OULUN YLIOPISTO
UNIVERSITY of OULU

Supporting Study Selection of Systematic Literature Reviews in Software Engineering with Text Mining

University of Oulu
Information Processing Science
Master's Thesis
Qianhui Zhong
Date 07.04.2017

Abstract

Since Systematic Literature Review (SLR) was introduced to Software Engineering (SE) field in 2003, it has gained plenty of attention and became a quite popular method to collect scientific evidence. However, SLR is a very time-consuming process, especially the study selection procedure.

This study was concentrating on how to support study selection process. Design science research methodology was used. An SLR was conducted to reveal the existing evidence of supporting SLR process. The review results indicated that study selection supporting evidences can be classified into three categories: tools, information visualization techniques, and text mining techniques. However, in SE field all those existing methods or tools are still in early stage, either the effectiveness is not good or the proposed tools and techniques are too difficult to use. Thus, there is a big requirement of supporting SLR, especially study selection procedure in SE.

Based on the SLR results, an iterative framework was proposed from a new perspective: refining the selection criteria to improve the selection results. There is research has indicated that inadequate SLR protocol can influence the SLR quality. Different from the previous research work, which mostly adopts automating or semi-automating selection process with various techniques, this new framework will be focused on refining the inclusion/exclusion criteria by extracting valuable terms from candidate papers. Text mining technique was chosen as the potential technique applied in the extracting process.

To prove the effectiveness of the proposed framework, two SLR study selection experiments were conducted and the results were evaluated with precision, recall and F_1 score. The first experiment used the existing SLR that was conducted in this thesis, while the second one used another existing large SLR. The results from both experiments indicates that the proposed framework performs better comparing to the traditional method, with higher F_1 score which combines recall and precision.

In conclusion, this thesis found a new effective method to support the SLR study selection process, which is aiming at refining the selection criteria before conducting reviewing work. Further researches can focus on improving the information extracting or combining the proposed iterative framework with other study selection support method.

Keywords
Systematic literature review, Software Engineering, study selection process, positive, negative, selection criteria, text mining, recall, precision, F_1 score

Supervisor
Dr. Jouni Markkula

Foreword

This study has been a big challenge to me for the past two years, despite of the fact that I had some research experience during my previous bachelor degree study. The whole master thesis work requires quite independent thoughts and strongly self-management. I believe I can benefit a lot from all the efforts even in the future work. Also, it is a good opportunity for me to conduct researching work independently. Moreover, I am familiar with SLR process and some related topics. Above all, I am glad that the method I proposed is creative and effective in several experiments.

I must say I sincerely thank my supervisor Jouni, who did give me so much support from both the practical and spiritual side. I received most of my inspirations in our discussion. I own a million thanks to him. Also, I would like thank my friends and families who have been supporting me all the time.

Lastly, three years studying in university of Oulu was the happiest time in my life, getting to know a wonderful country, various cultures, and impressive people. I learnt much knowledge in Software Engineering field as well. I appreciate all these so much.

Qianhui Zhong

Helsinki, Mar 22, 2017

Abbreviations

ATR	Automatic term recognition
DSR	Design science research
EC	Exclusion criteria
FN	False negative
FP	False positive
IC	Inclusion criteria
SE	Software Engineering
SLR	Systematic literature review
SLuRp	Systematic Literature unified Review Program
SM	Systematic Mapping
SM-VTM	Systematic Mapping based on Visual Text Mining
StArt	State of the Art through systematic review
TM	Text mining
TP	True positive
IT	Information technology

Contents

1. Introduction

Since Kitchenham introduced the concept of evidence-based software engineering from Medical field to Software Engineering (SE) field in 2004, systematic literature review (SLR) has gained huge attention and became popular among software engineering researchers (Kitchenham, 2004; Fabio, 2011). SLR is a way of gathering knowledge about a certain research question or topic (Kitchenham, 2004). The most common reasons for undertaking SLR are to collect the available evidence concerning a treatment or technology, to identify gaps in the existing research to suggest areas for later investigation, and to provide a framework in order to identify new research activities (Brereton & Kitchenham, 2007).

According to Kitchenham's guideline (2004), SLR involves several discrete activities. SLR process contains three main phases: review planning, review conducting, review reporting. Review planning phase is consisting of identification of the need of the review and development of a review protocol. Review conducting is a time-consuming phase, which associates five stages: research identification, primary studies selection, study quality assessment, data extraction & monitoring, data synthesis (Kitchenham, 2004). However, in SE domain, various SLR research projects can have different study selection process, while most of them are highly similar; usually a SLR process has three rounds, which are consist of title & keywords study selection, abstract study selection and introduction & conclusion study selection (Kitchenham, 2007).

However, comparing to Medical researching field, SLR process in SE domain is not mature enough (Kitchenham & Brereton, 2013). Even though Kitchenham has provided a guideline of preceding SLR process, which so far has been widely used in Software Engineering field, SLR is still facing big challenges (Kitchenham, 2007; Kitchenham, 2013). SLR is known as a slow and resource-intensive robust process, which requires a large amount of resources (Guy, 2014). The most tedious and heavy work recognized in SLR process is the primary study selection phase (Adeva & Atxa, 2014). Currently, most of the SLR work is still done manually by human experts. However, it takes plenty of time and can be biased because of the heavy work; this issue is getting more critical due to the rapid growing of research literatures (Sophia & Rob, 2007).

Therefore, heavy resource consuming and biased results are the two most serious issues in SLR, which mostly appear in the study selection stage (Sophia & Rob, 2007). In the study selection phase reviewers need to go through different parts of the identified potential studies and select the most related primary studies during the process. Hundreds or thousands of papers need to be reviewed in order to identify primary studies that related to the research question, but unfortunately SLR is still in lack of good automatic techniques/tools support (Ahmed, 2014).

The objective of this thesis is to research how to support the study selection process of SLR in Software Engineering. There is a general paper structure among the scientific papers in SE papers: titles, keywords, abstracts, introduction and conclusion; the standard paper structure makes it possible to automate or semi-automate the study selection process. Existing research work in supporting SLR process can be categorized into two types. The first one is facilitating the study selection process by providing easier study selection interface or simplifying the operation by means of software and the other one is facilitating the study selection with the assistance of text processing techniques (Kitchenham, 2013). It is notable that currently many studies have achieved better performance in optimizing the study selection phase, which shows less time and better accuracy rate; however, none of them have shown enough effectiveness (Kitchenham, 2013; Zubidy & Carver, 2014). All those existing methods or tools are still in an early stage, which is also the main reason why there is no unified study selection techniques/tools for SLR. Another problem of the previous tools and techniques is that many of them are too difficult to use due to the requirement of extra knowledge, like text mining. For example, a tool named 'Revis' is used to support visual text mining technologies in the study selection of SLR (Felizardo & Salleh, 2011). The precondition is to use Revis is that researchers should have some background in visual text mining mechanism.

To sum up, there is still no unified techniques/tools to support SLR process. The study topic of this thesis is limited to the study selection process so as to more concentrating on the real problem. Literature review of supporting SLR indicates that text mining and visualized techniques are the most popular techniques (Marshall & Brereton, 2014). Techniques like natural language processing and text mining are highly recommended for this study selection phase (Cruzes, 2007; Kitchenham, 2013). Therefore, applying various techniques gives possibilities to replace the human manual work. The most likely stage of automating the manual work is the "primary studies selection", which costs lots of time and may lead a biased result.

Design-science research (DSR) method was selected as the main research method of this thesis. DSR model is classified into three parts: problem identification, solution design and evaluation (Hevner, 2004). For each part, different research methods were applied. Systematic Literature Review is used for discovering the existing evidence of supporting SLR. The SLR process follows Kitchenham's SLR guideline in SE. Two reviewers (thesis author and supervisor) were responsible for the study selection process to avoid biases. In the evaluation stage, experimental evaluation was performed to validate the SLR support solution.

2. SLR in Software Engineering Field

A systematic literature review is used for identifying, evaluating and interpreting the existing research that is relevant to a certain research topic or question, or phenomenon of interest (Kitchenham, 2007). It is useful for exposing the existing research results and providing a good view to conduct the following research work; also, it can be used for identify any gaps in the existing research so as to provide suggestion for the further investigation (Kitchenham, 2007). In Software Engineering research field, SLR provides a rigorous method for evidence analysis that related to a certain topic of SLRs (Marshall & Brereton, 2013). The traditional SLR refers to the manual review process, while nowadays SLR automation has been developed to ease the process.

2.1 Traditional systematic literature review process

In Kitchenham's SLR guidelines (2007), SLR conducting process adapted for Software Engineering consist of three phases: review planning, review execution and review reporting. In Figure 1, a traditional SLR process is presented. Typically, the phases and steps are appearing to be sequential; however, it is important to recognize that many of the steps involve iteration. Many of the activities are planned and specified in the phase 1 as protocol, but they may need to be refined in the later process.

