2013 Third International Workshop on Replication in Empirical Software Engineering Research

The Use of Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews: A Replication Study

Katia Romero Felizardo
Comp. Systems Department
University of São Paulo
São Carlos, SP - Brazil
katiarf@icmc.usp.br

Simone R. S. Souza
Comp. Systems Department
University of São Paulo
São Carlos, SP - Brazil
srocio@icmc.usp.br

José Carlos Maldonado
Comp. Systems Department
University of São Paulo
São Carlos, SP - Brazil
jcmaldon@icmc.usp.br

Abstract*—*Background: **Systematic literature reviews (SLRs) are an important component to identify and aggregate research evidence from different empirical studies. One of the activities associated with the SLR process is the selection of primary studies. The process used to select primary studies can be arduous, particularly when the researcher faces large volumes of primary studies.**
Aim: **An experiment was conducted as a pilot test to compare the performance and effectiveness of graduate students in selecting primary studies manually and using visual text mining (VTM) techniques. This paper describes a replication study.**
Method: **The same experimental design and materials of the previous experiment were used in the current experiment.**
Result: **The previous experiment revealed that VTM techniques can speed up the selection of primary studies and increase the number of studies correctly included/excluded (effectiveness). The results of the replication confirmed that studies are more rapidly selected using VTM. We observed that the level of experience in researching has a direct relationship with the effectiveness.**
Conclusion: **VTM techniques have proven valuable in the selection of primary studies.**

***Keywords*-Controlled Experiment; Experimental Replication, Laboratory Package, Systematic Literature Review, Visual Text Mining.**

I. Introduction

An SLR is a "means of identifying, evaluating and interpreting available research relevant to a particular research question, topic area, or phenomenon of interest" [11]. Since its introduction in the Software Engineering (SE) field in 2004, SLR has been increasingly used as a method for conducting SE-related secondary studies [13, 17]. While the number of SLRs on various topics within the SE domain has been increasing, related studies have also been carried out to report researchers' experiences and consider the challenges faced by those conducting SLRs. For a summary of the problems and experiences reported by various researchers, the reader can refer to the work of Riaz et al. [14]. A particular issue concerns the selection of primary studies, especially when many, mainly irrelevant, search results are returned; consequently, it can be challenging to read, evaluate, and synthesize the state of the art of a particular topic of interest. This issue not only makes the primary study selection process very cumbersome, but could also introduce selection bias [14, 18]. The quality of a primary study selection impacts on the overall quality of the SLR, therefore, to ensure better quality outcomes of the SLR as a whole, it is important to conduct the primary study selection activity as reliably as possible. The traditional process of selection of primary studies is divided into two steps: reading of the title, abstract and conclusions and reading of the full text. In these circumstances the process used to select primary studies can be arduous and time-consuming and must be often conducted manually.

In recent years, there has been an increasing interest in the use of Visual Text Mining (VTM) techniques as supporting tools for SLRs [1, 4, 5, 12]. This interest has been motivated by the fact that humans show strong visual processing abilities. Visual-based techniques make use of these abilities by employing the human system to support knowledge discovery [8]. VTM is an extension of Text Mining (TM), a well-established practice commonly used to extract patterns and non-trivial knowledge from unstructured documents or textual documents written in a natural language [16]. VTM is the association of mining algorithms and information visualization techniques that support visualization and interactive data exploration [3]. In the SLR context, VTM would be potentially beneficial regarding the systematic discovery of relevant primary studies.

The fact that VTM might be useful to explore vast amounts of data motivated Felizardo et al. [6] to employ this technique in their study. They have investigated the use of visualization techniques to help carry out SLRs specifically by employing VTM techniques for the selection of primary studies.

The VTM techniques proposed by Felizardo et al. [6] have been extensively used in their SE laboratory in the context of local research and by international collaborators, such as, the SE group of the University of Groningen/Hollanda in the project titulated "Empirical Software Engineering for Critical Embedded Systems". However, the authors con-

978-0-7695-5121-0/13 $26.00 © 2013 IEEE
DOI 10.1109/RESER.2013.9

ducted only one experiment to compare the performance and effectiveness of doctoral students in selecting primary studies manually and using their approach. One of the potential threats to the internal validity of this experiment is related to the sample used (four subjects). It is often difficult to draw general conclusions from small-sample data. Replications increase the validity and reliability of the results yielded in an initial experiment [7]. Moreover, threats to the validity of an experiment can be addressed by the replication of this experiment. Therefore, the experiment of Felizardo et al. [6] should be replicated with a larger sample of participants.

The remainder of this paper is organised as follows: Section II provides background information on the previous experiment; Section III reviews the replication planning, users' task, metrics and experiment conduction. Section IV summarises the results of the replication in isolation; whereas Section V compares them to those of the first execution. Conclusions are discussed in Section VI.

II. Description of the Previous Experiment

The previous experiment [6] evaluated the use of VTM techniques to support the primary study selection activity. Felizardo et al. [6] investigated whether or not the use of VTM techniques affects the productivity of subjects who select primary studies during an SLR. The remainder of this section summarizes the previous experiment.

A. Research Questions (RQ)

The research questions were:

1) **RQ1:** *Do VTM techniques (document map, edge bundles, and citation network) improve the performance (time taken) of the study selection activity in the SLR process?*
2) **RQ2:** *Do VTM techniques improve the effectiveness (correctness of the inclusion/exclusion) of the study selection activity in the SLR process?*

Based on these research questions, the experiment tested the following hypotheses. In each case both null hypothesis and alternate hypothesis were provided.

