2015

Visual text mining: ensuring the presence of relevant studies in systematic literature reviews

International Journal of Software Engineering and Knowledge Engineering Sciences and Engineering,Singapore : World Scientific Publishing,v. 25, n. 5, p. 909-928, 2015
http://www.producao.usp.br/handle/BDPI/50214

International Journal of Software Engineering
and Knowledge Engineering
Vol. 25, No. 5 (2015) 909–928

DOI: 10.1142/S0218194015500114

Visual Text Mining: Ensuring the Presence of Relevant Studies in Systematic Literature Reviews

Katia Romero Felizardo

Department of Computer Systems
Federal Technological University of Paraná — UTFPR/CP
Avenida Alberto Carazzai 1640
CEP: 86300-000 – Cornélio Procópio-PR, Brazil
katiascannavino@utfpr.edu.br

Ellen Francine Barbosa*, Rafael Messias Martins†,
Pedro Henrique Dias Valle‡ and José Carlos Maldonado§

Department of Computer Systems, University of São Paulo
ICMC/USP, Avenida Trabalhador São-carlense 400
CEP: 13566-590 – São Carlos-SP, Brazil
**francine@icmc.usp.br*
†*rmartins@icmc.usp.br*
‡*pedrohenriquevalle@usp.br*
§*jcmaldon@icmc.usp.br*

Received 14 January 2014
Revised 26 November 2014
Accepted 6 January 2015

One of the activities associated with the Systematic Literature Review (SLR) process is the *selection review* of primary studies. When the researcher faces large volumes of primary studies to be analyzed, the process used to select studies can be arduous. In a previous experiment, we conducted a pilot test to compare the performance and accuracy of PhD students in conducting the *selection review* activity manually and using Visual Text Mining (VTM) techniques. The goal of this paper is to describe a replication study involving PhD and Master students. The replication study uses the same experimental design and materials of the original experiment. This study also aims to investigate whether the researcher's level of experience with conducting SLRs and research in general impacts the outcome of the primary study selection step of the SLR process. The replication results have confirmed the outcomes of the original experiment, i.e., VTM is promising and can improve the performance of the *selection review* of primary studies. We also observed that both accuracy and performance increase in function of the researcher's experience level in conducting SLRs. The use of VTM can indeed be beneficial during the *selection review* activity.

Keywords: Systematic literature review; visual text mining.

1. Introduction

Systematic Literature Review (SLR) is a "means of identifying, evaluating and interpreting available research relevant to a particular research question, or a topic, or a phenomenon of interest" [1]. Controlled experiments, case studies and surveys are examples of primary studies which compound the information source of SLRs. These empirical studies are grouped and summarized by SLRs, composing the secondary studies [2]. Kitchenham [2] proposed a process for SLRs in Software Engineering (SE) that involves three phases: (i) planning the review, (ii) conducting the review, and (iii) reporting the review. During the planning phase, the need for a review is identified and the review protocol is developed. The protocol includes items, such as sources selection, search methods and keywords, inclusion, exclusion and quality criteria for primary studies. The activities of the second phase include the identification of relevant research, selection of primary studies based on the inclusion and exclusion criteria, *selection review*, assessment of study quality and data extraction. Finally, the third phase comprehends data synthesis and dissemination or reporting of the SLR's results to interested parties including researchers and practitioners.

According to the literature, a potentially problematic aspect of the SLR process is the primary study selection [3], which is both challenging and time-consuming. The selection of primary studies is usually a three-stage process: (i) initially the selection is based on a review of titles, abstracts and keywords. The studies are selected against the inclusion/exclusion criteria defined in the protocol and studies that can answer the specified research questions are included and irrelevant papers are rejected; (ii) full copies of the papers classified as included in the first stage are obtained and selected against the same set of inclusion/exclusion criteria used previously (if necessary, new-more specific-criteria can be defined); (iii) the studies should be reviewed (*selection review* activity) to ensure that relevant studies have not been eliminated.

The *selection review* activity aims to prevent the exclusion of relevant studies and can be conducted in two different ways [1]: (i) performed by two or more reviewers — uncertainties about the inclusion or exclusion should be investigated by sensitivity analysis, which involves repeating the selection activity in the studies divergently classified by reviewers; and (ii) performed by an individual — the researcher should consider discussing her/his decisions with other researchers or, alternatively, the researcher can re-evaluate a random sample of primary studies to determine the consistency of the decisions.

Consequently, the *selection review* activity implies additional effort to re-read the studies, mainly if more than one reviewer is considered. A highly successful approach to support tasks involving the interpretation of a large amount of textual data suitable to be applied to the *selection review* activity is known as Visual Text Mining (VTM) [4]. VTM is an interdisciplinary field of research that combines visualization techniques, human factors (e.g. interaction, cognition and perception) and data mining algorithms to support visualization and interactive exploration of large sets of text documents [5–7].

Felizardo *et al.* [8] have proposed an approach based on VTM techniques to assist the *selection review* activity in SLR. The techniques proposed by the authors offer, for example, clues about the studies to be doubly reviewed for inclusion or exclusion when an SLR is performed by only one reviewer, replacing the random choice strategy. The authors conducted an experiment to compare the performance and accuracy of PhD students in reviewing the selection of primary studies manually and using the VTM-based approach. The major limitation of the original experiment was the small sample used (four subjects). Since the limitations of an experiment can be addressed by performing replications [9], this study focuses on the replication of the initial experiment conducted by Felizardo *et al.* [8], involving a larger sample size of subjects with different levels of experience in researching levels of experience in conducting SLRs.

The remainder of this paper is organized as follows: Section 2 describes the original experiment. In the sequence, Sec. 3 reports the replication we have performed, and Sec. 4 summarizes and discusses the results achieved. Section 5 brings a different view of VTM techniques used in the replication study. Conclusions and future work are discussed in Sec. 6.

2. Description of the Original Experiment

Replication is an essential component of experimentation. The term replication has come into use to refer to a systematic repetition of an original experiment to double-check its results [10]. This definition implies that a replication must be explicitly related to a previous experiment. As mentioned before, the original experiment to assess the utility of VTM techniques in the *selection review* activity was conducted by Felizardo *et al.* [8] in 2012. This section summarizes the original experiment, which involved two research questions:

(1) **RQ1:** Do VTM techniques improve the performance (time spent) of the *selection review* activity in the SLR process?
(2) **RQ2:** Do VTM techniques improve the accuracy (agreement between systematic reviews as to which primary studies they should include) of the *selection review* activity in the SLR process?

