer the research questions previously described, we carried out hypothesis tests on a consolidated value of the data obtained from all the observations made in the experiment, i.e. each observation of the consolidated data is formed from the mean of 30 observations for each ASE. For each MR, 33 observations were collected, which corresponds to a total of 990 executions for each MR in each ASE. Table IV presents the Standard Deviation, Average and Median for each MR. Each column represents the results for each ASE. All the data used for plotting the graphs, as well as for the statistical tests, are available in the experiment repository [6].

For all the experiments, two independent variables were controlled. The first is the search string that will perform the search in the ASEs and the second is the ASE itself. The dependent variables collected by the treatment are the metrics ROCOF, ROCOA and *Jaccard similarity coefficient*. ROCOF and ROCOA metrics are relevant for operating systems where the system has to process a large number of similar requests that are relatively frequent. Thus, such metrics reflect the rate of occurrence of failure or anomalies in the system [7]. A ROCOF or ROCOA of 0.06 means two failures or anomalies are likely in each 33 observations. In other words, a ROCOF or ROCOA near 1 indicates the high rate of failures or anomalies found in the system. On the other hand, a ROCOF or ROCOA near 0 means the low rate of failures or anomalies found in the system. Finally, the *Jaccard similarity coefficient* measures the similarity between finite sample sets [3]. Table III shows the independent and dependent variables for each MR.

TABLE III
INDEPENDENT AND DEPENDENT VARIABLES.

Metamorphic Relation	Independent variables	Dependent variables
MPublished	(1) *source query A* and *follow-up query B*; (2) ASE: ACM, IEEE v1, IEEE v2, Springer and ScienceDirect	ROCOF
MPTitle	(1) *source query A* and *follow-up query B*; (2) ASE: ACM, IEEE v1, IEEE v2, Springer and ScienceDirect	ROCOF
MPShuffleJD	(1) *source query A* and *follow-up query B*; (2) ASE: ACM, IEEE v1, IEEE v2, Springer and ScienceDirect	*Jaccard similarity coefficient*
Top1Absent	(1) *source query A* and *follow-up query B*; (2) ASE: ACM, IEEE v1, IEEE v2, Springer and ScienceDirect	ROCOA

The experiment was executed twice for *IEEE* ASE, because the *IEEE*'s search engine was updated on 31 July[7], and we carried out the first round of experiments in June. Then, we decided to maintain the 2 sets of results in order to compare the performance in this ASE before and after the upgrade. Therefore, *IEEE v1* matches the first execution (before the upgrade) and *IEEE v2* the second one.

A. Experiments for Research Question 1

The first research question aims to investigate whether the proposed MRs are capable of identifying possible failures or unexpected behaviors in the ASEs evaluated. To do so, we

[5]https://dblp.uni-trier.de/

[6]https://github.com/mt-ase/mt_project

[7]https://github.com/mt-ase/mt_project/blob/master/other/ieee_upgrade.png

TABLE IV
DATA SAMPLE USED IN HYPOTHESIS TESTING.

Data	Academic Search Engine				
	ACM	IEEE v1	IEEE v2	S.Direct	Springer
MPublished					
Std Dev	0.0822	0.0681	0.0920	0.0145	0.0434
Average	0.2970	0.1707	0.3980	0.0081	0.1051
Median	0.3000	0.1667	0.4000	0.0000	0.1000
MPTitle					
Std Dev	0.0205	0.0621	0.0434	0.0434	0.0360
Average	0.0081	0.1273	0.0485	0.0485	0.0222
Median	0.0000	0.1333	0.0333	0.0333	0.0000
MPShuffleJD					
Std Dev	0.0675	0.0867	0.1034	0.0804	0.0298
Average	0.2975	0.4190	0.3636	0.3189	0.9699
Median	0.3033	0.4253	0.3613	0.3333	0.9800
Top1Absent					
Std Dev	0.0081	0.0705	0.0737	0.0477	0.0328
Average	0.0020	0.1556	0.8242	0.0626	0.0232
Median	0.0000	0.1667	0.8333	0.0667	0.0000

evaluated MPublished and MPTitle concerning their capability to detect unexpected failures/behaviors. Therefore, we not only identified if MRs are useful for the purpose that they are proposed for, but we also assessed which ASE behaves more suitably for end users.