$H_{0\_Performance}$: The use of VTM has no effect on the effort devoted to the selection of primary studies.

$H_{A\_Performance}$: The use of VTM has reduced the effort devoted to the selection of primary studies.

$H_{0\_Effectiveness}$: The use of VTM has no effect on the correctness of the primary studies selection task.

$H_{A\_Effectiveness}$: The use of VTM has improved the correctness of the primary studies selection task.

B. Subjects

The subjects were four PhD's students with prior experience in conducting SLRs.

C. Materials

– Datasets: The experiment was organized in two sessions: (i) training and (ii) execution. For training purposes, a small set of data (set 1, containing 20 primary studies) and a specific set of inclusion and exclusion criteria were used. To ensure that first impressions from the training would not interfere with the experiment, a different and larger set of data (set 2, containing 37 primary studies) was used for the execution session.

–VTM techniques: The VTM techniques used were a document map, edge bundles and a citation network.

A *document map* (see Figure 1(a)) is a 2D visual representation of the primary studies that enables users to investigate content and similarity relationships among these studies. Each primary study is mapped to a graphical element represented by a circle. Similar documents, in terms of content (i.e. titles, abstracts and keywords) are placed close to one another and dissimilar documents are positioned far apart. The document map also shows regions where primary studies are grouped according to their similarity.

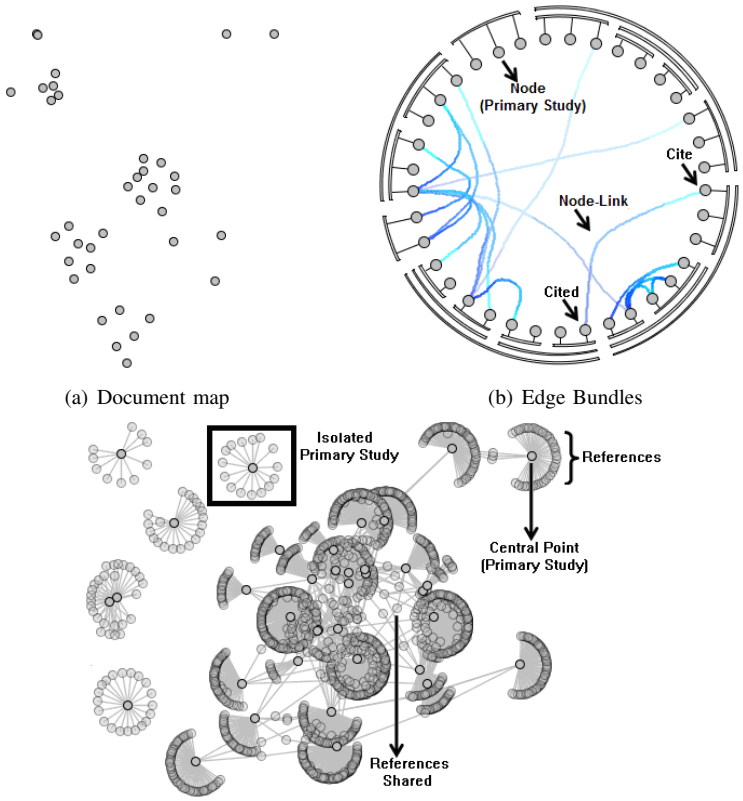
An *edge bundle* is a hierarchical tree visualization technique that shows both nodes and node-links (relationships between nodes) at the same time. The nodes (small circles, Figure 1(b)) are the primary studies and the node-links (blue lines[1]) are the citations among them.

The *citation network* shows the primary studies (central points – circles) and their cited references (circles around the central point connected by edges). It is possible to see citations among the primary studies with their own references and also citations among primary studies and references of other primary studies (shared references – see Figure 1(c)). The process used to create the three visual representations can be found in the work developed by Felizardo et al. [6].

The VTM strategies used to select primary studies are detailed as follows:

- **Document Map:** Three VTM techniques (see Figures 2(a), 2(b), and 2(c)) can be applied to a document map:
  1) **Clusters and Topics** classifies primary studies to identify the regions (clusters) of documents with similar content. Using this technique, clusters are created automatically followed by the formation of their associated topics. These topics are labels that represent the content of the documents contained in the clusters. In order to efficiently include groups of primary studies, a user can concentrate their reading on documents that belong to the clusters labeled with topics that most closely match the SLR's research questions. Similarly, in order to exclude studies, a

[1]In general, visualization techniques employ colour in order to add extra information to a visual representation. We suggest the reading of a colour print version of this paper



(a) Document map

(b) Edge Bundles

(c) Citation Network

Figure 1. Examples of visualizations to support the selection of primary studies in the SLR process.

user can read (perhaps less thoroughly) the documents belonging to clusters labeled with topics that do not match their SLR's research questions. Figure 2(a) shows a document map after the application of the clusters and topics technique. The colour of each point represents the cluster it belongs to and topics appear inside boxes.

2) **Expression Occurrence** changes the colour of each point on a document map to show the frequency of occurrence of specific user-defined expressions in the primary studies. In this case, the colour scale ranges from black (no occurrence) to white (many occurrences). A user can then prioritize their reading towards documents coloured in white (or closer to white) in order to consider whether these documents should be included in the SLR. Conversely, a user can read the documents coloured in black (or closer to this colour) to determine whether they should indeed be excluded from the SLR (assuming, of course, that the user-defined expression is relevant). Figure 2(b) shows a document map after the application of the expression occurrence technique. The white point on the map indicates the maximum occurrence of an expression.