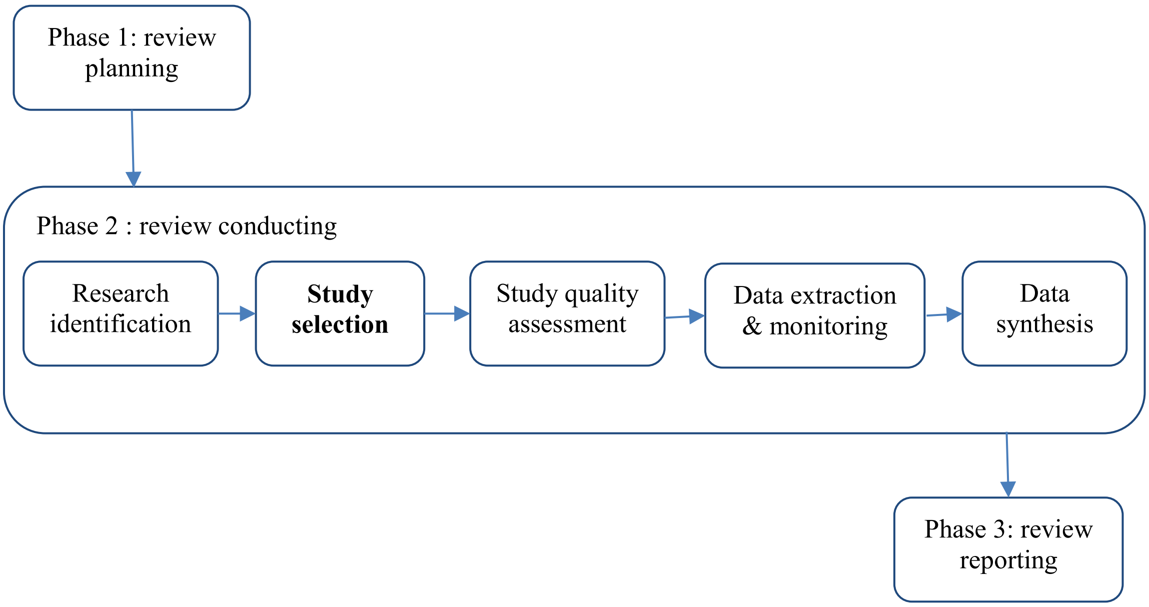


Figure 1. Traditional SLR process

In the review planning phase, two stages are included: identification of the need for a review, and development of a SLR protocol. Before performing a SLR, researchers should make sure that this SLR is necessary. The demand for a SLR derives from the requirement of researchers to aggregate all available information regarding some phenomenon, in a rigorous and impartial manner. A review protocol is predefined and specifies the methods that would be used in the review process, which is necessary and reduces the possible researcher bias. Also, it is very common that the review protocols go through peer review. All the elements of the SLR and some additional planning information should be included in the protocol: research background, research questions, studies searching strategy (search terms and resources), study selection criteria and procedures.

The second phase review conducting is a complicated phase, which contains five stages: identification of research, study selection, study quality assessment, data extraction & monitoring, and data synthesis. Review conducting can start once the review protocol is agreed.

The aim of Identification of research is to define research questions that are related to primary studies. A search strategy was generated and was followed, and the searching process must be transparent and replicable. After this stage, the initial primary studies are collected and all the search and process should be documented, for instance the search string and data sources.

Primary study selection is a multistage process. Typically, the study selection process could be consisted of several rounds (1, 2, …N), while in each round human experts screen different component of the target set of papers based on the inclusion/exclusion criteria (Kitchenham, 2007). In the traditional study selection process, the period could last quite long, even for years, which highly depends on the research problem and retrieved initial papers. A common study selection sequence is starting from title, keywords, to abstracts, to introduction and conclusion, and finally ends in the full-page reading. A common way to conduct this process is that more than one human expert get involved in the first several rounds, and solving the conflicts to make consensus through discussion. After the study selection process finished, papers are selected as the primary studies for extracting the results. Figure 2 shows process of election of primary studies (Adeva & Atxa, 2014). A critical element for the study selection process is the inclusion and exclusion criteria, which should be well defined in the planning phase. The principle of the criteria design is keeping objective and clear, as the purpose is to guarantee the objectiveness of the SLR results.

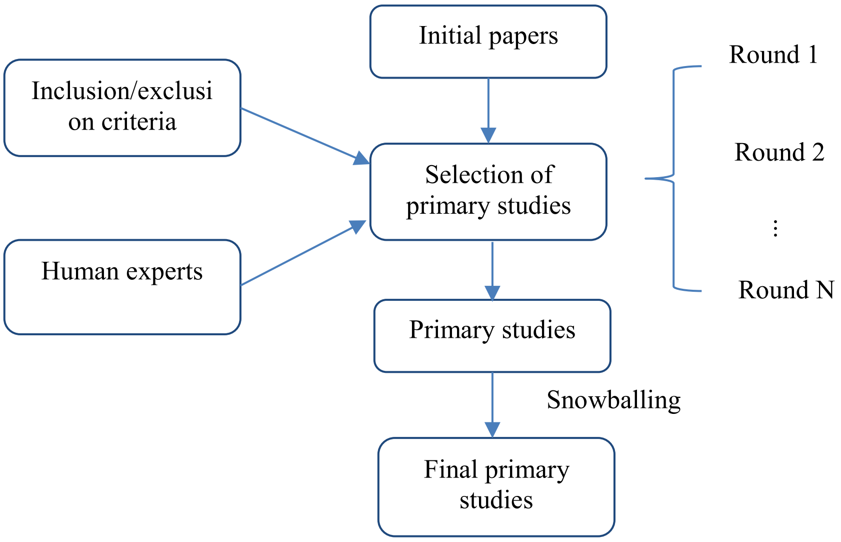


Figure 2. Study selection process.

Study quality assessment is generally considered to evaluate the "quality" of the selected primary studies, which is based on the inclusion/exclusion criteria. However, there is no agreed definition of the study quality, while in general the quality is suggested to be related to the extent that a study minimizes bias and maximizes the validity.

In stage data extraction and monitoring, the extraction forms are designed to record the gained information from the primary studies. Data synthesis summarizes the results from the extracted information. It is possible to present the results with a quantitative summary with statistic techniques, while descriptive synthesis is very common.

Review reporting is a single stage phase. It is important to effectively present the results of system reviews. The structure and contents for SLRs are suggested as: title, authorship, executive summary or structured abstract, background, review questions, review methods, included and excluded studies, results, discussion, conclusions. Contents like acknowledges, conflicts of interests and the related references and appendices can also be reported. Nevertheless, the suggested reporting form is not strict.

2.2 Issues related to traditional process

The most common criticisms of SLRs are that they cost plenty of time (Kitchenham & Brereton, 2013), while study selection phase takes a lot of the time when conducting the whole process (Marshall & Brereton, 2013). For example, a typical case is 2000-5000 papers are retrieved in the beginning of the study selection phase, and human experts would need to go through a three rounds' study selection phase with inclusion/exclusion criteria, which are predefined in the protocol; each round corresponding to different components of the study (title, abstract, keywords, introduction and full paper). The problem is getting even worse due to the growing number of scientific papers in Software Engineering field in the recently years. Most tools for supporting the study selection process are still at the early stages, and they haven't performed well in speeding up the phase or improving the systematic review results (Marshall & Brereton, 2013). So far, none of the existing tools or methods is widely used as a unified tool.

The time-consuming issue can be caused by various reasons. A very known cause is that large number of articles requires lots of time and effort to review one by one; and the scale of the initial collected articles determines the time for study selection process. Another cause is quite practically seen in the process: review repeating caused by immature SLR protocol. Different reviewers are usually equipped with various knowledge regarding a certain research topic. Protocol is designed and agreed in the planning phase, thus there is a high possibility that the protocol is not well designed, especially when the researchers are lack of enough knowledge. Therefore, as researchers gaining knowledge during the study selection process, inevitably the review protocol needs to be revised.

The other serious but typical issue of SLR is the biased results of primary studies selection. Inclusion/exclusion criteria are predefined and used by reviewers to make the primary studies including or excluding decision. For many reasons a reviewer can make a wrong decision: improper inclusion/exclusion criteria, subjective mistake, or reviewer's own knowledge. Researchers have been working on improving the SLR result reliability, while "pair review" is used widely in the SLRs that two reviewers study selection the same studies and solve the conflicts when they occur. But still the manual and heavy SLR study selection work makes the process very difficult and exhausting for reviewers.

2.3 The relationship between inclusion/exclusion criteria and improving study selection process

In the traditional method, study selection process is divided to several rounds of selection, and each round selection is attached with inclusion/exclusion criteria, which is targeting at different components of papers. Typically, selection criteria affect the results of included studies, therefore they influence the quality of the SLR results. However, the process of making selection criteria is varying among different human experts with various knowledge.

Inadequate protocol from the planning phase leads to biased studies selection results easily, which happens particularly often when human experts are new to a research field; they usually lack enough knowledge to include the related information in the criteria (Brereton, 2007). Inadequate inclusion/exclusion criteria can result in various troubles in the study selection process (Brereton, 2007). On one hand, the precision of including positive papers can be low when there are wrong including/excluding decisions made by the human experts based on criteria; on the other hand, inadequate protocol especially selection criteria can result in frequent criteria revision during the review, and the worst part is that the researchers would need to go back to the previous stages of the whole process (Brereton, 2007). A great amount of human resources would be wasted because of the work repeating.