This sub-section provides descriptive statistics and hypothesis test for RQ_1. Figure 2 shows the *box plot* of the distribution of the MPublished concerning an interval of 30 executions of ROCOF scores. The *box plot* is good for visualizing the dispersion and skewness of samples. The *box plot* is constructed by indicating different percentiles graphically [11].

*1) **MPublished**:* MPublished focuses on the reliability of the ASE in retrieving scientific documents that contain exactly one word or phrase. Based on this, lower ROCOF values indicate the higher reliability of the ASE. We can see, in Figure 2 that *ACM*, *IEEE v1*, *IEEE v2*, and *Springer* present higher ROCOF values. These values show that these three ASEs have more chance to produce failures. The most reliable service was *ScienceDirect* search, whereas the least reliable service was *IEEE v2* search that reached a ROCOF as high as around 0.4 in the worst case.

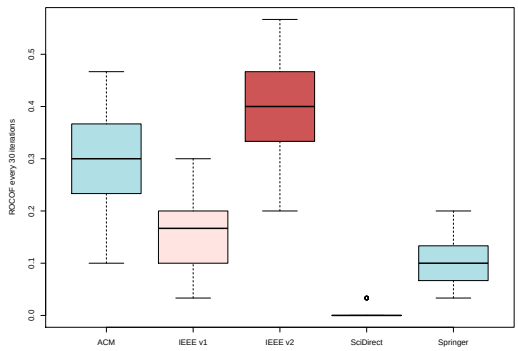

Fig. 2. Distributions of MPublished results. ROCOF: Rate of occurrence of failure

The Kruskal-Wallis test leads to the following result: *(p-value = 2,87E-29)*. The **null hypothesis $\mathbf{H}_0$** for MPublished states that "There is no difference in MPublished performance in detecting flaws in the ASEs". However, as the *p-value* is less than the margin error (0.05%), which corresponds to the complement of the significance level established at 95%, statistically the null hypothesis H_0 can be rejected. Then, according to our **alternative hypothesis $\mathbf{H}_1$**, we have evidence that MPublished performed differently in detecting failures in an ASE.

TABLE V
SIGNIFICANCE OF *p-value* IN SHAPIRO-WILK TEST.

Data	Academic Search Engine				
	ACM	IEEE v1	IEEE v2	S.Direct	Springer
MPublished					
Shapiro	0.9654	0.9568	0.9647	0.5335	0.8350
p-value	0.3641	0.2100	0.3480	0.0000	0.0002
MPTitle					
Shapiro	0.4379	0.9080	0.8738	0.8738	0.6487
p-value	0.0000	0.0087	0.0012	0.0012	0.0000
MPShuffleJD					
Shapiro	0.9818	0.9234	0.9448	0.9691	0.8712
p-value	0.8401	0.0227	0.0937	0.4569	0.0010
Top1Absent					
Shapiro	0.1677	0.9651	0.9509	0.9167	0.7188
p-value	0.0000	0.3571	0.1414	0.0148	0.0000

Looking at the results we can see that MPublished has performed differently in detecting failures in an ASE. Related to this MR, the most reliable ASE was *ScienceDirect* search because it presents a lower ROCOF value that indicates higher reliability, whereas the least reliable was the *IEEE v2* search that shows a higher ROCOF value among the ASEs used. It can be observed that after the upgrade, *IEEE* presents a poor performance related to this MR. MPublished is interested in the reliability of the ASE in retrieving scientific documents that contain exactly one word or phrase. According to [4], reliability refers to the "degree to which a system, product or component performs specified functions under specified conditions for a specified period". Based on this, we could say that *ScienceDirect* presents reliability in their executions, then from the users' perspective, this ASE meets their specific needs when used to conduct systematic literature reviews.