3) **KNN Edges Connection (Neighborhood Relationship)** connects primary studies with their neighbors to support study inclusion by association, i.e. the closer the neighbors of an included study, the more likely to be relevant to the SLR. Likewise, the neighbors of an excluded study are more likely to be irrelevant and should not be included. Figure 2(c) shows the primary studies connected with their neighbors.

- **Edge Bundles:** A relevant paper is usually cited by other papers. The edge bundles show the number of times that a paper has been cited and papers cited many times are strong candidates to become primary studies to be included in the SLR, or at least to be given due attention by the user. On the other hand, papers that are not cited, or cited few times, may be indicative of studies that should not be included in the SLR.
- **Citation Network:** This visualization offers important information besides that related to the initial set of documents, in particular primary studies' references and the connections between papers via a set of references they share. Reference lists from relevant primary studies could be other sources of evidence to be searched [9]. Therefore, papers that share references with a

relevant paper could be more appropriate for inclusion in the SLR. On the other hand, primary studies that are not connected to any other studies (studies that do not share citations or references) are more likely to be irrelevant documents in terms of the research question and may therefore be more readily excluded from the SLR.

Moreover, the user can combine the above-mentioned strategies of exploration using **coordination**, which represents an interaction among the different views (i.e. document maps, edge bundles, and citation network). Using coordination, once a point or a group of documents in a view has been selected, the corresponding point (or points) is then highlighted in the other views. Felizardo et al. [6] implemented a supporting tool, named *Revis – Systematic Literature Review Supported by Visual Analytics* to enable users to explore a collection of documents (primary studies) using VTM techniques.

D. Definition of Users' Task and Metrics

The subjects were randomly split into two groups, one to conduct the study selection activity manually (manual group) and another to use the VTM techniques (tool group). The subjects in the manual group classified the primary studies as included or excluded based on their reading of the abstracts. The subjects in the tool group used the Revis tool and applied the VTM techniques.

To answer the first research question (RQ1), the subjects were required to record the time they spent to conduct the selection activity, i.e. to make their decisions, which does not include the time required to prepare the data for the tool. The time spent on the selection activity was used as an indicator of *performance*. The studies analysed originated from an SLR conducted and were double-checked by an expert in SLRs, whose opinion was used to define the studies that should be either included or excluded. The *effectiveness* was calculated as the number of included/excluded studies that agreed with the opinion of the expert in SLR.

E. Previous Experiment Conduction

The four PhDs who participated in the experiment were randomly split into two groups of two students, one to conduct the selection activity manually (manual group) and another to conduct it using the VTM techniques (tool group). As previously mentioned, the experiment comprised two sessions, training and execution.

Only the subjects involved in the VTM-based task (tool group) participated in the training. During the training session all subjects received an overview of the experiment and an explanation on their task. Subjects from the tool group were trained on how to use the Revis tool. During the training, the subjects' doubts about the tool and the VTM techniques were clarified. Dataset 1 was used for training purposes.

In the execution session, the users utilized dataset 2 to carry out their task. The manual group was given the list of the papers to be selected. The subjects from the tool group received the visualizations (i.e. document map, edge bundles and citation network) containing the same papers used by the manual group. Both groups obtained the inclusion and exclusion criteria and a form to summarize their decisions.

F. Results

The results of the previous experiment showed that the two PhDs in the manual group spent 85 and 54 minutes to correctly classify 25 and 22 studies (of a total of 37), respectively. The two PhDs in tool group spent 30 and 58 minutes to correctly classify 27 and 28 studies, respectively. In summary, the number of studies correctly included/excluded using the manual reading approach was lower than that produced by the VTM techniques, and the use of VTM improved the performance of the study selection activity in the SLR in comparison to a manual reading approach.

III. Replication

The next subsections introduce a more detailed description of the replication.

A. Subjects, Materials, Definition of Users' Task and Metrics

The subjects involved in this replication were 15 graduate students (6 PhDs and 9 Master's students) of an SE course at the USP (University of São Paulo), Brazil. They were randomly divided into two groups: one with 7 subjects (manual group) and another with 8 subjects (tool group). The subjects in the two groups were not significantly different from each other in terms of experience in conducting SLRs.