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

2.1. *Materials*

2.1.1. *VTM techniques*

The VTM techniques used were content and citation maps. The process used to create the maps can be found in [8, 11].

A content map (see Fig. 1) is a two-dimensional (2D) visual representation, where each study (document) of an SLR is graphically represented as a circle on the plane. The documents' positions in a map reflect the similarity relationships between their

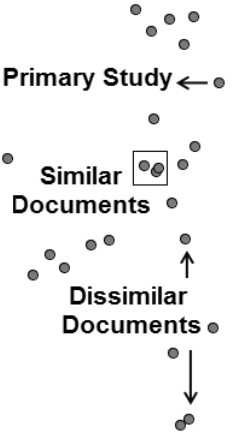


Fig. 1. Content map, where each circle (point) represents one textual document — primary study.

content. Therefore, studies positioned closely together have similar content and documents placed far apart are dissimilar. Details about the stages to create a content map (i.e. pre-processing; similarity calculation; and projection) can be found in [12, 13].

One of the techniques to review the selection activity is to create a content map containing the studies collected and analyzed in an SLR and highlight them using different colors[a] as a strategy to identify the two possible classes of studies — included or excluded (red points are studies excluded from the review and blue points represent the included ones). A clustering algorithm can be applied to the content map, creating groups of highly related (similar) documents. The resulting clusters are analyzed in terms of included and excluded documents in order to find inconsistencies. In this analysis, the possible situations a cluster can be configured and the possible consequences for the review process are:

- **Pure Clusters** — all documents belonging to a cluster have the same classification (all included or excluded). Normally, such cases do not need to be reviewed;
- **Mixed Clusters** — there are documents with different classifications in the same cluster. These cases are hints to the reviewer that there are similar documents with different classifications. The studies grouped there should be reviewed following the manual method;
- **Isolated Points** — there are documents that are not similar to most or all other documents. These cases are also hints to the reviewer, and the isolated study, if classified as included, must be reviewed.

[a]In general, visualization techniques employ color to add extra information to a visual representation. Therefore we suggest the reading of a color printing version of this paper for fully understanding the pictures.

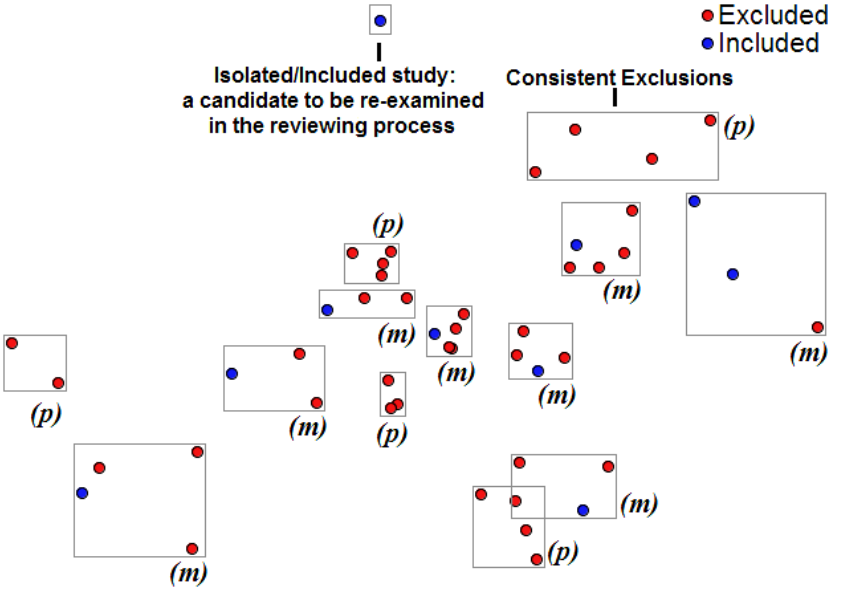


Fig. 2. Example of a content map.

Examples of pure clusters are identified in Fig. 2 as p. Mixed clusters are identified as m. The evaluation of these clusters can be refined with the help of other content-based strategies, detailed in [8].

Another technique to review the selection activity is to use the citation map (see Fig. 3), which shows the primary studies (central points — circles), their cited references (grey circles connected by edges) and how documents are related to each other through direct citations or cross-citations.

Similar to the content map, the primary studies contained in the citation map (see Fig. 4) can be highlighted using different colors to identify the two possible classes of studies: red points are studies excluded from the review, and blue points represent included studies. Using the colored citation map it is possible to visualize, for instance, studies that are not connected to any other, that is, studies that do not share citations. These studies, which are isolated in terms of references, deserve special attention from experts (reviewers) if they are included in the review. Another situation, which requires attention, arises when a highly connected study, sharing citations with included studies, is not selected for inclusion. In this case, important studies may be missing since co-citation is also a valid criterion. In summary, 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 (i.e. they do not share citations or references and are referred to as isolated primary studies) are more likely to be irrelevant documents in terms of a research question and may therefore be more readily excluded from the SLR.

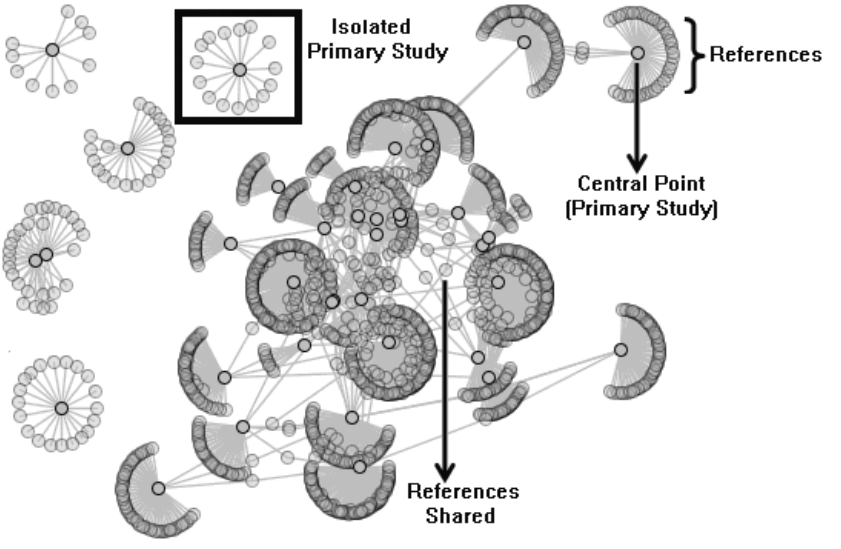


Fig. 3. Citation map: Primary studies that do not share references (isolated primary studies) are disconnected from the other studies in the network.