Plenty of research work has been done regarding performing systematic literature review in SE domain. SLR protocol is essential in study selection and data extraction, but study selection criteria need to be revised to guarantee the quality of the selection results (Brereton, 2007). Piloting the research protocol are very necessary because the process of testing will find mistakes in the protocol (Brereton, 2007). Also, in the guidelines of performing SLR in SE domain, Kitchenham indicates that in the SLR protocol, criteria for studies selection process can affect the SLR results. The suggestion is testing the selection criteria on a small amount of primary studies, which would be helpful in practical (Kitchenham, 2007).

Figure 3 describes the traditional process of how the criteria are formulated and applied in the study selection process. Criteria are defined separately for each stage of the selection, while the three rounds selection is the mostly highly used method in the traditional study selection process.

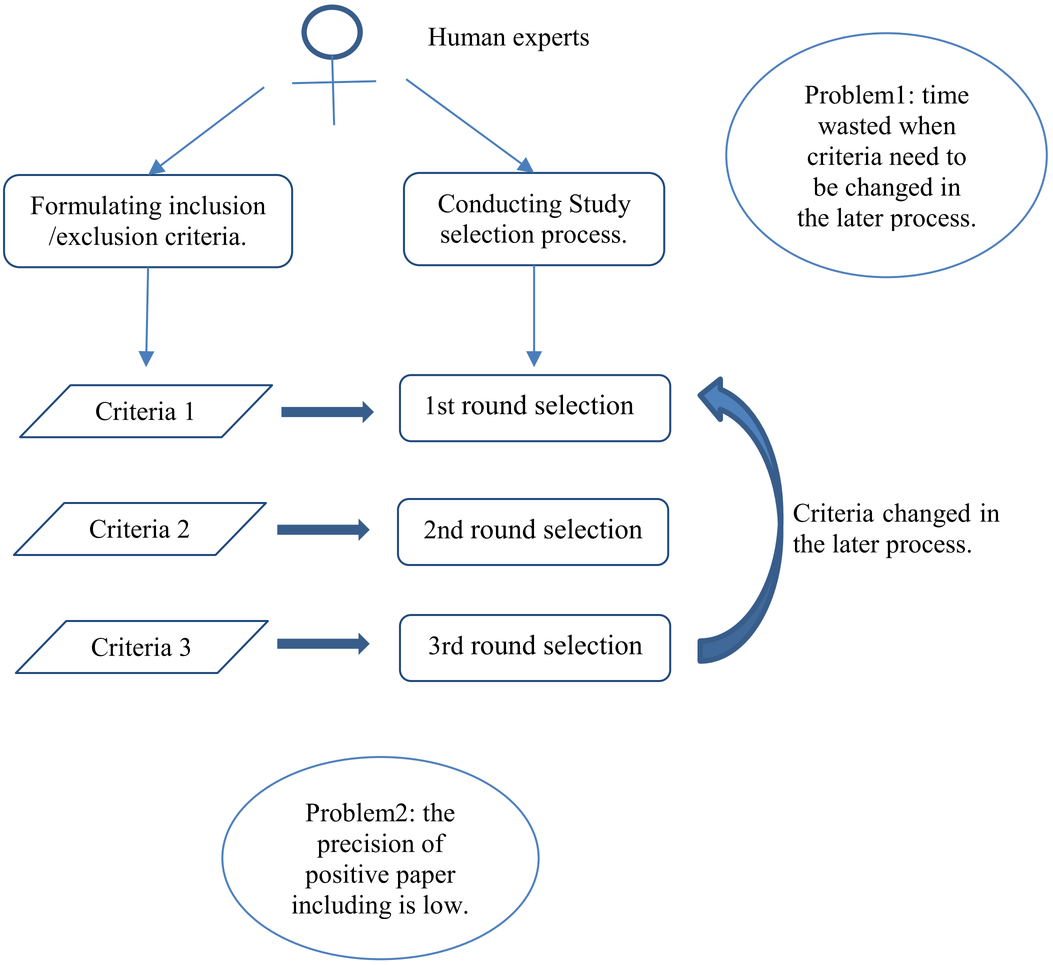

Human experts
Problem1: time wasted when criteria need to be changed in the later process.
Formulating inclusion /exclusion criteria.
Conducting Study selection process.
Criteria 1
1st round selection
Criteria 2
2nd round selection
Criteria 3
3rd round selection
Criteria changed in the later process.
Problem2: the precision of positive paper including is low.

Figure 3. Traditional selection criteria formulating, study selection process and potential problems.

In the figure 3, there are two potential problems in the traditional study selection process, which caused by the inadequate inclusion/exclusion criteria, though factor of human experts is inevitably affecting them as well. Considering the problem 1, criteria need to be changed in the later study selection case is a quite common phenomenon in the practical SLR process, which is mainly because the criteria is not adequate for conducting the study selection process. In the problem 2, limited inclusion/exclusion criteria provide the biased standard for human experts to make the including/excluding decision, which caused the reviewing papers are not correctly classified.

According to the above analysis, it is reasonable to refine the inclusion/exclusion criteria before the actual selection process starts, in order to solve the time wasting and low selecting precision problems.

3. Research questions and methods

This study is mainly aiming at supporting systematic literature review and mapping studying in Software Engineering field, by reducing the manual work and improving the results. Most of researchers are still conducting SLR in a traditional way according to Kitchenham's instructions (2007), while the process takes a lot of time especially in primary study selection phase.

3.1 Research problem

The main goal of this thesis is to study supporting the study selection of SLR in SE field. The study scope was restricted to the second phase of SLR and Software Engineering discipline to focus on solving the real difficulties of SLR in SE. The existing research indicates a very common way to support SLR is to semi-automate the processes which involves both manual work and developed tool running on machine. Besides, since the target review papers are all written in natural language, text mining is the highly adopted in semi-automating SLR. Therefore, to study how to support study selection process of SLR in SE field, two research questions are defined:

RQ1: What are the existing tools and methods in supporting SLR process?

RQ2: How can study selection phase of SLR in SE be improved?

The studying of RQ1 is the basis of RQ2. The objectives of RQ1 are to unveil the problems in SLR process and discover the existing achievements and potentials in support SLR process. Software Engineering discipline and study selection process are not restricted for the RQ1 because it is likely the evidences in other research fields (particularly in medicine discipline) and other phases in SLR can be referred for solving the RQ2, which is limited to SE field and study selection process.

As RQ1 is very wide and general, it is divided to two sub research questions so as to study the specific phenomena:

RQ1-1: What are the existing tools and methods in semi-automating SLR?

RQ1-2: What text mining related techniques and tools can be used in supporting SLR process?

This thesis used SLR as a research method for collecting the existing evidences regarding the two sub research questions. Studying RQ1-1 is the most important preparation for RQ2, and it is aiming at collecting the existing research results in supporting semi-automating SLR. The current research achievement in supporting semi-automating SLR process should be learned and summarized, particularly the advantages and disadvantages. RQ1-2 results are another important basis for RQ2. The aiming of RQ1-2 is to find out the tools and techniques that have been applied in supporting SLR process, and to study the potential techniques that could be applied in the future SLR process improvement.

RQ2 is aiming at find out how to reduce the manually work and improve the SLR performance. The knowledge from the above three research questions were analyzed and results were used for find out the solution to improve study selection of SLR in SE.

3.2 Research methodology

Design-science research is used as the research method in this thesis. DSR is known as developing technology-based solution for important business problems (Hevner, 2004). Typically, the motivation of DSR comes from two aspects: the study object must be interesting and real; also, the developed design science artifact should be relevant and useful (Hevner, 2004). Therefore, DSR is a suitable research method for this thesis, because the research problems and objective fulfill the motivation of applying DSR. The thesis aims to solve a practical problem in SLR process in SE domain.

To use design science paradigm, IT researchers must: identify important problems, informed by existing knowledge related to natural, behavioral and design science; then develop creative artifacts to address the problems (Hevner, 2004). Clear and reliable contributions of design artifact, design foundations or design methodologies should be provided by effective design science research (Hevner, 2004). Philip Offermann proposed an outline of a design science research process (2009), which combines quantitative and qualitative research and references well-known research methods. Figure 4 presented the adapted DSR process proposed in the guideline.