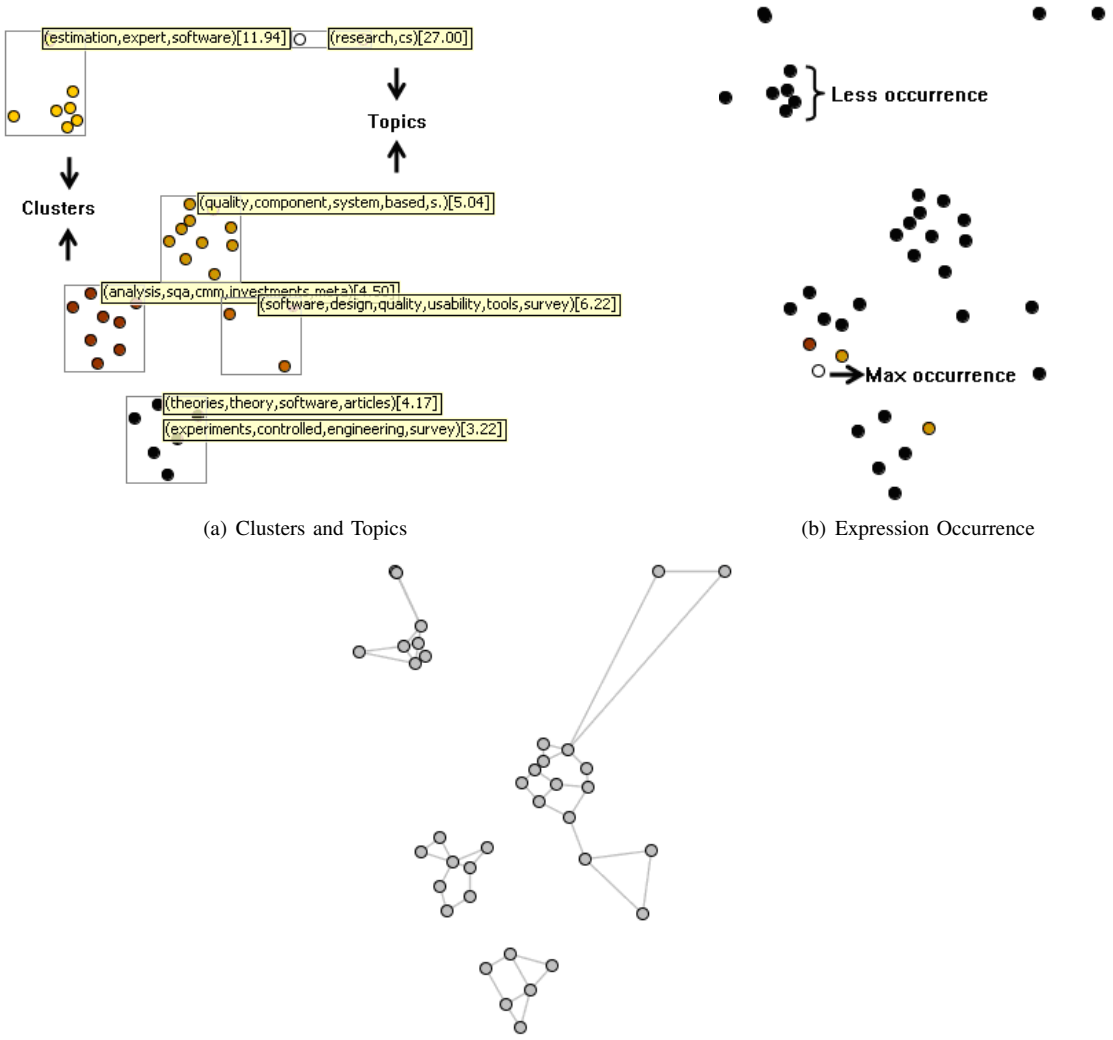
The VTM techniques, datasets 1 and 2 of primary studies, and set of inclusion and exclusion criteria from the previous experiment were used in the replication. The only change introduced in the replication was the increase in the sample size, from 4 to 15 students – Masters and PhDs.

The users' task was to mark the 37 studies from dataset 2 as included or excluded based on the inclusion and exclusion criteria. Subjects from the manual group conducted the study selection activity manually whereas subjects from the tool group were supported by the VTM techniques and the Revis tool.

The same metrics from the previous experiment were used. The subjects were required to record the time they spent to execute the selection activity as an indicator of *performance* and the *effectiveness* was measured as the number of studies correctly included/excluded.

B. Replication Conduction

The design of the previous experiment was duplicated for the replication without changes (2 groups, 2 sessions). Therefore, during the training session, the subjects in the



(a) Clusters and Topics

(b) Expression Occurrence

(c) KNN Connection

Figure 2. The three different VTM strategies that can be applied to the document map.

Table I
GROUPS AND TASKS.

Group	Subjects	Training (60 minutes)	Execution (no-limit time)
Reading papers (Manual Group)	7	no training	Manual selection (dataset 2)
VTM techniques (Tool Group)	8	Revis tool (dataset 1)	VTM selection (dataset 2)

manual group were given the list of papers to be analyzed (set 1), the inclusion and exclusion criteria, and a table to summarize the decision on whether to include or exclude a study.

Only the subjects in the tool group were trained on how to use the Revis tool so that they could be familiar with the VTM techniques and the experiment form. The subjects' doubts about the reading approach, the tool, the VTM techniques and the form were clarified. During the execution session, the subjects selected the primary studies in set 2. As in the training session, the manual group read the abstracts and the tool group used the Revis tool and applied the VTM techniques. No time limit was imposed and the participants were not allowed to communicate with each other. Table I shows a summary of the replication setup.

IV. RESULTS

This section reports the results of the replication addressing the specific research questions (RQ1 and RQ2).

A summary of the results is shown in Table II. To answer the first research question (RQ1), the subjects' performances were measured (see third column of Table II). The time

Table II
SUMMARY OF RESULTS.