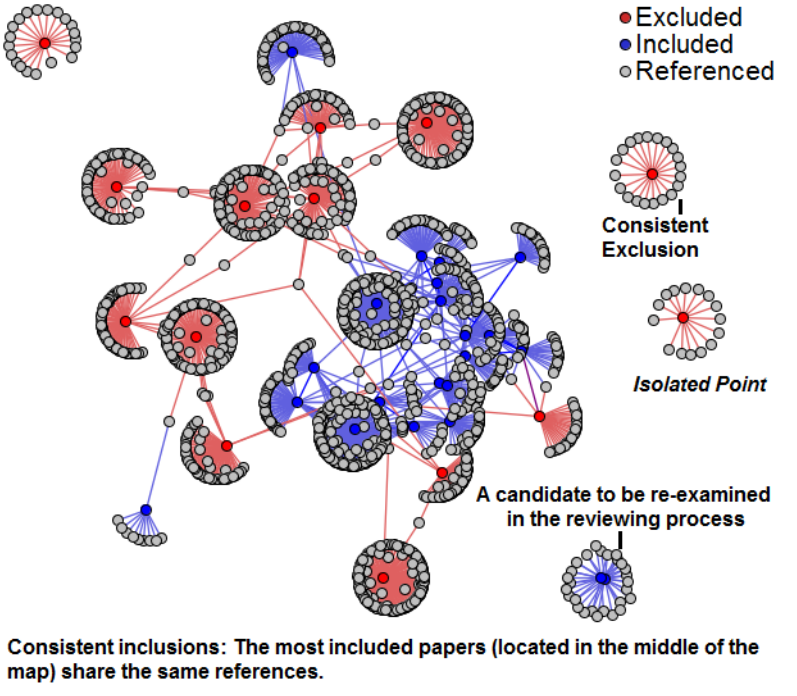


Fig. 4. Example of a citation map.

The Revis tool [8] — Systematic Literature Review Supported by Visual Analytics — enables users to explore a collection of studies using VTM techniques, such as content and citation maps. It takes Revis only a few seconds to create and present content and citation maps with a few hundred documents. Examples of functionalities offered by Revis are: (i) it creates the views: content and citation maps; (ii) it creates clusters; (iii) it allows changing of the color of the studies to represent their classification: included (blue points) and excluded studies (red points), among others.

Datasets

The experiment design was organized in two sessions: training and execution. For training purposes, a small set of data (set 1, containing 20 primary studies — scientific papers on inductive logic programming/case-based reasoning) 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 set of data (set 2, containing 41 scientific papers on concurrent software testing) was used for the execution session. Set 2, including papers of periodics and conferences, originated from an SLR conducted and double-checked by experts in SLRs on the domain of software testing. The purpose of this SLR was to identify testing criteria and testing tools used in concurrent program testing.

2.2. *Definition of users' task and metrics*

The users' task was to review the studies and either confirm the previous classification — conducted by experts — or change them, that is, to ensure that the studies marked as included were in accordance with the inclusion criteria and those marked as excluded were in accordance with the exclusion criteria.

Subjects were required to record the time they spent to perform the task, therefore their performance was calculated using the metric: $\frac{chosen\_and\_relevant\_articles}{review\_time}$. The articles marked as included by two or more subjects who participated in the experiment were considered relevant and taken as the oracle. The accuracy was calculated as the number of studies included that belonged to the oracle.

2.3. *Experiment conduction*

Subjects were split randomly into two groups: (i) Group 1, to conduct the *selection review* activity manually; and (ii) Group 2, to use the VTM techniques. Only the participants involved in the VTM-based task (Group 2) were trained on how to use the VTM techniques and the Revis tool. In the execution session, Group 1 was given the list of the papers to be reviewed, based on their reading of the abstracts and the previous classification of the papers (included or excluded). Subjects from Group 2 received the visualizations (content and citation maps) containing the same papers used by Group 1 (included papers were colored in blue and excluded papers in red).

Both groups were given the inclusion and exclusion criteria and a form to summarize their decisions.

2.4. *Original results*

The main results achieved by performing the original experiment were:

(1) the performance of the subjects that used the VTM is higher than that of the subjects using the manual method; and
(2) there is no difference in accuracy that used VTM or reading the papers.

3. Replication

This section describes the replication of the original experiment (detailed in Sec. 2). The same VTM techniques (content and citation maps), datasets 1 (20 studies) and 2 (41 studies), and the set of inclusion and exclusion criteria from the original experiment were used in the replication. The same users' task (*selection review* manually and using VTM) and metrics from the original experiment were used. The design of the original experiment was duplicated for the replication without changes (2 groups, 2 sessions).

It is important to remember that one of the threats to the validity of the original experiment was related to the small sample used. Therefore, the only change introduced in the replication was the increase in the sample size, i.e., from 4 to 15 students — 10 PhD and 5 Master students — of a SE course at the University of São Paulo (USP), Brazil. We replicated the experiment using 10 subjects with the same level of experience in conducting SLRs of those who participated in the original experiment, and added 4 Master students.

They were divided into two groups: (i) Group 1, containing 8 subjects; and (ii) Group 2: containing 7 subjects. Each group contained 5 PhD students. It is worth mentioning that all the subjects had prior experience in conducting SLRs. No time limit was imposed for the experiment and the participants were not allowed to communicate with each other.

4. Replication Results

Table 1 shows a summary of the results. The time (see the fourth column) spent by the subjects of Group 1 to perform the *selection review* activity on the basis of reading the abstracts varied between 57 and 87 minutes, whereas the time spent by the subjects of Group 2 to perform the same activity using the VTM techniques varied between 32 and 62 minutes.

To answer the first research question (RQ1), the subjects' performances were measured (see the fifth column). The results show that subject 2 reviewed 0.25 articles/min using manual review (Group 1), subject 8, also from Group 1, reviewed 0.13 articles/min, and subjects 11 and 12 reviewed 0.18 articles/min and

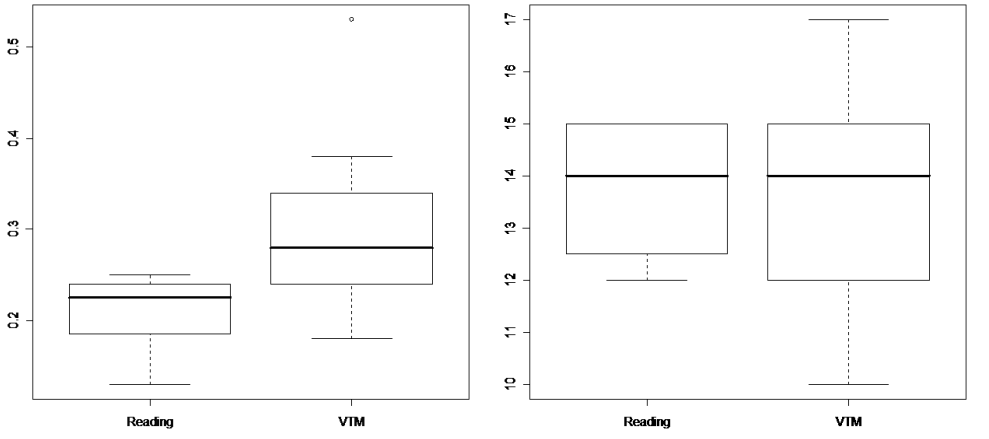
Table 1. Summary of results.