*2) **MPTitle**:* MPTitle aims to understand the ability of the ASEs to understand and abstract scientific works through the title of the study results. Figure 3 shows the *box plot* of the distribution of the MPTitle which also measures ROCOF scores. A visual analysis shows that none of the ASEs were perfect as their ROCOF scores were all above 0. They all produced failures in a given interval of execution. To this MR, the most reliable service was *ACM*, which was one of the ASE that showed a poor performance in MPublished. In this case, the least reliable service was *IEEE v1* that reached a ROCOF as high as 0.20 in the worst scenario.

The hypothesis testing process applied to MPTitle is similar to the one described previously. We applied the Shapiro-Wilk test to verify the distribution of our data. The results in Table V suggest that the data collected do not follow a normal distribution ($p\text{-value} < 0.05$) and the application of the non-parametric tests for statistical evaluation is recommended. Using these results, we applied the Kruskal-Wallis hypothesis

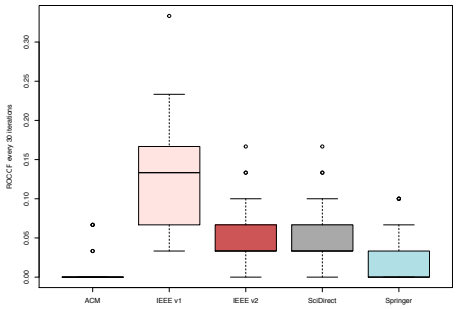

Fig. 3. Distributions of MPTitle results. ROCOF: Rate of occurrence of failure

test.

The Kruskal-Wallis test reached the result of: *(p-value = 2.33E-16)*. It is less than the margin error (0.05%), which matches the complement of the significance level established at 95%. Statistically, it is possible to reject the **null hypothesis** $\mathbf{H}_0$. Then, according to the **alternative hypothesis** $\mathbf{H}_1$, we have evidence that MPTitle performed differently in detecting failures in an ASE.

In MPTitle, we are also interested in understanding whether it is possible to detect possible flaws or bad behaviors in the ASEs. To achieve this goal, the ROCOF metric was also used to evaluate the error rate in each ASE. A high ROCOF score means a high rate of failures. Based on the hypothesis test applied, we were able to reject the null hypothesis. Thus, the results of MPTitle performed differently in detecting failures in each ASE. In the experiments, all the ASEs produced failures in a given interval of execution. In this MR, the most reliable service was *ACM* and the least reliable was *IEEE v1*. It can be observed that with the upgrade, the *IEEE v2* performed better equaling the performance of *ScienceDirect*. MPTitle seeks to understand the ability of the ASE to understand and abstract scientific articles through the title of the returned studies. This MR is focused on checking the reliability of the ASE. From the users' perspective, *ACM* performs better and meets their needs when used to find an article.

After analyzing the results presented in the descriptive analyses and hypothesis tests for MPublished and MPTitle, we can finally answer the (RQ_1) that guides the conduction of this experiment. The RQ_1 the following: *"Is it possible to point out possible faults in ASEs using MRs?"*. Based on the results and statistical tests presented, it is possible to attest with a degree of confidence that MRs were able to demonstrate possible unexpected behaviors in ASEs, and were alert to potential problems in ASEs ranking algorithms. Using randomly chosen keywords that always return valid queries (return at least one valid result) further increases the degree of reliability of the data presented to answer this research question.

B. Experiments for Research Question 2

The second research question (RQ_2) is not intended to identify flaws, but to expose possible inadequate behaviors of ASEs, regarding how they present the results of the queries, therefore the main focus of this research question is to investigate possible anomalies. To do so, we measure the ranking stability of ASE using MPShuffleJD and Top1Absent. Investigating this research question allows us to understand how each of the ASEs behaves and which one has the best results for each of the analyzed MRs.