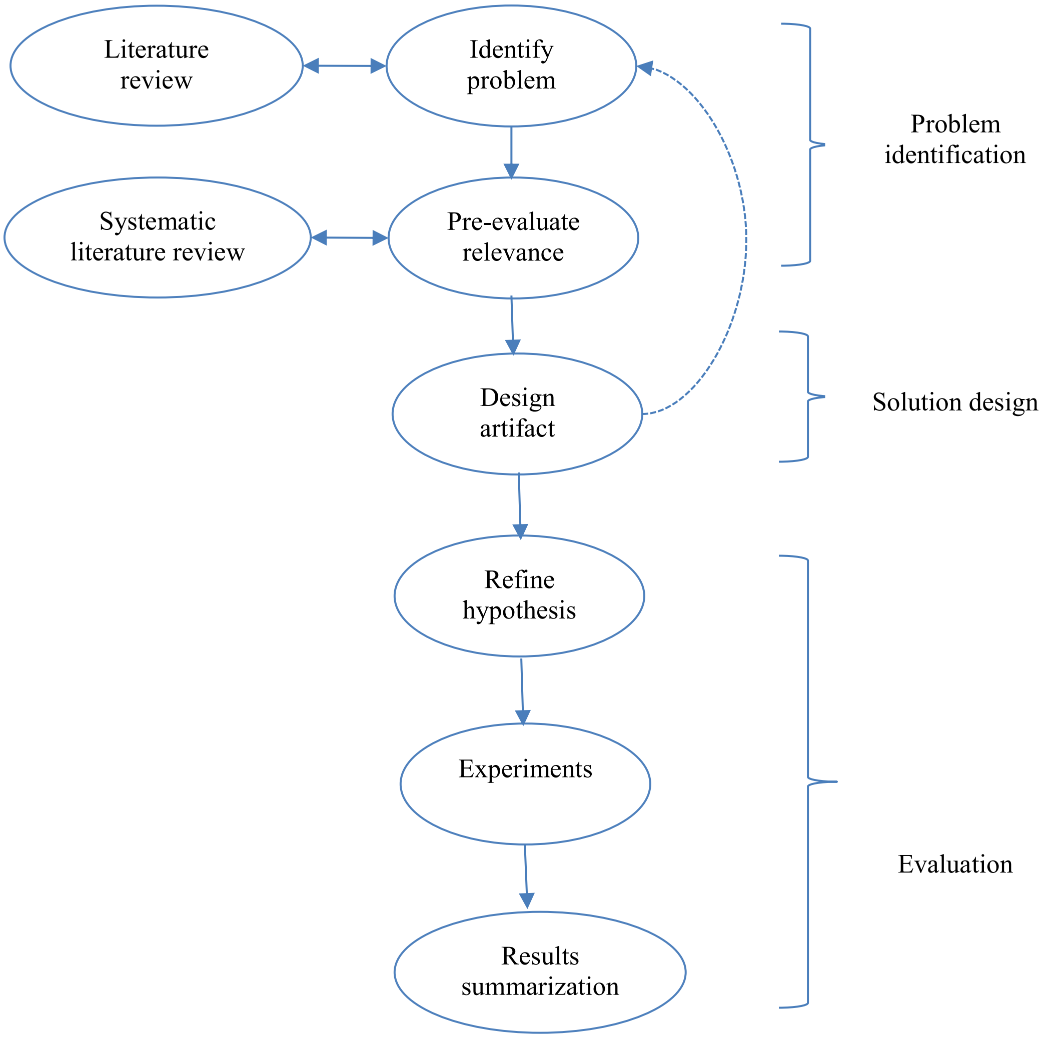


Figure 4. Adapted DSR process.

The DSR contains three parts: problem identification, solution design and evaluation, while each part can interact with others in the process. Each part of the process is divides to steps and the arrows indicate the transition relations (Offermann, 2009). In this thesis, this DSR process is used for planning and conducting the study. The steps of the process are adapted as follows:

Identify problem. Problem needs to be identified and refined during the current phase (Offermann, 2009). The initial literature research (Literature research-part I) of SLR process is only for identifying research problem. Since SLRs method introduced from Medical discipline to SE field, it became widely used but mostly manually done. There are very common criticisms about the heavy work and the whole process is very time consuming. The research problem is finalized as supporting study selection process of SLR in SE. Before come up with solution, three sub problems need to be addressed, which is aiming at find out the existing relevant evidence and potential techniques.

Pre-evaluate relevance. After the thesis's study problem specified, the extracted evidence from the initial literature review should be presented. A relevance pre-evaluate will conducted to estimate the possibilities for the study selection phase supporting. This part will be presented in section 4.4 as implications for improving study selection.

Systematic literature review. An individual SLR will be conducted in this step, which is aiming at answering RQ1, and discovering all the related evidences for “artifact design” for RQ2. The sub questions of RQ1 will be used in SLR planning; the extracted information from the data extraction phase should answer the above research questions. The SLR protocol is presented in chapter 3.3, and the results will be presented in chapter 4.

Design artifact. Designing artifact is a creative engineering process, while the existing solutions and state-of-art needs to be under consideration (Offermann, 2009). This thesis is looking for solutions for supporting study selection phase of SLR in SE, while this developing process is highly depending on the evidences found (issues, techniques, or possible solutions.).

Refine hypothesis. When a reasonable solution is found for supporting study selection phase, a hypothesis will be refined for the experiment evaluation. The hypothesis should be that the solution can improve the study selection phase when using SLR in SE, in terms of the efficiency and result accuracy.

Experiment. In the experiment, paper selection accuracy and positive papers’ recall will be calculated under the designed solution, and then be compared to the traditional SLR study selection process. The experiment is aiming at evaluating the proposed method for study selection process, thus it will be carried out between the traditional SLR study selection process and the proposed study selection procedure. The purpose for the first experiment is to verify the hypothesis that the new study selection process will show better results in terms of time and effectiveness. For the further validation, this thesis will carry out another experiment with more practical SLR. Regarding the research problem, selection precision and recall are recognized as the main target of optimizing study selection process, or even the whole systematic literature review procedure. Both the two measures can be measured with quantitative data.

Summarize results. The findings of existing SLR research work, together with solution developed for supporting study selection phase of SLR in SE will be summarized and presented.

3.3 Systematic literature review

A SLR was undertaken in this thesis as a very important research method. The purpose of this systematic review is to solve the first research questions of this thesis. Also, this SLR could help others researcher if they would continue to solve this study selection issue in the future.

Background and related literature review work

The issue of supporting systematic literature review process has been gaining more and more attention in recent years in SE field, due to the increasing popularity of SLR applied in preparing and carrying out research work. Under this circumstance of lacking a unified useful SLR tool, researchers are in a need of tool to support their SLR work, this also made to be a big requirement in SE field.

There are several literature reviews about supporting SLR process existing, however none of them have done a comprehensive review about supporting the study selection process. Thus, the SLR was conducted, while focusing on discovering the existing knowledge in automating study selection process. By analyze their feasibility, effectiveness and shortcoming; the thesis continued studie how to improve the SLR study selection phase.

Research questions and purposes

As defined in the research problem, RQ1 is related to study the existing evidence concerning issues, techniques or any research work in general. To study RQ1, this thesis undertook a SLR process based on Kitchenham's guidelines, as presented in the chapter 2.1.

RQ1: What are the existing tools and methods in supporting SLR process?

a. *RQ1-1: What are the existing tools and methods in semi-automating SLR?*

b. *RQ1-2: What text mining related techniques and tools can be used in supporting SLR process?*

RQ2: How to improve study selection phase of SLR in SE?

Research questions played an important role in discovering and reporting the possible knowledge for supporting our research work. Specifying the research questions is a critical step for the whole SLR process. As this thesis is focusing on improving the existing SLR process, in this SLR process the sub questions of RQ1: RQ1-1, RQ1-2 are specified as the research questions. The final SLR findings should address RQ1, which is also a very important basis for studying how to improve SLR process (RQ2).

SLR process

This SLR was adapted from the traditional SLR, as discussed in chapter 2. Figure 1 presents the traditional SLR process in SE.

Search string

The main keywords used for searching papers are "Text mining", "SLR" and "tool". Term alternatives are defined in table 1.

Table 1. Searching keywords.

Keyword	**Alternatives**
Text mining	Text analysis, text analytics, information extracting, natural language processing
SLR	SLRs, systematic literature review, systematic literature reviews, mapping study
Tool	tool AND automat*

Based on the searching terms, the searching string is formulated. The final string is as below:

Search string:

("systematic literature review" OR "systematic literature reviews" OR "mapping study")

AND

((“text mining" OR "text analysis" OR "text analytics" OR "information extracting" OR "natural language processing”) OR (tool AND automat))*

Literature resources

According to recommendation of the experienced researchers, combining the analysis of our research topic, four literature resources are defined for searching for literature.

a. ACM Digital Library

b. IEEE Xplore

c. Scopus

d. ScienceDirect

Search process and study selection

The primary study selection process is presented in figure 5:

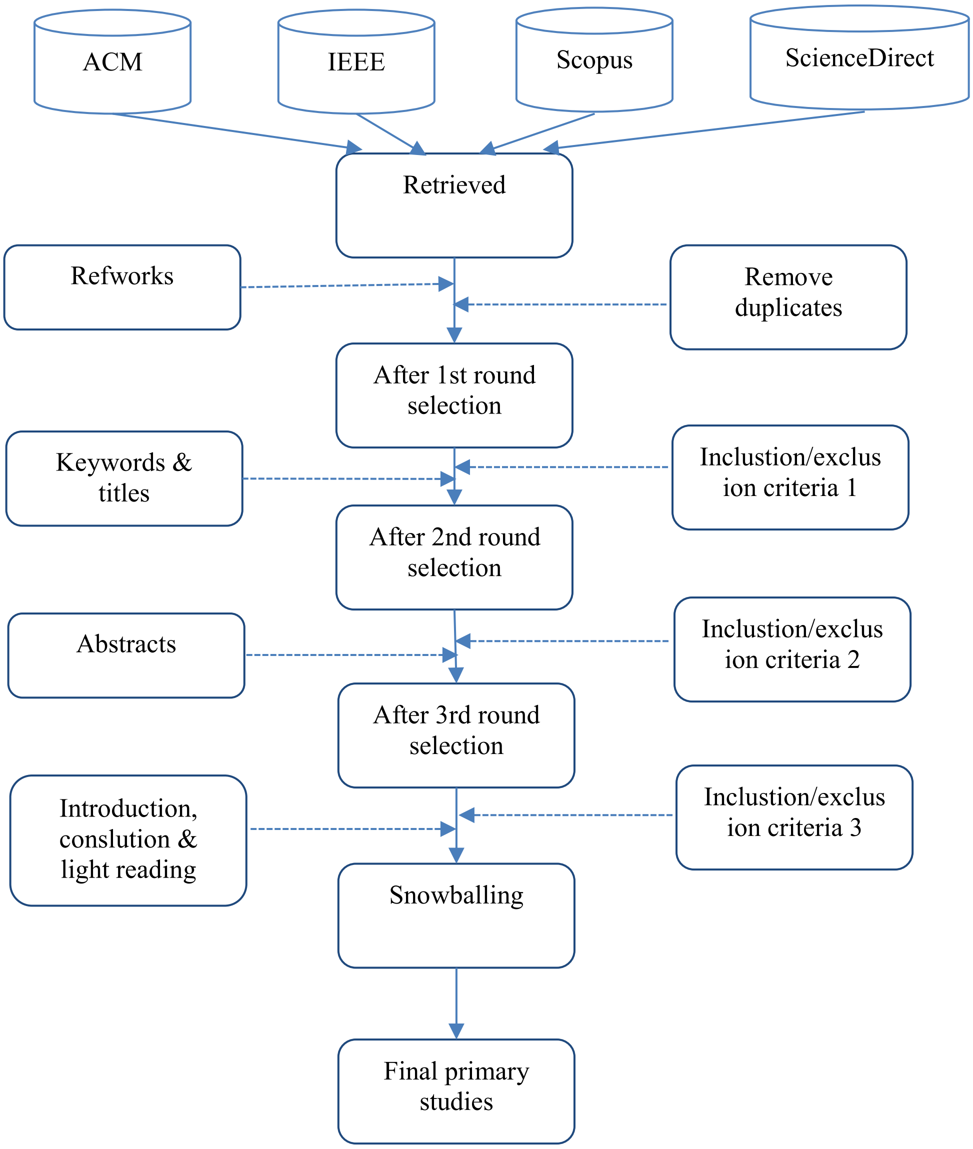

ACM
IEEE
Scopus
ScienceDirect
Retrieved
Refworks
Remove duplicates
After 1st round selection
Keywords & titles
Inclustion/exclusion criteria 1
After 2nd round selection
Abstracts
Inclustion/exclusion criteria 2
After 3rd round selection
Introduction, conslution & light reading
Inclustion/exclusion criteria 3
Snowballing
Final primary studies

Figure 5. Primary studies selection.

Snow balling is known as a first search strategy, may be good for replace the use of databases searches (Wohlin, 2014). Once obtained primary studies after 3nd round study selection, a backward snowballing was conducted to find more useful primary studies. Backward snowballing refers to use the references of a paper and select the relevant papers also as primary studies (Wohlin, 2014). The process can be done with four steps: (1) look at title in reference list; (2) look at reference's place; (3) look at the abstract of referenced paper; (4) read the full reference paper (Wohlin, 2014).

Four types of the selected articles are identified as the primary studies: (1) review study: reviewing study of supporting SLR, (2) tool: the articles presented or discussed using tools to support study selection process, (3) theory: study presented specific theory or concept, which can be, applied to support reviewing, (4) other aspects.

Data synthesis

The criteria used for performing the data synthesis process are showing in following table 2. Each of the articles were scored, while “Yes” corresponds to point 1, “Partly” to point 0.5 and “No” with point 0.

Table 2. Article quality assessment criteria.

No	Item	Answer
1	Is it clear what techniques have been used in the article?	Yes/No/Partial
2	Have the techniques or tools been validated?	Yes/No/Partial
3	Is it clear which SLR phase the techniques or tools could be applied to?	Yes/No/Partial
4	Is it clear how the tools or techniques applied?	Yes/No/Partial

Selecting procedure and inclusion/exclusion criteria

The study selection procedure is a complex process that requires plenty of paper study selection and decision making work. After pre-selection (Refworks will be used), human experts will go through the searched papers three rounds while study selection different components of the papers. In round 1, titles and keywords will be screened for each paper, while in round 2, only abstract will be read. In round 3, human experts will read introduction, conclusion and more slightly reading when it is needed. The basic study selection criteria are specified as: (1) after the pre-selection through the RefWork’s function, the primary study selection will be taken place in three rounds; (2) when there’re duplicate papers, consider the first one; (3) when there’re multiple papers for the same study, consider the latest one. Based on the research questions, inclusion/exclusion criteria are predefined as follows:

Inclusion criteria (IC) and Exclusion criteria (EC):

- IC0: If the article is a: Conference paper/ journal paper/book chapter/technical report/Ph.D. thesis.
- EC0: If the article is a: Duplicate/not written in English/opinion.paper/interview/summary/extended abstract/master thesis.
- IC1: Fulfills inclusion criteria from the previous rounds AND Is related to SLR/systematic review/mapping study, or is related to SLR/systematic review/mapping study tool/method, or is related to techniques such as text mining/natural language processing/information extraction/machine learning/pattern recognition, or is related to any other possible techniques/application related to the above mentioned terms of techniques, e.g. "neural network".) AND fulfills the inclusion criteria from the previous rounds.
- EC1: Fulfill the exclusion criteria from the previous rounds.
- IC2: Fulfills inclusion criteria from the previous rounds AND (Is a SLR or mapping study that used tools or techniques for supporting automating of the process OR is a

SLR or mapping study of supporting automatic SLR/mapping study OR is a study about techniques or methods that can be used for supporting automatic SLR or mapping study process), OR is a study that discussed how to support text mining.

- EC2: Fulfill the exclusion criteria from the previous rounds.
- IC3: Fulfills inclusion criteria from the previous rounds AND (Is a study about how to support the automatic selection/semi-automatic selection of primary study Or including the methods/techniques that supporting the automatic selection/semi-automatic selection of primary studies).
- EC3: Fulfill the exclusion criteria from the previous rounds.

3.4 Evaluating measures

The experiments will be evaluated with experiments while comparing the results to the traditional method. The defined evaluation process and measures will be used for testing the hypothesis, which refers to the proposed framework improves the study selection results.

Precision (p) and recall (q) are two common measures for assessing how successful a text categorizer is, they are selected for evaluating the study selection process. Precision indicates the probability that a document assigned to a certain category by the classifier that belongs to that category. On the contrary, recall estimates the probability that a document belongs to a certain category got correctly assigned during the categorization process.

When interpreting the study selection results, recall is the most important measure which indicates how successfully the positive papers are included in the selected papers. The higher recall, the more valuable evidence is collected for the SLR. Precision will be used for measuring the possibility that a study included by a reviewer turn out to be positive. In the study selection process, the higher selecting precision, the more irrelevant papers (negative papers) can be excluded in the early selection rounds, which also means the faster reviewers can complete positive papers inclusion. Therefore, recall and precision are useful measures for evaluating the study selection results. F-measure is a simple measure that combines the recall and precision, also will be measured in the later experiments.

Precision (p) and recall (q) are defined by

$$p = \frac{TP_i}{TP_i + FP_i}$$

$$q = \frac{TP_i}{TP_i + FN_i}$$

where TP_i indicates the number of true positives or how many documents were correctly classified under category C_i. Similarly, FP_i indicates the number of false positives and FN_i corresponds to false negatives.

Precision and recall can be generally combined into a single measure called F-measure: F_β, with $\beta \epsilon (0,1)$. The parameter b, which is used to find the appropriate balance between the importance of p and q, is expressed by

$$F_\beta = \frac{(\beta^2 + 1)pq}{\beta^2 p + q}$$

When $\beta = 1$, the function is known as F1-score measure. It is a well-known measure of effectiveness and it is common used, which combine the contribution of precision and recall. The function is defined as:

$$F_1 = \frac{2pq}{p + q} = \frac{2TP}{2TP + FP + FN}$$

Apart from evaluating the results improvements, another important measure index is the time consumed in the study selection process. During the process of the SLR experiments, every selecting round of the study selection process will be record of the time consumed. However, recording consumed time during SLR is a very demanding and difficult procedure. Indicator “time” will not be used in the experimental evaluation in this thesis.

4. Previous studies

To discover the evidence about supporting study selection phase of SLR in SE, a SLR was conducted according to Kitchenham's guidelines. The SLR protocol is defined and validated before carried out. Each included primary study is summarized and synthesized, which would be interpreted to determining their applicability and supporting the later research. Eventually 31 articles are selected as the primary studies. This chapter will present the results from the analysis of primary studies.