Group	ID	Level of expertise	Time	Performance	Accuracy
Group 1	1	PhD	60 min	0.20 articles/min	12 (66.6%)
	2	PhD	58 min	0.25 articles/min	15 (83.3%)
	3	PhD	65 min	0.23 articles/min	15 (83.3%)
	4	PhD	62 min	0.24 articles/min	15 (83.3%)
	5	PhD	57 min	0.22 articles/min	13 (72.2%)
	6	Master	62 min	0.24 articles/min	15 (83.3%)
	7	Master	75 min	0.17 articles/min	13 (72.2%)
	8	Master	87 min	0.13 articles/min	12 (66.6%)
Group 2	9	PhD	50 min	0.30 articles/min	15 (83.3%)
	10	PhD	34 min	0.38 articles/min	13 (72.2%)
	11	PhD	60 min	0.18 articles/min	11 (61.1%)
	12	PhD	32 min	0.53 articles/min	17 (94.4%)
	13	PhD	56 min	0.26 articles/min	15 (83.3%)
	14	Master	35 min	0.28 articles/min	10 (55.5%)
	15	Master	62 min	0.22 articles/min	14 (77.7%)

0.53 articles/min respectively, applying VTM. The performance of the subjects that used VTM appeared to be better than that of the subjects which used the manual method. Therefore, the results suggest that the use of VTM can help to improve the performance of the selection activity in the SLR in comparison to a manual reading method.

18 articles were marked as "included" by at least two subjects. 14 matched the previous experts' classification of the articles that composed dataset 2. However, four articles were not in accordance with such a classification. The four divergent papers were checked by a senior researcher/specialist in concurrent software testing domain, who classified them as relevant. Therefore, 18 articles were considered the oracle. Table 1 (see sixth column) shows the comparison between the VTM and the manual reading approaches in terms of accuracy (RQ2). Regarding accuracy, researchers 2,3,4 and 6 (Group 1) correctly classified 15 articles of a total of 18 studies included as oracle using manual review, that is, researchers 2,3,4 and 6 correctly classified 83.3% of the articles. Researchers 9 and 13 (Group 2) also correctly classified 15 articles (83.3%) using VTM techniques. Researcher 12 showed a 94.4% precision (17 articles correctly classified).

Boxplots were used to show the distribution of the performance and accuracy of the subjects in reviewing the primary studies. They are based on non-parametric statistics and can help explaining the behavior of the summary statistics. The bar in the box shows the median (central tendency for the distribution) and the length of the box indicates the spread of the distribution. Figure 5(a) shows that there is no equal variance within the data (the variance of Group 2 — using VTM is higher than those of Group 1 — reading abstracts) and that the medians for both groups are different. The highest performance (0.53 articles reviewed/min — an outlier, i.e. a point distant from the rest of the data) was obtained by one subject of Group 2. The second highest performance (0.38 articles reviewed/min) was also achieved by one



(a) (b)

Fig. 5. Boxplots showing the distribution of (a) performance and (b) accuracy.

subject of Group 2. The lowest performance (0.13 articles reviewed/min) was obtained by one subject of Group 1. Regarding accuracy (see Fig. 5(b)), the boxplots show that there is similar variance within the data and that the medians for both groups are the same (18 studies correctly included as an oracle).

To formally evaluate the results the Man-Whitney test [14], also called Mann-Whitney-Wilcoxon test, a non-parametric statistical hypothesis test, was used. Regarding performance, our results have shown that (see Table 2 — Performance) there is a statistically significant difference (P-Value = 0.0487 < 0.05) between the performance averages with the use of VTM and the manual method. Therefore, we can state that the use of VTM can improve the performance of the primary studies review activity.

A plausible explanation to the significant difference in the performance using VTM is that VTM techniques usually allow a faster data exploration helping to address the challenges that arise in the exploration of large datasets [15]. Moreover, VTM techniques facilitate the extraction of high-quality information from a large amount of primary studies usually through content/similarity and citation relationships. It is important to highlight that our aim is not to eliminate the manual method, i.e. reading the abstracts or full-texts, to review the primary studies. Rather, our hypothesis is that exploratory visualization techniques may augment the manual review approach. The employed visual representations can be used to support the decisions made by reviewers regarding inclusions and exclusions.

Table 2. Results for Man-Whitney test.

Variable	*P-Value*	Statistically significant?
Performance	0.0487	Yes (*P-Value* < 0.05)
Accuracy	0.9521	No (*P-Value* > 0.05)

Regarding accuracy, the results have shown that (see Table 2 — Accuracy) there is no statistically significant difference (P-Value = 0.9521 > 0.05) between the accuracy averages with the use of VTM and the manual method. Therefore, we can affirm that the use of VTM exerts no effect on the accuracy of the primary studies review activity. The reason for the no significant difference in the accuracy using VTM may be related with different factors, including: (i) the subjects who participated of the replication were partly Masters. The level of experience of the subjects in conducting SLRs could affect their capability to select studies. Only after a few years of experience in a certain research field, researchers are capable to review studies more effectively [16, 17]; (ii) SLRs on the same topic may reach different conclusions [18]; (iii) How easy it is to review papers for selection in an SLR depends on the domain and the papers to be examined. The subjects were not specialist in concurrent software testing. Sometimes it is really hard to decide whether to include a paper or not, independent of whether using VTM or not.

We compared the replication results with the outcomes of the original experiment. In both experiments, the results showed that the incorporation of the VTM into the SLR study *selection review* can improve the performance of this activity and did not increase the accuracy in comparison to a manual reading method.