Group	ID	Time	Correctly Included/23	Correctly Excluded/14	Total/37	Incorrectly Included	Incorrectly Excluded	Total
Manual	1	70	13	11	24 (64.86%)	3	10	13
	2	70	8	8	16 (43.24%)	5	15	20
	3	63	12	6	18 (48.64%)	8	11	19
	4	60	12	9	21 (56.75%)	5	11	16
	5	68	13	10	23 (62.16%)	4	10	14
	6	95	12	13	25 (67.56%)	1	11	12
	7	65	12	13	25 (67.56%)	1	11	12
Tool	8	62	12	11	23 (62.16%)	3	11	14
	9	63	15	11	26 (70.27%)	3	8	11
	10	35	12	12	24 (64.86%)	2	11	13
	11	59	15	12	27 (72.97%)	2	8	10
	12	66	16	11	27 (72.97%)	3	7	10
	13	38	11	13	24 (64.86%)	1	12	13
	14	56	18	7	25 (67.56%)	7	5	12
	15	57	15	5	20 (54.05%)	9	8	17

average of the manual group was 70.14 minutes and the time of tool group was 54.5 minutes. Both groups had similar standard deviation, i.e., 11.56 minutes (manual group) and 11.60 minutes (tool group). The high standard deviation indicates that the data points are spread out over a large range of values. The results showed that the time spent by the subjects of the manual group to perform the selection activity on the basis of reading the abstracts ranged between 60 and 95 minutes and the time spent by the subjects of the tool group to perform the same activity using the VTM techniques varied between 38 and 66 minutes. The performance of the subjects that used the VTM appeared to be higher than that of the subjects that used the manual approach.

We agree that speed is not important in an SLR if the studies are not correctly included and excluded (effectiveness). Table II (see sixth column) shows a comparison between the VTM and the manual reading approaches in terms of number of studies correctly included/excluded. The average values of studies correctly included/excluded were 21.7 and 24.5 in the manual and tool group, respectively. The standard deviations were 3.54 studies for the manual group and 2.32 studies for the tool group. The number of studies correctly included/excluded using VTM is likely to be as good as that achieved by using the manual reading approach. The average value of studies incorrectly included/excluded in the manual group was 15.14 studies, whereas in the tool group it was 12.5 studies. The standard deviations were 3.28 studies for the manual group and 2.81 studies for the tool group. Regarding primary studies incorrectly judged (see ninth column of Table II), the effectiveness of the subjects that used the VTM appeared to be similar to that of the subjects that used the manual approach.

Boxplots were used to show the distribution of the time spent by the subjects to select primary studies. Figure 3(a) shows that there is no equal variance within the data and the averages for both groups (manual group – reading abstracts and tool group – VTM) are dissimilar. Regarding the number of studies correctly included (see Figure 3(b)), the boxplots show that there is no equal variance within the data and the variance of the manual group (reading abstracts) was higher than that of the tool group (VTM). The same situation was observed (see Figure 3(c)) for primary studies incorrectly excluded (false-negative), i.e. the variance in the number of studies incorrectly excluded by the subjects of the manual group was higher than that of the tool group.

To formally evaluate the results, the Man-Whitney test, also called Mann-Whitney-Wilcoxon test, a non-parametric statistical hypothesis test, was used. Regarding performance, our results (see Table III) show that there is a statistically significant difference between the time averages for the use of VTM and the traditional approach (reading abstracts). Therefore, we can reject the null hypothesis $H_{0\_Performance}$ (the use of VTM has no effect on the effort devoted to the selection of primary studies). The results of effectiveness show that there is no statistically significant difference between the number of primary studies correctly/incorrectly included/excluded using VTM and the traditional approach (reading abstracts). Therefore, we cannot reject the null hypothesis $H_{0\_Effectiveness}$ (the use of VTM has no effect on the correctness of the primary studies selection task).

Table III
SUMMARY OF RESULTS OF THE MAN-WHITNEY TEST.

Variable Compared	*P-Value*
Time (Performance)	0.010
Studies correctly included	0.104
Studies correctly excluded	0.814
Studies incorrectly included	0.628
Studies incorrectly excluded	0.104

Tables IV and V show part of the 37 studies from dataset 2 highlighting the original classification performed by the expert and the one conducted by the subjects who participated in the replication. Table IV shows examples in which most of the classifications performed by the subjects

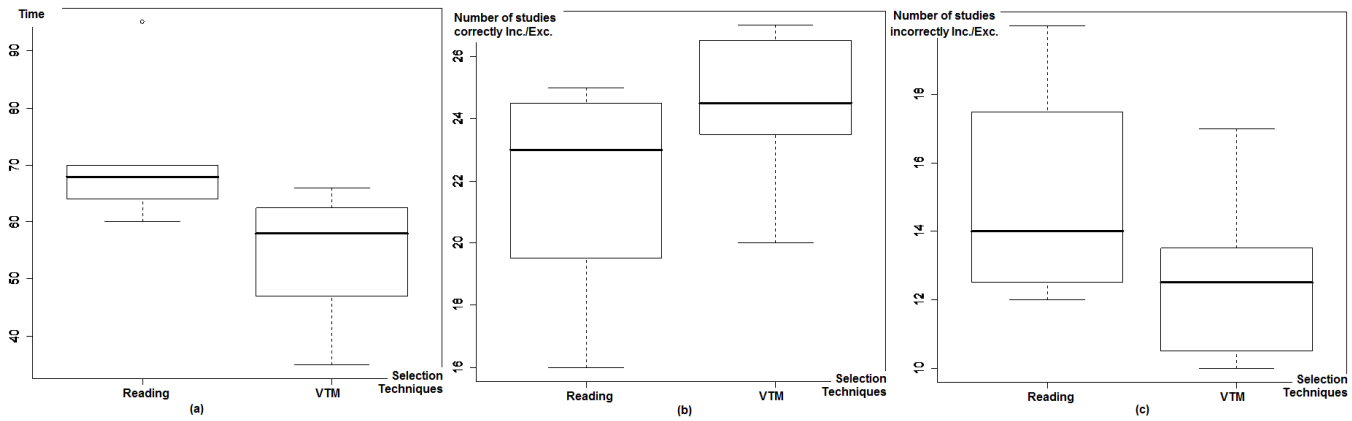


Figure 3. Boxplots showing the distribution of (a) time – performance; (b) effectiveness – studies correctly included/excluded and; (c) studies incorrectly included/excluded – false-negative.