*1) **MPShuffleJD**:* MPShuffleJD is based on the justification that a good ASE should return similar results for similar queries. Users prefer stable search results, and low stability can result in poor user experience through perception by users of search results. In Figure 4, we have the box plot showing the overall patterns of response for each group considering *Jaccard similarity coefficient*.

All the ASEs presents the *Jaccard similarity coefficient* average below 1, that is, the stability quality is not perfect for any of the ASEs. Figure 4 shows that *Springer* was superior to the other three ASEs in ranking stability and similarity. Scientists are the main users of these ASEs, and they should pay attention to the word order when searching with *ACM*, *IEEE*, and *ScienceDirect* because these ASEs are much more sensitive to this than Springer. The users of these three ASEs may also consider changing the word order to query again when they are not satisfied with the initial search result.

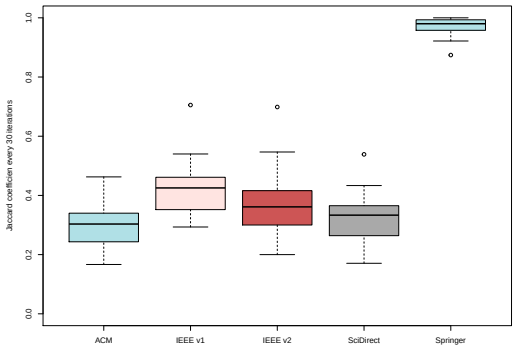

Fig. 4. Distributions of MPShuffleJD results. JC: *Jaccard similarity coefficient*.

To verify the reproducibility of the presented results, and arrive at a valid conclusion to RQ_2. Table V shows the results of the Shapiro-Wilk test applied to our collected data in order to verify the normality of the data distribution. It can be observed that some data collected do not follow a normal distribution ($p\text{-value} < 0.05$) and the application of non-parametric tests for statistical evaluation is recommended. Based on these results, we decided to apply the Kruskal-Wallis hypothesis test [6]. The Kruskal-Wallis test leads to the following result: *(p-value = 1.55E-20)*. The *p-value* is less than the margin error (0.05%), which corresponds to the complement of the significance level established at 95%, and statistically we can reject the null hypothesis H_0. Discarding the null hypothesis, we have evidence that MPShuffleJD

proves that the ASEs show a difference in stability or similarity in its operation.

The results showed that Springer was superior to all other three ASEs in ranking stability and similarity. The others, namely ACM, IEEE and ScienceDirect, present a lower result to the *Jaccard similarity coefficient* that is measured in this MR. A large *Jaccard coefficient* presented in Springer indicates greater similarity, and therefore better stability. Whereas users prefer stable search results, low stability showed in ACM, IEEE and ScienceDirect can result in poor user experience through perception by users of search results. This MR is also related to user error protection, where user error protection refers to the "degree to which a system protects users against making errors" [4]. It occurs because, in many situations, the users are unaware of what word order would be the best for them to type their query. Users should pay more attention to the word order when searching in ACM, IEEE, and ScienceDirect because these ASEs are more sensitive than Springer. Users may also consider changing the word order to query again when they are not satisfied with the initial search result.

2) ***Top1Absent****:* This MR also investigates aspects related to ASE rankings. The main aim of analyzing the ranking capacity of the ASEs is not to find a possible failure in the search algorithms, but to explore their behavior from the users´ point of view. On the other hand, MPShuffleJD is concerned that users are not aware of the fact that the order of the keywords may influence the query result. Top1Absent explores ASEs taking into account the premise that most users focus their expectations on the first result returned by the ASE. Therefore, when query results do not deliver what the user expected, this may indicate frustration on the part of the user, who may consider redesigning the query terms, that might have been initially correct or search for other ASEs.