4.1. *Discussions*

We believe that a higher level of experience in conducting SLRs positively the accuracy. Particular issues encountered by novice researchers (Master students) involve the primary study selection and *selection review* activities, especially when many, and mostly "irrelevant", search results are returned [19]. SLR is a complex process and demands a range of skills [16, 17]. Often, Masters do not have all the knowledge and skills required (e.g., select/review relevant papers). In principle, a Master's student might be able to learn or acquire all skills needed to conduct an SLR, but if two or more researchers collaborate, there is a greater probability that they will possess a more complete range of skills.

Based on the scenario previously described, we suspected that the inclusion of Master's students affected the accuracy of the *selection review* activity. Thus, we reanalyzed the data as two separate groups, i.e., PhD and Master's students. PhDs were users 1, 2, 3, 4 and 5 in Group 1 and users 9, 10, 11, 12 and 13 in Group 2 (see Table 1). A summary of the results is shown in Table 3.

The analysis has revealed that the judgment of PhDs in comparison to the judment of Master's (manual review – Group 1) was better regarding studies correctly classified, i.e., 14.0 and 13.33 studies, respectively. The PhDs (VTM — Group 2) correctly classified, on average, 14.20 studies, whereas the Master's correctly classified, on average, 12.0 studies. In both groups the number of studies correctly classified by the Master's was lower than that correctly classified by the PhDs, which shows the PhDs achieved the best results.

Table 3. Summary of results: PhDs versus Masters.

Level of expertise	Accuracy	Performance
Manual review - Master students	Median = 13.33	Median = 0.18 articles/min
Manual review - PhD students	Median = 14.00	Median = 0.22 articles/min
VTM - Master students	Median = 12.00	Median = 0.25 articles/min
VTM - PhD students	Median = 14.20	Median = 0.33 articles/min

The performance of PhD students (manual review — Group 1) was, on average, 0.22 articles/min, whereas the performance of Master's was, on average, 0.18 articles/min. The performance of PhD students (VTM — Group 2) was also better in comparison to the performance of Master's, i.e., 0.33 and 0.25 articles reviewed/min, respectively. The results show that, in general, the performance increases with an increase of the researcher's experience level in conducting SLRs, i.e., in both groups the performance of Master's was lower than that of PhD students. Similar to accuracy, the PhD students achieved the best results.

The findings of the current study are consistent with those of Brereton [17] who concluded that undergraduates can perform SLRs (specially if undertaken by groups), but the selection activity is clearly quite challenging and time-consuming.

4.1.1. *Content maps and clusters of SLRs conducted by Master and PhD students*

In order to evaluate if the results of our replication could be confirmed in other contexts, the outcomes of the primary study selection activity conducted by students with different levels of experience and practice in SLRs have been visually analyzed using the concept of pure clusters, mixed clusters and isolated points (see Sec. 2.1). We analyzed content maps of SLRs conducted by three students: (i) one Master's student, who conducted two SLRs. Firstly, the student had no experience in conducting SLRs; (ii) two PhD students, who conducted one SLR each. Both students had experience in conducting SLRs. The generated maps and the results are presented as follows. Note that the number of clusters created for each of the examples was suggested by the Revis tool, based on the total of studies contained in each SLR. The clusters are sequentially numbered.

The clusters of the first SLR conducted by the Master's student are presented in Fig. 6(a). The map is composed of 66 studies. There are 41 red points, which are studies excluded in the second stage (reading the full-text) of the selection activity, and 25 blue points, which are the included studies. All the clusters (100%) are mixed, similar, i.e., studies that have different classifications (included or excluded) but have similar content. Upon examining the content map, it is clear that the included and excluded studies are placed closely together in the map and the maps are very mixed with respect to included and excluded studies. This suggests that the outcome of the primary study selection decision for studies with similar content was not the same, thus pointing towards the possibility of discrepancies in the primary study selection activity carried out by the Master's student. The classification of 25 studies (37.87%)

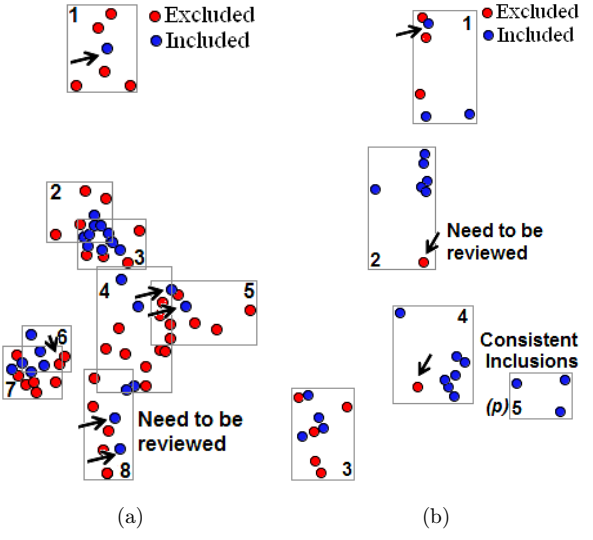


Fig. 6. Content map and clusters of SLRs conducted by Master's students: (a) 100% of mixed clusters; (b) 80% of mixed clusters.

should be reviewed. The points highlighted by an arrow in Fig. 6(a) are examples of these studies.

The clusters of the second SLR conducted by the same Master student are presented in Fig. 6(b). The map is composed of 33 studies analyzed in an unpublished SLR on games engines. There are 23 red points (excluded), and 10 blue points (included). Note that 4 of a total of 5 clusters (80%) are mixed whereas only 1 (20%) is a pure cluster, therefore the mixed clusters are evident. In spite of the experience acquired in the previous SLR, there is an small increase in the number of pure clusters. This result may be explained by the fact that the student was not specialist in game engines. There are similarities between the first and second conduction, such as the large number of mixed clusters, however, in this case, an improvement can be observed in the number of studies that should be reviewed (9 studies — 27.27%). The points highlighted by an arrow in Fig. 6(b) are examples of these studies.

The clusters of the SLR conducted by one of the PhD students are presented in Fig. 7(a). The map is composed of 40 studies. There are 32 red points — excluded studies, and 8 blue points — included studies. Note that 5 of a total of 7 clusters (71.42%) are pure, i.e., studies that have same classification (included, blue points; or excluded, red points) and are similar in terms of content. The classification of 4 studies (10.0%) should be reviewed following the manual method, re-reading the full-text.