4.1 Overview of primary studies

Primary studies were selected from the initial 997 papers that retrieved from four databases: ACM, IEEE, Scopus and ScienceDirect, with the pre-defined search strings. Scopus contributes the largest number of studies (503), which even exceed half of the total studies. ScienceDirect also returned a large amount of studies (399), while ACM and IEEE contributed only 85 and 10 papers respectively. After pre-section, selection round 1, round 2 and round 3, only 20 (2.0%) papers were selected as the primary studies. Table 3 shows the paper distribution during the selection process.

Table 3. Paper distribution during the selection process.

Database	**Initial**	**Pre-selection**	**1st**	**2^{nd}**	**3^{rd}**
ACM Digital Library	85	85	20	5	2
IEEE Xlpore	10	10	8	4	1
Scopus	503	501	202	50	13
ScienceDirect	399	397	129	17	4
Total	997	993	359	76	20
Percentage	100.0%	99.6%	36.0%	7.6%	2.0%

However, the 20 selected primary studies were not the final primaries studies. While reviewing the full papers, there are more studies found fulfills the inclusion criteria of SLR selection. During this snow balling process, 11 studies are found (See figure 6); therefore 31 papers are selected as the final primary studies.

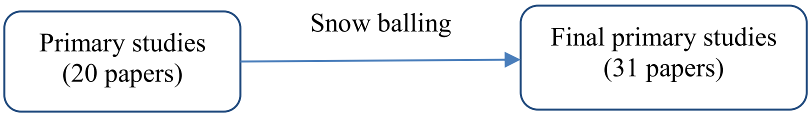


Figure 6. Snow balling for selected primary studies.

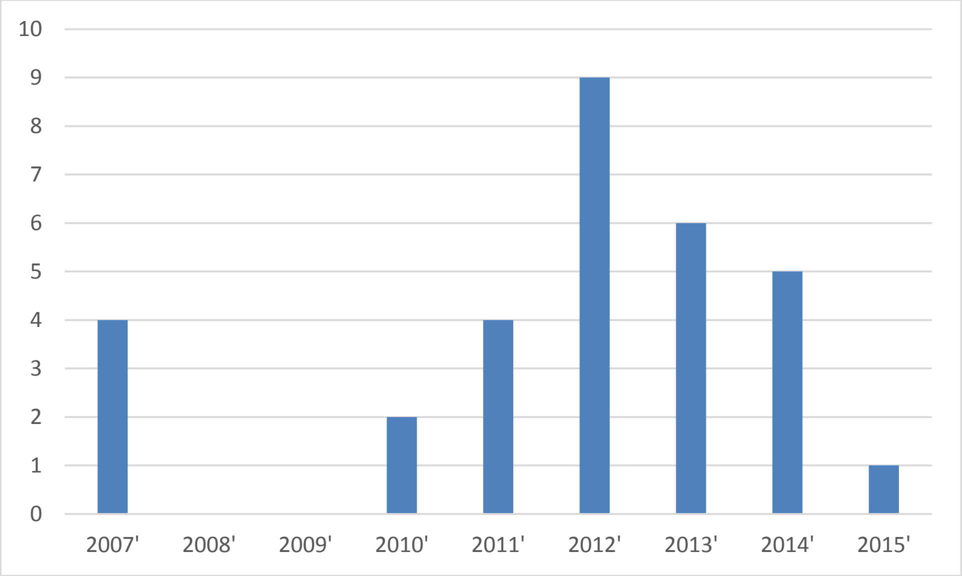


Figure 7. Primary study publication year distribution.

It has been found that all final primary studies were published during the past 8 years and the number of publications shows an increasing trend in the recent years. Figure 7 shows the primary studies publication year distribution. As can be seen from the figure 7, year 2012, 2013 and 2014 are the years produced most of the papers, in other words, recently years researching on supporting SLR process has gained more attentions than before. Year 2015 is an exception because the initial papers are collected in the beginning of that year. Besides, there were not any publication in year 2008 and 2009. A possible explanation could be that in 2007, Kitchenham published a paper about the SLR process and after that it took other researchers a couple of years to conduct more research for the afterwards work.

4.1.1 Primary Quality Analysis

The quality of the primary studies was analyzed according to the article quality assessment criteria presented in Table 3. By collecting the answers for each question, the total points were calculated for each article. The higher the score, the higher quality of the article. Quality score of the primary study can be found in Appendix B.

The quality score results indicate that most of the articles have answered the quality criteria questions quite well, especially in Q1. Some articles are not clear about the validation of using the tools and techniques, which is not apparently shown in the articles. Besides, Q2 is not clear in some articles as well, while it is mostly because of these primary studies selected as potential papers and originally, they were not used for supporting study selection phase in SLR.

4.1.2 Primary studies analysis

All the primary studies are analyzed based on the research questions. Each primary study involves one or multiple review phases, while the most common phase is the review conducting phase, which is also the most time consuming phase. A few studies even cover all the three phases. Four study types are defined to categorize the studies: tools, review study, theory/concept, and others. A study marked with study type tools should present a tool or tools that relate to support SLR study selection process.

"Review study" type studies are SLRs or mapping studies of how to support SLR. "Theory/concept" studies discuss the theories that relates to support SLR review conducting. If a paper is not one of the above three types study, it is categorized to "others" type. Table 4 presents the overall analysis of the selected studies.

Table 4. Overall analysis of the selected studies

ID	Review phases	Study type	Technique	Details
P01	Planning, conducting, reporting	Theory or concept	Text mining	Text mining techniques are discussed applying for improving different stages of the SLR process (searching, study selection, synthesis). Concluded semi-automated techniques can be useful for performing time consuming tasks of SLRs.
P02	Conducting	Theory or concept	Text mining	Described application of four text mining technologies in supporting review conducting phase of SLR: automatic term recognition, document clustering, classification and summarization.
P03	Conducting	Tool	Text mining and information visualization techniques	Approach SM-VTM (systematic mapping based on visual text mining) is proposed for supporting categorization and classification stages in SM, aiming at reducing time and effort. The evaluation results indicate VTM seems to be very relevant in the context of SM. VTM tool PEX is used.
P04	-	Review study	-	A review study of supporting or automating the processes of SLR or each of the tasks of the SLR
P05	Conducting	Other	Other	Introduced automatic coding system that code medical report written with natural language. Coded reported can used for automatic classification. The TREEBOOST.MH learning algorithm is adapted for classifying.
P06	Conducting	Other	Text mining	Using semantic atoms and topology for automated clinical coding (ACC). ACC transforms narrative text in clinical records into a structured form.
P07	Planning, conducting, reporting	Tool	Text mining and information visualization techniques	VTM techniques are used in study selection stage, which can speed up the entire process. VTM tool Revis is used.
P08	Planning, conducting, reporting	Tool	Other	Tool SLuRp is introduced for performing SLR, especially large complex SLR. SLuRp supports management of SLR.
P09	-	Review study	-	Review study of using visual data mining in supporting SLR.

P10	Conducting,	Theory or concept	Text mining and information visualization techniques	A novel approach SLR-VTM is proposed to support study selection.
P11	Planning, conducting, reporting	Theory or concept	Text mining and information visualization techniques	Applying VTM techniques in study selection stage. Proved VTM techniques are valuable in the selection o primary studies.
P12	Planning, conducting, reporting	Tool	Text mining	Presented tool SLR-Tool, a free-available tool for conducting SLR.
P13	Conducting	Theory or concept	Text mining	Uses automatic text classification to automate study selection process of SLR in medicine. Results are generally positive.
P14	-	Review study	-	A review study of technologies that used to automate SLR processes. Results show most of literature authors use machine learning classifiers to automate study selection process.
P15	-	Review study	-	A review study of SLR research in SE.
P16	-	Review study	-	A mapping study of tools to support SLR in SE.
P17	-	Review study	-	A feature study of tools to support SLR in SE
P18	Conducting	Other	Text mining and information visualization techniques	A knowledge browsing environment is developed to support literature analysis. Visualization capabilities of formal concept analysis are used.
P19	-	Review study	-	A review study of SLR tools
P20	-	Review study	-	A systematic review study of using text mining for study selection process.
P21	Conducting	Theory or concept	Text mining	Semi-automate the study selection process to reduce manual work and subjectivity bias. Approach “Linked Data” in proposed.
P22	Conducting, reporting	Theory or concept	Information visualization techniques	Analyzed the use of graphs to present the results of SLR in SE.
P23	Conducting, reporting	Theory or concept	Text mining and information visualization techniques	Introduced VTM in SLR and showed VTM can make the SLR process more effective.
P24	Conducting, reporting	Tool	Text mining and Information visualization techniques	Presented SLR tool StArt (State of the Art through Systematic Review). Information visualization and text mining are used.

P25	Conducting, reporting	Tool	Information visualization techniques	Evaluation of SLR tool StArt with GQM paradigm and TAM model.
P26	Conducting	Theory or concept	Other	
P27	Conducting, reporting	Theory or concept	Text mining and Information visualization techniques	Uses visualization and clustering algorithms to explore similarities and differences among empirical studies.
P28	Planning, conducting, reporting	Theory or concept	Text mining	Discussing possibilities of automatically extraction information from empirical SE literature. It is argued that the use of information extraction tools can support SLRs.
P29	Planning, conducting, reporting	Theory or concept	Text mining	Discussing automatic results identification in SE papers. Analyzed of the main methods for sentence classification in scientific papers
P30	Planning, conducting, reporting	Tool	Other	Supporting SLR process with proposed ontology SLRONT.
P31	Planning	Tool	Other	Presented a federated search tool to provide an automated integrated search mechanism in SE database.