From the data shown in Figure 5, it can be observed that at a given moment all the ASEs, excluding *ACM*, presented an anomaly rate higher than 0, which indicates that for each of the ASEs the MR was violated at least in one execution. Therefore, for this MR, the ASE that presented the best results was the *ACM*, while *IEEE v1*, *ScienceDirect*, and *Springer* presented low and similar rates, however, showing some irregularities at some point. Analyzing the results presented by *IEEE v1* and *IEEE v2*, whereas for the first version of the ASE algorithm presented a low anomaly rate, for *IEEE v2* that number grew considerably. It may indicate that the changes made to the ranking algorithm directly influenced how the results are presented to the user and this should be investigated.

Related to the hypothesis test that will be used to help respond to RQ_2, we conducted the Kruskal-Wallis hypothesis test [6], and the results *(p-value = 3.27E-34)* indicate that it is statistically possible to reject the null hypothesis H_0. Thus, the alternative hypothesis H_1 shows evidence that Top1Absent presents differences of quality ranking in their operations in different ASEs.

Therefore, gathering the results presented by MPShuffleJD and Top1Absent, it is possible to investigate RQ_2 in detail.

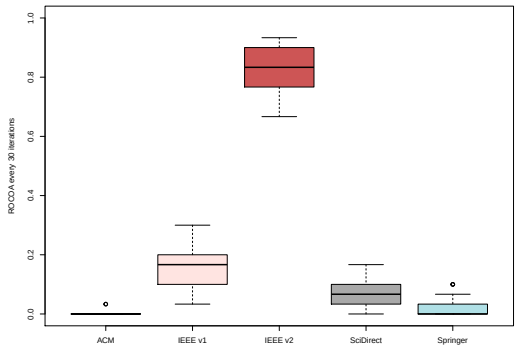

Fig. 5. Distributions of Top1Absent results. ROCOA: Rate of occurrence of anomaly

As discussed earlier, RQ_2 investigated the following *"Is it possible to detect possible anomalies in ASE algorithms using MRs?"*. The results presented by MPShuffleJD and Top1Absent in dicated differences between their ability to reveal possible anomalies in ASEs. Moreover, through the descriptive analysis, it could be noted that in all the examined instances there was the presence of some anomaly. The results led us to the conclusion that it is possible to detect possible irregularities in the ASEs through the MRs evaluated.

This answers the gaps investigated in our study, which sought to evaluate the possibility of identifying failures (RQ_1) and anomalies (RQ_2) in ASEs, using MRs, from the point of view of an end-user who does not have access to any source code or documentation information of the systems. Our research has primarily shown that ASEs are different from each other and perform differently for each of the characteristics investigated. Therefore, it is not possible to choose one ASE over another. Our study indicates that all ASEs have points that should be investigated in more depth so that they can provide a better end-user experience since the target audience of the ASEs relies/depends on the results presented, and consequently failures in these mechanisms could impact the outcome of scientific research.

VI. Threats to Validity

The validity of the results achieved in experiments depends on factors in the experiment settings. Different types of validity can be prioritized depending on the goal of the experiment [11]. There are four types of validity: (i) conclusion; (ii) internal; (iii) construct; and (iv) external.

Conclusion validity: in order to guarantee the reliability of the experiment, we carefully followed the experiment guidelines proposed in [11]. All the results were interpreted by the authors to avoid misinterpretation.

Internal validity: to avoid this threat, pilot tests were conducted before carrying out the experiment to ensure that the results obtained for implementing the experiment were correct. All the queries performed in each ASE were built based on the official documentation, available on the website of each ASE.

Similarly, the ASE was tested using samples from a data set that comprises of keywords from scientific articles that were published in the computer science field and the queries from keywords that did not return articles were excluded from the results to avoid misinterpretation.

Construct validity: all the variables were controlled to minimize the threats. The data used correspond to the average of approximately 1000 executions, for each of the hypotheses investigated. To increase the confidence, the data was analyzed not only in tables and graphs but also by statistical tests to ensure that the results were correctly interpreted and to ensure that the relations presented in the tables and charts were actually significant and not generated merely by chance.