The clusters of the SLR conducted by the other PhD students are presented in Fig. 7(b). The map is composed of 109 studies. There are 71 red points — excluded studies, and 38 blue points — included studies. Note that 8 of a total of 10 clusters (80.0%) are pure (see Fig. 7(b)). The classification of 12 studies (11.0%) should be

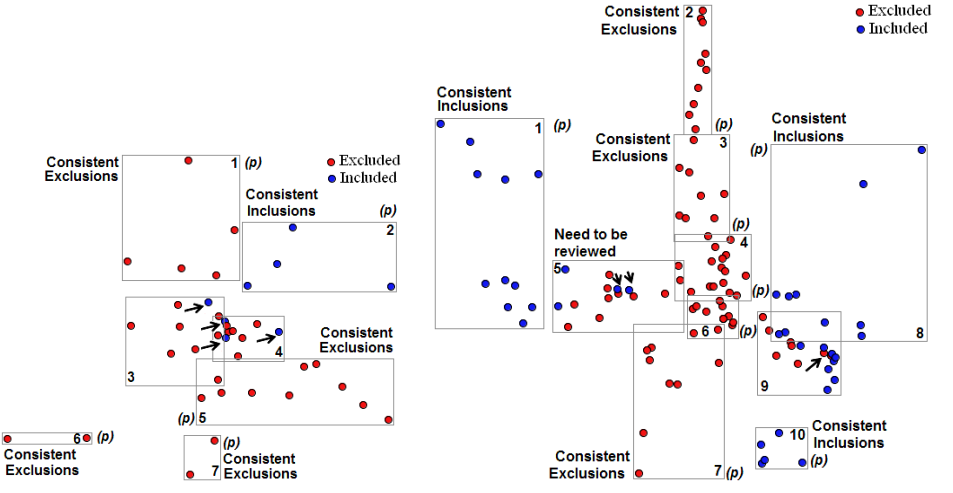


Fig. 7. Content map and clusters of SLRs conducted by PhD students: (a) 28.58% of mixed clusters and 71.42% of pure clusters; (b) 20% of mixed clusters and 80% of pure clusters.

reviewed. Points highlighted by an arrow in Fig. 7(b) are examples of these studies. Upon examining both content maps related to PhD students, it is clear that the included and excluded studies are not as mixed as it was for the Master's student. This suggests that the PhDs' decisions regarding primary study selection for studies with similar content were more likely to be consistent.

We have investigated the possible effect of researchers' experience on the outcomes of primary study selection when conducting SLRs. For this purpose, content maps and clusters were created after collecting data on SLRs conducted by students having varying levels of experience with conducting SLRs and with research in general. The results suggest that there is a relationship between the researcher's level of experience in conducting SLRs and the outcomes of the study selection activity. This was evident in the maps and clusters, where the regions of included studies were better separated from the regions containing excluded studies for PhD students compared to that found for Master's student.

Given the results based on the maps, one might ponder upon the reason behind the difference. One reason, and the one assumed here, is the difference in experience level. Another possible reason may be due to the time available for conducting SLRs. The time period for the completion of a Master's degree is generally one to two years, clearly less than the available time to PhD students. Therefore, one might argue that the overall available time for conducting an SLR could also impact the study selection outcomes. The PhD students of our example had already completed a Master's degree comprising a research dissertation, and so they had some hands-on prior experience with literature reviews. In contrast, the Master's student was completely novice to research.

On considering Kitchenham's guidelines for conducting SLRs (Kitchenham and Charters 2007), in which the importance of properly conducting the primary study selection activity has been emphasized, one can deduce that the quality of the primary study selection step impacts the overall quality of the SLR. Therefore, in order to ensure better quality outcomes of the SLR as a whole, it is important to conduct as completely and reliably as possible the primary study selection step. Our results suggest that the quality of the primary study selection activity carried out in SLRs conducted by Master students is unlikely to be as good as that of PhD students. The content maps can visually support the primary selection step in that a researcher

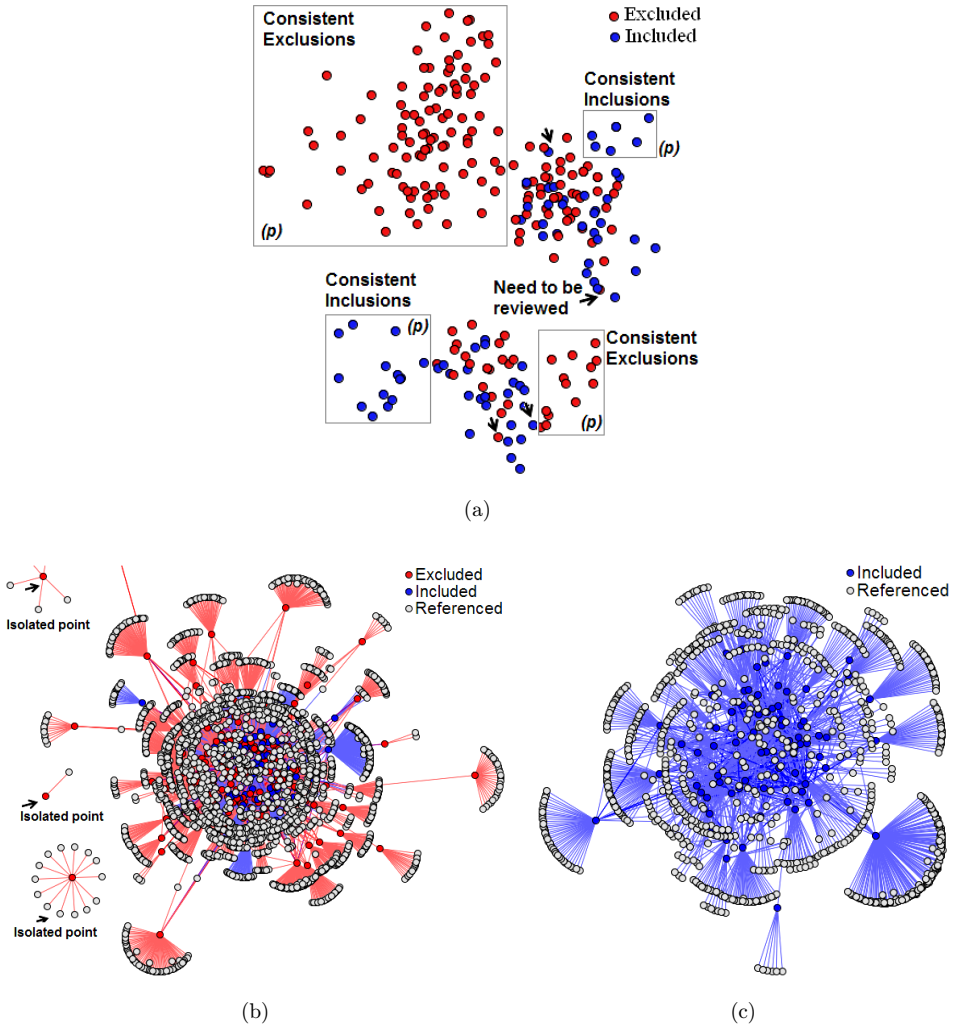


Fig. 8. Example of an SLR containing more than 100 studies in two different perspectives: (a) content map and (b–c) citation map.

conducting an SLR can see and then investigate whether studies having similar content (and so placed together) have different or the same selection outcome. With the help of content maps, one can revisit the primary selection step and possibly improve the quality of this activity and hence the overall quality of the SLR.