Table IV
PART OF RAW DATA: PAPERS CORRECTLY CLASSIFIED BY THE SUBJECTS.

Group	ID	List of Primary Studies (dataset 2) / Status:I/E						
		#1 I	#2 I	#3 I	#4 E	#5 I	#6 E	#7 E
Manual Group	1	◯	◯	◯	◯	◯	◯	◯
	2	⊗	◯	◯	◯	◯	◯	◯
	3	◯	◯	⊗	◯	◯	◯	◯
	4	◯	◯	◯	◯	◯	◯	◯
	5	◯	⊗	◯	◯	◯	◯	◯
	6	◯	◯	◯	◯	◯	◯	◯
	7	◯	◯	◯	◯	◯	◯	◯
Tool Group	8	◯	⊗	◯	◯	◯	◯	◯
	9	◯	◯	◯	◯	◯	◯	◯
	10	◯	◯	◯	◯	◯	◯	◯
	11	◯	◯	◯	◯	◯	◯	◯
	12	◯	◯	◯	◯	◯	◯	◯
	13	◯	◯	◯	◯	◯	◯	◯
	14	⊗	◯	◯	◯	◯	⊗	◯
	15	◯	◯	◯	⊗	◯	◯	⊗

Legend: ◯ – Correct classification; ⊗ – Incorrect classification.

Table V
PART OF RAW DATA: PAPERS INCORRECTLY CLASSIFIED BY THE SUBJECTS.

Group	ID	List of Primary Studies (dataset 2) / Status:I/E									
		#8 I	#9 I	#10 I	#11 I	#12 E	#13 I	#14 I	#15 I	#16 I	#17 I
Manual Group	1	◯	⊗	⊗	◯	⊗	⊗	⊗	◯	◯	⊗
	2	⊗	⊗	⊗	⊗	⊗	⊗	◯	◯	⊗	⊗
	3	⊗	⊗	⊗	◯	◯	◯	⊗	◯	◯	⊗
	4	⊗	⊗	⊗	⊗	⊗	⊗	◯	⊗	⊗	⊗
	5	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	◯	⊗
	6	◯	⊗	⊗	⊗	⊗	◯	⊗	⊗	⊗	⊗
	7	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	◯
Tool Group	8	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	◯	⊗
	9	◯	⊗	⊗	◯	◯	⊗	⊗	◯	◯	⊗
	10	◯	⊗	⊗	⊗	⊗	⊗	⊗	◯	⊗	⊗
	11	⊗	◯	⊗	⊗	⊗	⊗	◯	⊗	⊗	⊗
	12	◯	◯	⊗	◯	⊗	⊗	◯	⊗	⊗	⊗
	13	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	14	◯	◯	⊗	◯	⊗	◯	⊗	◯	◯	◯
	15	◯	⊗	⊗	⊗	◯	⊗	◯	◯	◯	⊗

Legend: ◯ – Correct classification; ⊗ – Incorrect classification.

match the classification conducted by the experts and Table V provides examples in which most of the classifications were divergent (incorrectly judged).

An interesting point about the data is that paper #5 was classified as included by the expert who conducted the SLR and all subjects who participated in the replication. Paper #4 (see Table IV) was originally classified as excluded by the expert and only one subject (#15 – tool group) did not classify it the same way. Moreover, 93.3% of the subjects agreed on the classification of papers #3, #6 and #7 performed by the expert and 86.6% of the subjects agreed on the classification of papers #1 and #2 conducted by the expert.

Although the expert classified paper #10 (see Table V) as included, 100% of the subjects classified it as excluded, using the traditional approach or VTM techniques. Paper #17 was originally classified as included by the expert and only two participants (#7 – manual group and #14 – tool group) did not classify it as excluded. The results show that 53.3% of the subjects disagreed on the classification of papers #8, #15 and #16 conducted by the expert, 66.6% of the subjects disagreed on the classification of papers #11 and #14 performed by the expert and 80% of the subjects disagreed on the classification of papers #9, #12 and #13 performed by the expert.

One of the potential threats to the internal validity of our replication is related to our assumption that the researchers who originally conducted the SLR made 100% correct decisions on the inclusion and exclusion of studies. However, SLRs conducted by different researchers on the same question sometimes lead to different conclusions. If 50% or more subjects who participated of the replication disagreed with the classification performed by the expert, then the expert's opinion could be incorrect. Based on the scenario described above, a total of 10 papers (papers #8, #9, #10, #11, #12, #13, #14, #15, #16 and #17 – see Table V) could had a different classification. In this perspective, we reanalysed the data calculating effectiveness as the number of included/excluded studies that agreed with the opinion of 50% or more subjects.

The average value of studies correctly included/excluded in the manual group changed from 21.7 to 29 studies. In the tool group the effectiveness changed from 24.5 to 31.12 studies. The average value of studies incorrectly included/excluded in the manual group changed from 8 to 15.14. In tool group the effectiveness did not change (12.5 studies). Considering the "new classification" of the studies, the findings revealed that there is no statistically significant difference *(p-value = 0.29)* between the number of primary studies correctly included/excluded using VTM and the traditional approach. There is also no statistically significant difference *(p-value = 0.09)* between the number of primary studies incorrectly included/excluded using VTM and reading abstracts.