External validity: to avoid this threat to validity, the four MRs were observed in the four different ASEs and the results obtained for each of these bases were observed and discussed individually.

VII. Conclusion and Future Work

In our experiment, we verified that when we analyze the ASE concerning its capacity in the recovery of scientific documents containing exactly the terms specified in the search, *ScienceDirect* was the ASE that presented the best results. When the focus of the search was to observe the ASEs that best return the results, based on the keywords and paper title desired, the ASE that presented the best performance was *ACM*. Regarding the stability of the ASEs concerning the ability to return similar results for similar queries, *Springer* was the one that presented better results compared to the others. It is a characteristic that we think is crucial because the user does not expect that the order of the keywords submitted will influence the results of their searches.

Moreover, we also investigated the behavior of ASEs concerning their consistency of result ranking. *ACM* was the ASE that presented better results. It is also important to note that the ASE with less consistent results was the *IEEE v2*, which shows that the upgrade performed in its algorithm may have changed in the ranking performance compared to *IEEE v1*, and it needs to be verified.

Regarding future works, we plan to increase the number of queries made in the experiment in each of the ASEs. We also plan to add new MRs that analyze other features of ASEs, such as checking their behavior by applying search filters to refine search results. We also plan to evaluate other ASEs, such as *Scopus*, and compare them with the results obtained in this first analysis to understand the operation of these search engines better and in the future to create a guide that can support users to achieve better results when using such tools.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Stevão Andrade and Ítalo Santos´ research was, respectively, funded by FAPESP (São Paulo Research Foundation), process numbers 2017/19492-1 and 2018/10183-9.

References

[1] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, and Doug Downey. Construction of the literature graph in semantic scholar. In *NAACL*, 2018. URL https://www.semanticscholar.org/paper/09e3cf5704bcb16e6657f6ceed70e93373a54618.

[2] Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, TH Tse, and Zhi Quan Zhou. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys (CSUR)*, 51(1):4, 2018.

[3] Y. Dodge. *The Concise Encyclopedia of Statistics*. Springer reference. Springer, 2008. ISBN 9780387317427.

[4] ISO ISO. IEC25010: 2011 systems and software engineering–systems and software quality requirements and evaluation (square)–system and software quality models. *International Organization for Standardization*, 34:2910, 2011.

[5] Tahir Jameel, Mengxiang Lin, and Liu Chao. Test oracles based on metamorphic relations for image processing applications. In *Proceedings of the 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 1–6. IEEE, 2015.

[6] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[7] Michael R. Lyu. *Handbook of software reliability engineering*. IEEE Computer Society Press, 1st edition, 1996.

[8] Sergio Segura, Amador Durán, Javier Troya, and Antonio Ruiz Cortés. A template-based approach to describing metamorphic relations. In *Proceedings of the 2nd International Workshop on Metamorphic Testing (MET)*, pages 3–9. IEEE, 2017.

[9] A. Signorini and T. Imielinski. If you ask nicely, i will answer: Semantic search and today's search engines. In *Proceedings of the International Conference on Semantic Computing*, pages 184–191, Sept 2009. doi: 10.1109/ICSC.2009.31.

[10] Qiuming Tao, Wei Wu, Chen Zhao, and Wuwei Shen. An automatic testing approach for compiler based on metamorphic testing technique. In *Proceedings of the 17th Asia Pacific Software Engineering Conference (APSEC)*, pages 270–279. IEEE, 2010.

[11] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers, Norwell, MA, USA, 2000. ISBN 0-7923-8682-5.

[12] Z. Q. Zhou, S. Xiang, and T. Y. Chen. Metamorphic testing for software quality assessment: A study of search engines. *IEEE Transactions on Software Engineering*, 42(3):264–284, March 2016. ISSN 0098-5589. doi: 10.1109/TSE.2015.2478001.