The examples used in Sec. 4.1.1 contain a rather small number of studies (dozens or a few hundreds articles). However, in most secondary studies, especially systematic mappings [20, 21], a large number of candidate studies are considered — even reaching the thousands. According to VTM experts [22], VTM tools can also be useful in SLRs with larger numbers of articles. In the sequence, we present a example to illustrate this assertion.

We selected an unpublished SLR on software testing methods formed by 264 studies — 190 excluded and 74 included. Figure 8 shows the content and citation maps related to this SLR. The regions marked by squares in the content map (see Fig. 8(a)) bring together sets of studies with the same classification and which are also similar in terms of content, pointing to clues that the decision for inclusion/exclusion of these groups of studies are consistent. Points indicated by arrows are examples of studies that need to be reviewed to ensure their classification.

Based on another perspective, it is possible to observe in the citation map (see Fig. 8(b)) that the studies that do not share citations (isolated points), as expected, were excluded. In order to facilitate the visualization, a new version of the citation map containing only the included studies is represented in Fig. 8(c). This figure shows that all points are connected, reassuring the reviewer about his/her choices for inclusion, since the inclusion of a study that shares references with another included (relevant) study is a recommended practice.

5. Using VTM to Update SLRs

The same VTM techniques validated in our replication study can be used in different SLR settings. One of these perspectives is related to supporting users in updating SLRs. In order to verify the application of VTM for the inclusion of "new" studies in an SLR, we have also conducted a case study using a real and previously published SLR on software effort estimation models [23]. This SLR, which we shall call test-SLR, is formed by 185 potentially relevant studies, including duplicates (147 papers excluded, 28 repeated and 10 included).

MacDonell *et al.* [24], the authors of the test-SLR, noted that they did not include one relevant study in their SLR because this paper had not been published yet, although it was in press. We used it here to create the content and citation maps (see Figs. 9 and 10), to test where they would be positioned by the VTM techniques.

Figure 9 shows that the "missed" paper (highlighted with an arrow) was allocated in the content map next to other included papers, a strong indication that it should receive the same treatment. Figure 10 shows that the "missed" paper also shares citations only with included papers (blue points), illustrating quite effectively the utility of the VTM representations. In both visualizations the reviewer has clues that

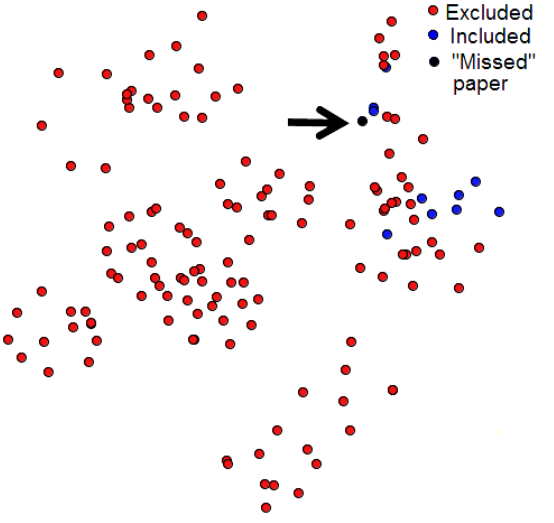


Fig. 9. Content map after the inclusion of the "missed" paper, which is a strong candidate to be considered as an included paper because it is similar in content with other papers already included.

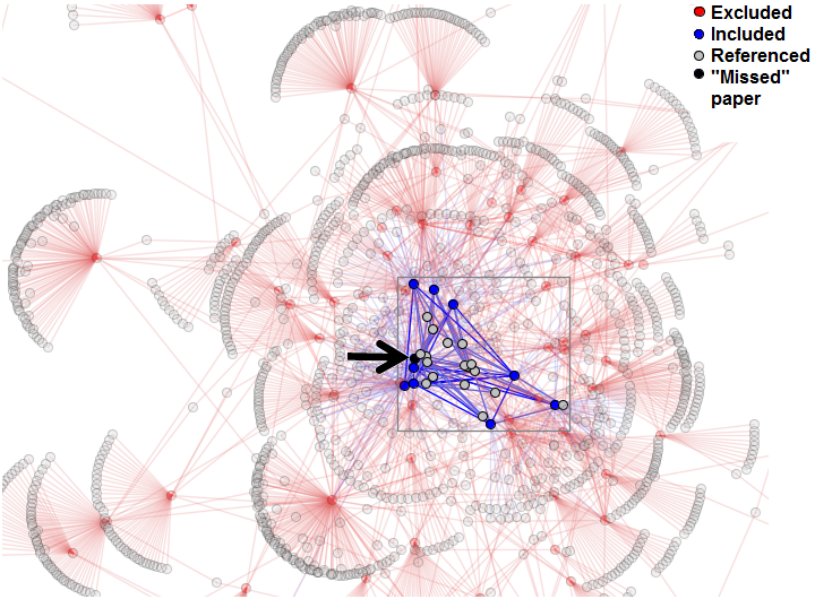


Fig. 10. The "missed" paper also shares citations with other included papers, a strong indication that it should be considered in the same light regarding the inclusion/exclusion decision.

the "missed" paper is relevant in the context of the SLR question and that it should be included.

6. Conclusions and Future Work

Visual data exploration has been used in many applications (e.g., fraud detection and data mining) [15], and it has been applied here to support improved data selection in the SLR context of SE research. Note, however, that it is not our intention to eliminate the manual method — reading the abstracts and full texts — to select primary studies. Rather, it is our view that exploratory visualization techniques may augment the manual selection approach, helping users to better understand the primary studies. In particular, reviewers can use the VTM techniques to judge their inclusion and exclusion decisions. That is, the employed visual representations can be used to complement the decisions made by reviewers, giving support to guide the researchers to a consistent treatment regarding inclusions and exclusions. In a scenario of a group of reviewers conducting an SLR, the VTM techniques are valuable tools to reach a conclusion on what should and should not be included. The employed visual representations can be used to compare and analyze the decisions made by the different reviewers, giving support to guide the group to a common sense about the inclusions and exclusions. In the special case of an SLR executed by only one reviewer, the VTM techniques eliminates the need for random choices of papers to be re-evaluated. Instead, such a selection is based on similarities and citations criteria revealed by the content and citation-based layouts.