A. Limitations of the study

A limitation of our study is that typically, many SLRs involve a greater number of studies to be considered during the selection stage (more than 100). However, in our replication we used the same SLR employed in the previous experiment, which contained 37 primary studies. We made this choice on the assumption that adding too many studies to our replication could might affect the motivation and performance of the subjects in carrying out the assigned tasks, influencing the results. Although our dataset contains a rather small number of studies, the Revis and VTM techniques can be used in real SLRs, which consider a large number of candidate studies – hundreds and even thousands. In fact, according to the VTM experts, VTM tools work better with more articles [12].

V. Discussions: Relationship with the Previous Experiment

In both experiments, the results show that the incorporation of VTM into the SLR study selection reduced the time spent on this activity. Keim [8] affirms that VTM techniques usually allow a faster data exploration and can help address the challenges that arise in the analysis of large data sets. VTM techniques facilitate the extraction of high-quality information from a large amount of data. The main advantage is the acceleration in the rate at which analyses of the high volume of primary studies can be undertaken.

The results from the previous experiment, in which only PhDs participated, also show that the use of VTM increased the number of studies correctly included/excluded. However, the results of the replication, in which PhDs and Master's students participated, show that the use of VTM did not improve the effectiveness of the study selection activity in comparison to a manual reading approach. One plausible explanation to the non-significant difference in the effectiveness is that the subjects who participated of this replication were partly Master's students. Their level of experience in researching could affect their capability for selecting the studies. Only after a few years of experience in a certain research field, researchers can select studies. In other words, we expected that the inclusion of Master's students would affect the effectiveness of the selection task, therefore, we reanalyzed the data as two separate groups: PhDs and Master's students. The PhDs were subjects #1, #5 and #6 in the manual group, and subjects #11, #12 and #13 in the tool group (see Table II).

A summary of the results is shown in Table VI. The "new" analysis shows that the time spent by PhDs and Master's students using the traditional approach is longer than the time spent using the VTM techniques. PhDs and Masters were faster using VTM than reading the abstracts. The standard deviation of time to conduct the selection activity was high for PhDs in comparison to the Master's students in both groups (manual and tool groups). Subject

Table VI
SUMMARY OF RESULTS: MASTERS VERSUS PHDS.

Independent Variables	PhDs	Masters
Time (Reading Group)	Median = 77.66 min; σ = 15.04 min	Median = 64.50 min; σ = 04.20 min
Time (VTM Group)	Median = 54.30 min; σ = 14.57 min	Median = 54.60 min; σ = 11.37 min
Studies Correctly I/E (Reading Group)	Median = 24.00 studies; σ = 01.00 study	Median = 20.00 studies; σ = 03.90 studies
Studies Correctly I/E (VTM Group)	Median = 26.00 studies; σ = 01.73 studies	Median = 23.60 studies; σ = 02.30 studies
Studies Incorrectly I/E (Reading Group)	Median = 13.00 studies; σ = 01.00 study	Median = 16.75 studies; σ = 03.59 studies
Studies Incorrectly I/E (VTM Group)	Median = 11.00 studies; σ = 01.73 studies	Median = 13.40 studies; σ = 02.30 studies

#6 (manual group) spent the longest time on conducting the activity, therefore, correctly included/excluded 67.56% of the studies. An interesting aspect to be analysed in future work is a possible direct correlation between the time spent on the selection activity and the effectiveness (number of studies correctly included/excluded).

The judgment of PhDs in relation to Master's students (reading group) was better in comparison to studies correctly included/excluded, i.e., 24 and 20 studies, respectivelly (see Table VI – line 4). The PhDs (tool group) correctly included/excluded, on average, 26 studies, whereas the Master's students correctly included/excluded, on average, 23.6 studies. In both groups the number of studies correctly included/excluded by the Masters was lower that of studies correctly included/excluded by the PhDs. The PhDs achieved the best results.

Regarding the studies incorrectly judged, the PhDs (reading group) incorrectly included/excluded, on average, 13 studies, whereas the Master's students incorrectly included/excluded, on average, 16.75 studies (see Table VI – line 6). The PhDs (tool group) incorrectly included/excluded, on average, 11 studies, whereas the Master's students incorrectly included/excluded, on average, 13.4 studies. These results show that the number of studies incorrectly included/excluded decreases with an increase in the researcher's experience in researching. The results also indicate that PhD students have presented a better ability to decide about include/exclude studies.

Our results are in agreement with the findings of other authors. Brereton [2] conducted a case study to explore the effectiveness of second-year undergraduate computer science students in carrying out an SLR. The results suggest that the students found the conduct phase, including the selection activity, more problematic than the planning phase. The author concluded that undergraduates can perform SLRs (specially if undertaken by groups), but the task is clearly quite challenging and time-consuming.

We believe that higher levels of experience in researching will positively influence the effectiveness (i.e., number of primary studies correctly included and excluded) of researchers in the selection of primary studies. Therefore, we grouped data from the previous experiment, in which only PhDs participated, and the data of the six PhDs who participated in this replication in order to validate the "new" hypothesis.

Table VII
SUMMARY OF RESULTS OF THE MAN-WHITNEY TEST: COMBINING DATA FROM THE TWO EXPERIMENTS (PHDS'S STUDENTS).