The synthesized data from all primary studies show that all the review phases have automation support, while review conducting is the most frequently discussed phase, which is also the most time consuming step. Comparing the other two phases, review planning has less automation support. Besides, various techniques are presented and (or) applied in these studies; text mining and visual techniques are the most common used ones. The "Detail" of each study is presented regarding how this paper related to answer research questions: *What is the existing research work in semi-automating SLR (RQ1-1) and What tools and text mining related techniques can be used in supporting SLR process (RQ1-2).*

4.1.3 Extracted results

To have a better overview of primary studies, all the studies were summarized as four types: review study, tool, theory/concept, and other types. Figure 8 presented an overview of the primary studies distribution in the four different types. The category of review study refers to the study review of supporting SLR, or mapping study, feature study, etc. Those review studies do not introduce the new method but presented the existing methods or tools or other materials about supporting the SLR process.

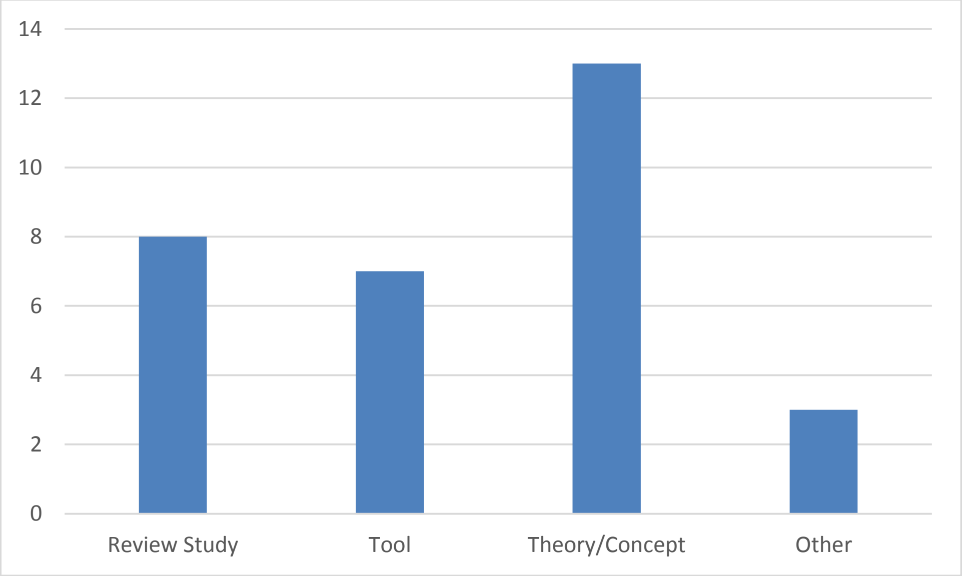


Figure 8. Primary study views of various study types distribution.

As can be seen from the figure 8, tools and theory or concept stand for most of the primary studies with the most numbers of selected papers. Theory/concept represents the largest category group (10 papers), followed by tool (10 papers) and review study (8 papers). The rest of studies (3 papers) do not belong to any of the categories above, which are defined as others.

By analyzing those primary studies, it has been found out that the study selection process has either been supported by tools or simply techniques, and those tools and techniques are mostly related to text mining techniques and visual techniques. However, the study selection process is still not well supported for many reasons in various primary studies. Some automating tool results in a problem that people too much trust in the machine so that they could reduce their tedious huge work, which is good for a reliable SLR. Some other tools or techniques has an issue that they require the users understanding the machine learning knowledge, which made it too hard to apply the developed tool.

Through the analyzing process, the primary studies were found out to use various techniques for supporting the study selection phase. Only non-review studies are analyzed in terms of the techniques, as the review studies not really provide a solution. This means 87.5% of the articles are analyzed regarding the techniques, according to the previous primary studies classification distribution.

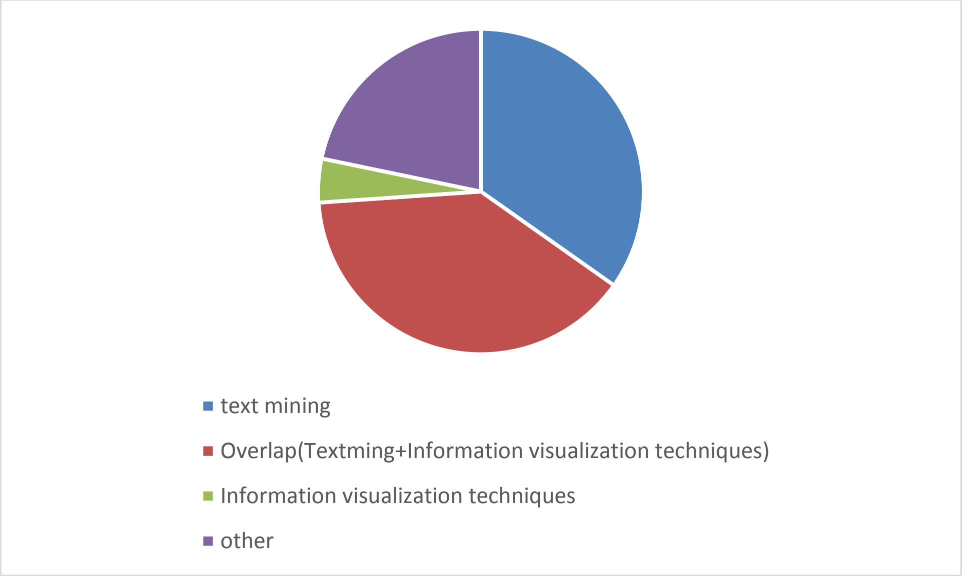


Figure 9. Technique used distribution.

The found techniques could be summarized to a few groups as presented in figure 9. Text mining and visual techniques are the most commonly used methods, which stand for 56% and 32% respectively (including the overlap). Combining text mining and visual techniques is also a popular approach presented in some studies, while the percentages reached 16% (overlap part in the figure 9). Other techniques such as data mining, pattern recognition is used as well.

4.2 Existing research work in supporting semi-automatic SLR

Based on the synthesized data of primary studies, there is plenty of research work has been done in supporting semi-automatic SLR. The studies can be categorized to four types: theory or concept, tools, review study and others.

Theory/concept

Text mining is the most frequently discussed techniques that can be applied in supporting SLR process. Text mining techniques can be applied for improving different stages of the SLR process (searching, study selection, synthesis), and it has been discussed that semi-automated techniques can be useful for performing time consuming tasks of SLRs (Ananiadou & Procter, 2007). Various text mining technologies in supporting review conducting phase of SLR, such as automatic term recognition, document clustering, classification and summarization (Thomas, 2011). In Medicine field, text mining has been also used in automating study selection process in SLR, and the evaluation results turn to be positive (Adeva & Atxa, 2014). Related to text mining, approach "Linked Data" in proposed to semi-automate the study selection process to reduce manual work and subjectivity bias (Tomassetti, 2011).

Visual techniques are also common considered to be applied in SLR process (Felizardo, 2012; Felizardo & Maldonado, 2013; Felizardo, 2011; Malheiros, 2007; Cruzes, 2007). It has been analyzed the use of graphs to present the results of SLR in SE (Felizardo, 2011). However, the most common way to use visual techniques is to combine it with text mining techniques (Felizardo, 2012, Felizardo & Maldonado, 2013; Malheiros, 2007; Cruzes, 2007). Visual text mining (VTM) has been applied in conducting SLR process, and there are existing results indicate that VTM can make the SLR process more effective (Felizardo, 2012, Felizardo & Maldonado, 2013, Malheiros, 2007; Cruzes, 2007).

Some other possibilities of supporting SLR have also been discussed. It is possible to conduct automatically information extraction from empirical SE literature, and the use of information extraction tools can support SLRs (Cruzes, 2007). Besides, automatic results identification in SE papers is also possible, for instance use sentence classification in scientific papers (Torres, 2012).

Tools

There is a list of SLR tools have been developed for supporting various phase of SLR, while some of them do support the whole review conducting process.

SLuRp refers to Systematic Literature unified Review Program. The main functionality of SLuPp is to support the management of the SLR process by using the database that is open source web-enabled (Bowes, 2012; Marshall & Brereton, 2013). This tool achieved saving time of SLR by easing the whole SLR process (Bowes, 2012; Marshall & Brereton, 2013). However, this tool is not supporting study selection process much, while still quite lots of time requires for study selection articles (Bowes, 2012; Marshall & Brereton, 2013).

StArt, also known as the State of the Art through systematic review, in fact supports each stage of the SLR process (Fernández-Sáez, 2010; Hernandes, 2012). In terms of supporting the selecting phase, StArt provides a better view for study selection, as the human experts are able to make including or excluding decision (Fernández-Sáez, 2010; Hernandes, 2012). Users can upload the data to StArt and input the protocol's inclusion and exclusion criteria (Fernández-Sáez, 2010; Hernandes, 2012).