Using VTM techniques users can explore different visual representations of the primary studies to have additional and complementary information that are not readily available directly from reading the study abstracts (e.g., similarity relationships, citations between primary studies). The visual representations can give solid clues about which studies should be checked, reducing the amount of documents that need to be re-evaluated and the time spent in the whole process. In addition, the manual approach implies additional effort to select studies for review.

The main contribution of this research is the replication of a controlled experiment to compare PhD and Master's students performance and accuracy in reviewing primary studies manually and using VTM techniques. The results show that the answer to RQ1 is "*Yes*" — suggesting that the performance of the subjects that used VTM is higher than that of the subjects that used the manual method. VTM techniques usually allow a faster data exploration, therefore the main advantage of using VTM is the acceleration of the rate at which the review of a large volume of primary studies can be undertaken. In other words, the use of VTM techniques speed up the *selection review* activity. In terms of accuracy, our results suggest that the PhD students' decisions regarding primary study *selection review* are more consistent in comparison to those made by the Master students.

One of the potential threats to the internal validity of our study is related to the fact that, typically, many SLRs involve a greater number of studies to be considered

during the review stage (more than 100). However, we chose to use in our replication the same SLR used in the original experiment, which contained 41 primary studies. We made this choice on the assumption that adding too many studies to our replication could similarly influence the results, affecting the motivation of the subjects to carry out the assigned tasks.

The key disadvantage of introducing VTM in the SLR process is the additional knowledge required, i.e., reviewers will need to become familiar with the visual tools, but the advantage, in our opinion, will be that it will result in improving the quality of the SLRs conducted by Master's students in particular.

The empirical SE comunity has been addressing several issues related to replication, including the role of laboratory packages to support replications [25]. The laboratory package of our experiment is available for replications upon request.

The presented results are promising and reveal the benefits of using VTM techniques in supporting the conduction of SLR. Further replications involving more subjects and a wider dataset will be conducted in order to identify effects provided by the use of VTM techniques. Conclusive results regarding the use of VTM to support the SLR process should be achieved before widely adopting the use of VTM techniques.

Acknowledgments

The authors would like to acknowledge the Brazilian agency FAPESP for the financial support provided to this research.

References

1. B. A. 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.
2. B. A. 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.
3. H. Zhang and A. B. Muhammad, Systematic reviews in software engineering: An empirical investigation, *Information and Software Technology* (2012).
4. A. A. Lopes, R. Pinho, F. V. Paulovich and R. Minghim, Visual text mining using association rules, *Computers and Graphics* **31**(3) (2007) 316–326.
5. M. C. F. de Oliveira and H. Levkowitz, From visual data exploration to visual data mining: A surveym, *IEEE Transactions on Visualization and Computer Graphics* **9**(3) (2003) 378–394.
6. D. A. Keim, F. Mansmann, J. Schneidewind and H. Ziegler, Challenges in visual data analysis, in *IV Conference on Information Visualization* (2006), pp. 9–16.
7. P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining* (Addison-Wesley, 2005).
8. K. R. Felizardo, G. F. Andery, F. V. Paulovich, R. Minghim and J. C. Maldonado, A visual analysis approach to validate the selection review of primary studies in systematic reviews, *Information and Software Technology* **54**(10) (2012) 1079–1091.

9. N. Juristo and O. S. Gomez, Replication of software engineering experiments, in *LASER Summer School*, LNCS, Vol. 7007 (2010), pp. 60–88.
10. N. Juristo and S. Vegas, Using differences among replications of software engineering experiments to gain knowledge, in *3rd International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2009, pp. 356–366.
11. K. R. Felizardo, E. F. Barbosa and J. C. Maldonado, A visual approach to validate the selection review of primary studies in systematic reviews: A replication study, in *XXV International Conference on Software Engineering & Knowledge Engineering*, 2013, pp. 141–146.
12. K. R. Felizardo, N. Salleh, R. M. Martins, E. Mendes, S. G. MacDonell and J. C. 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*, 2011, pp. 1–10.
13. F. V. Paulovich and R. Minghim, Text map explorer: A tool to create and explore document maps, in *X International Conference on Information Visualisation*, 2006, pp. 245–251.
14. M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods* (Wiley-Interscience, 1999).
15. D. A. Keim, Information visualization and visual data mining, *IEEE Transactions on Visualization and Computer Graphics* **8**(1) (2002) 1–8.
16. P. O. Brereton, B. A. Kitchenham, D. Budgen, M. Turner and M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, *Journal of Systems and Software* **80**(4) (2007) 571–583.
17. P. Brereton, A study of computing undergraduates undertaking a systematic literature review, *IEEE Transactions on Education* **54**(4) (2011) 558–563.
18. F. Peinemann, N. Mcgauran, S. Sauerland and S. Lange, Disagreement in primary study selection between systematic reviews on negative pressure wound therapy, *BMC Medical Research Methodology* **8** (2008) 16.
19. 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*, 2010, pp. 1–10.
20. K. Petersen, R. Feldt, S. Mujtaba and M. Mattsson, Systematic mapping studies in software engineering, in *12th International Conference on Evaluation and Assessment in Software Engineering*, 2008, pp. 1–10.
21. K. Petersen and B. A. Nauman, Identifying strategies for study selection in systematic reviews and maps, in *5th International Symposium on Empirical Software Engineering and Measurement*, 2011, pp. 1–10.
22. V. Malheiros, E. Hohn, R. Pinho, M. Mendonca and J. C. Maldonado, A visual text mining approach for systematic reviews, in *1st International Symposium on Empirical Software Engineering and Measurement*, 2007, pp. 245–254.
23. S. G. MacDonell and M. J. Shepperd, Comparing local and global software effort estimation models — reflections on a systematic review, in *International Symposium on Empirical Software Engineering and Measurement*, 2007, pp. 401–409.
24. S. MacDonell, M. Shepperd, B. A. Kitchenham and E. Mendes, How reliable are systematic reviews in empirical software engineering? *IEEE Transactions on Software Engineering* **36**(5) (2010) 676–687.
25. F. Shull, J. Carver, S. Vegas and N. Juristo, The role of replications in empirical software engineering, *Empirical Software Engineering* **13**(1) (2008) 211–218.