Variable Compared	*P-Value*
Time (Performance)	0.05
Studies correctly included/excluded	0.04
Studies incorrectly included/excluded	0.04

The Man-Whitney test was used and regarding performance, our results (see Table VII) indicate that, from the PhDs perspective, there is a statistically significant difference between the time averages for the use of VTM and reading abstracts. Regarding effectiveness (studies correctly/incorrectly included/excluded) the results show that there is a statistically significant difference between the effectiveness averages with the use of VTM and the traditional approach (reading abstracts). Therefore, we can affirm that, in the PhD's students context, the use of VTM affects the effectiveness of the primary studies review task.

These results are consistent with those of the previous experiment and suggest that the use of VTM can reduce the time spent to conduct the study selection activity. The evidence from this study also suggests that the level of experience in researching can help to improve the effectiveness.

VI. CONCLUSIONS

An SLR commonly involves a large set of primary studies to be analysed and interpreted. One of the key activities associated with the SLR process is the selection of primary studies, which is a time-consuming process and whose quality impacts on the overall quality of the SLR. VTM can support tasks that involve large collections of data, such as studies collected and evaluated in SLRs.

The main contribution of our research is the replication of a controlled experiment to compare the performance and effectiveness of 15 graduate students (Masters and PhDs) in selecting primary studies using the traditional approach (reading the abstracts) and VTM techniques.

Our results show that the use of VTM is promising in terms of effort reduction (i.e. time spent). VTM techniques usually allow a faster data exploration, therefore the main advantage of using VTM is the acceleration in the rate at which a large volume of primary studies can be reviewed.

In other words, the use of VTM techniques speed up the selection activity.

In terms of selection effectiveness (studies correctly included/excluded) our results suggest that the PhD students' decisions regarding primary study selection are more consistent in comparison to those made by the Master's students. Therefore, evidence suggests that the level of experience in researching impacts on the primary study selection activity in the SLR process. The effectiveness of the primary study selection activity carried out by PhDs using VTM is better than that achieved by reading the papers. VTM techniques can successfully assist the SLR process.

The empirical SE comunity has been addressing several issues related to replications, including the role of lab packages to support them [10, 15]. The lab package of our experiment is available for replications upon request.

ACKNOWLEDGMENT

The authors would like to acknowledge the Brazilian agency FAPESP(Process n.2012/02524-4) for the financial support provided to this research.

They are also indebted to the students that agreed to participating of this replication study.

REFERENCES

[1] S. Ananiadou, B. Rea, N. Okazaki, R. Procter, and J. Thomas. Supporting systematic reviews using text mining. *Social Science Computer Review*, 27(4):509–523, 2009.

[2] P. Brereton. A study of computing undergraduates undertaking a systematic literature review. *IEEE Transactions on Education*, 54(4):558–563, 2011.

[3] M. de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):378–394, 2003. ISSN 1077-2626.

[4] K. El Emam, E. Jonker, M. Sampson, K. Krleza-Jeric, and A. Neisa. The use of electronic data capture tools in clinical trials: Web-survey of 259 canadian trials. *Journal of Medical Internet Research*, 11(1):1–8, 2009.

[5] K. Felizardo, E. Nakwgawa, D. Feitosa, R. Minghim, and J. Maldonado. An approach based on visual text mining to support categorization and classification in the systematic mapping. In *14th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pages 1–10. BCS-eWiC, 2010.

[6] K. Felizardo, N. Salleh, R. Martins, E. Mendes, S. MacDonell, and J. Maldonado. Using visual text mining to support the study selection activity in systematic literature reviews. In *5th International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–10. ACM, 2011.

[7] N. Juristo and O. Gomez. Replication of software engineering experiments. In *LASER Summer School*, volume 7007 of *Lecture Notes in Computer Science*, pages 60–88. Springer, 2010.

[8] D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[9] B. Kitchenham. Procedures for performing systematic reviews. Joint Technical Report TR/SE-0401 (Keele) - 0400011T.1 (NICTA), Software Engineering Group - Department of Computer Science - Keele University and Empirical Software Engineering - National ICT Australia Ltd, 2004.

[10] B. Kitchenham. The role of replications in empirical software engineering – a word of warning. *Empirical Software Engineering*, 13(1):219–221, 2008.

[11] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University, UK, 2007.

[12] V. Malheiros, E. Hohn, R. Pinho, M. Mendonca, and J. Maldonado. A visual text mining approach for systematic reviews. In *1st International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 245–254. ACM, 2007.

[13] K. Petersen and B. Nauman. Identifying strategies for study selection in systematic reviews and maps. In *5th International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–10. IEEE Computer Society, 2011.

[14] M. Riaz, N. Sulayman, M. Salleh, and E. Mendes. Experiences conducting systematic reviews from novices' perspective. In *14th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pages 1–10, Keele University, UK, 2010. BCS-eWiC.

[15] F. Shull, J. Carver, S. Vegas, and J. N. The role of replications in empirical software engineering. *Empirical Software Engineering*, 13(1):211–218, 2008.

[16] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 1 edition, 2005.

[17] H. Zhang and A. Muhammad. An empirical investigation of systematic reviews in software engineering. In *5th International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–10. IEEE Computer Society, 2011.

[18] H. Zhang and A. Muhammad. Systematic reviews in software engineering: An empirical investigation. *Information and Software Technology*, page In Press, 2012.