The most obvious advantage of SLR-TOOL is its freeness to use, while it supports each phase of SLR process as well (Fernández-Sáez, 2010). To some extent it is similar to StArt that it is also more like a management tool, which eases the study selection process for human experts (Fernández-Sáez, 2010).

Different from the precious tools, Revis is tool based on text mining, which supports both the study selection and study reviews (Felizardo, 2012). Revis supports the analysis of the documents and metadata based on the content, and there are four strategies can be used for exploring the content map: KNN edges connection, coordination, clusters and expression occurrence (Felizardo, 2012). These text mining related features help it to be easier and faster to make including/excluding decision (Felizardo, 2012).

A federated search tool can be used for automated integrated search mechanism in SE database (Kitchenham, 2012). It bridges the gap between the spread of databases in SE and integrated search required by SLR; A partially-automated tool is developed based on the proposed model (Kitchenham, 2012). This tool supports the semi-automating the review planning process.

SLRONT was developed to support the key activities of SLR (Sun, 2012). It has been proved that using ontology can support SLR effectively and efficiently (Sun, 2012).

To conclude, most of the above tools are SLR management tools, which support the study selection phase by easing the study selection activity. Some techniques used among them as well, such as the open source database, text mining. Despite of the existence of these tools, there are still quite many researchers continue using the traditional SLR process, or some developed their own small tool to assist the process. In a word, to some extent some developed tools improved the efficiency, but there is no unified mature tool available for supporting the study selection.

Review studies

8 review studies that related to supporting SLR process are included in the primary studies, while they were targeting different research problems.

A review study of SLR research in SE is conducted to study the systematic review automation technologies (Tsafnat, 2014). Another similar view study of "SLR automation: a SLR review" was also aiming at surveying the literature about the technologies can used for automating SLR processes; this review found out lots of research was done to automate study selection process and machine learning classifiers are frequently applied (Hamad & Saim, 2014). A systematic review of SLR process research in SE was done to improve the SLR process (Kitchenham & Brereton, 2013).

SLR tools are studied in several review studies (Marshall & Brereton, 2013; Marshall & Brereton, 2014; Zubidy & Carver, 2014). The summarized tools from these studies have been discussed in this chapter. In addition, there is a review study about using visual data mining in supporting SLR (Felizardo, 2012); and another review study about using text mining in study selection process (O'Mara-Eves, 2015).

Other

Some other relative research work was also included in the primary studies, which might give implications for finding out solutions to support semi-automating SLR. Automatic coding system is for processing code medical report written with natural language. Coded reported can used for automatic classification. The TREEBOOST.MH learning algorithm is adapted for classifying (Baccianella, 2013). Automated clinical coding (ACC) transforms narrative text in clinical records into a structured form (Barrett, 2012). Visualization capabilities of formal concept analysis can be used in knowledge browsing environment, in order to support literature analysis (Poelmans & Kuznetsov, 2013).

4.3 Tools and techniques applied in supporting study selection phase

There are several SLR tools and techniques have been developed or adopted for supporting SLR. The most common discussed tools are: SLuRp, SLR-TOOL, Levis, StArt, while they have drawn quite lots of attention from the previous researchers (Marshall & Brereton, 2013; Felizardo, 2012). Techniques applied in the SLR primary studies were classified to three groups: visual techniques, text mining and other techniques. According to the final primary studies, visual techniques and text mining are the most common used techniques, while in some papers it could be found that two techniques are combined and performed well in generating the study selection results. In general, text mining is defined as the process of extracting knowledge and structure from unstructured data (Ananiadou & McNaught, 2006; Hearst, 1999).

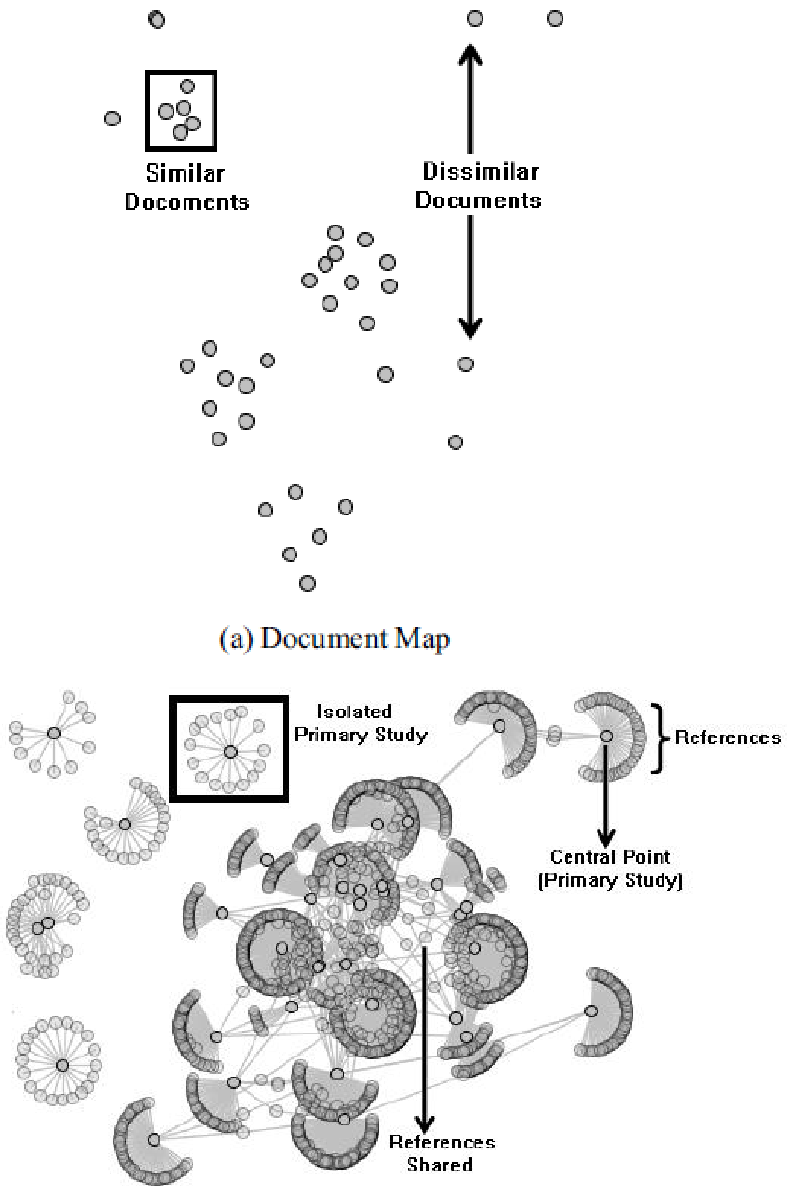
4.3.1 Tools

As presented in chapter 4.2, most of the existing tools are SLR management tools that ease activities of SLR. Many of them can be used for supporting the study selection process, while they are Revis, SluRp, StArt, SLRONT.

4.3.2 Information visualization techniques

Within the range of systematic literature review, typically information visualization refers to use visual text to present the information to improve the results. A visual text refers information that contains the imagery elements, which helps in emphasizing that critical information or attracts people (Felizardo, 2011).

Therefore, it is not hard to link the visual techniques to the selection process of SLR, due to its human manual study selection property. In the articles that used visual techniques, mostly the researchers present the key information or potential useful information with visual techniques, which facilitates reviewers make the including/excluding decision faster and easier, with a better view (Felizardo, 2012; Marshall & Brereton, 2013; Hernandes, 2012).



(a) Document Map

(c) Citation Network

Figure 10. Examples of information visualization techniques application (Felizardo, 2011).

In the figure 10, visual techniques are used with other data mining and some other techniques. The potential relationship between different articles are discovered with mining techniques and then presented with visual techniques. Similar documents would have quite high possibility to be included or excluded together. The citation network also gives a good view for people to make the selecting decision.

It is apparently that visual techniques have merits of improving the study selection results by providing human experts a better view. However, it also got a critical issue. It is difficult to start with the tool or techniques, particularly for people lack experience or related knowledge. The visual graphs are too complicated, and the developed tools like USR-VTM are difficult to use.

4.3.3 Text mining related techniques

Text mining as a type of data mining technologies is one of the most widely used techniques in developing SLR tools (Thomas, 2011; Cruzes, 2007). It includes various technologies such as document clustering, classification, summarization and automatic term recognition (Thomas, 2011). The main objective of text mining is to extract information from unstructured text, and then send useful knowledge to people with a concise form (Hearst, 1999; Ananiadou & McNaught, 2006; Thomas, 2011). The three major activities of text mining are information retrieval, information extraction and data mining, while these activities are mapped to some SLR's processes. There is a quite good possibility that the knowledge discovery with text mining supports SLR study selection process.

There are quite a set of articles that have used text mining in their research, while documents classification is widely used particularly (Thomas, 2011). Some research even almost entirely implemented automatic study selection, though it is used in the Medical field (Adeva, 2014).

Existing achievements indicate text mining is very useful technique applied in developing SLR tools or for supporting. However, the issue of text mining is that automating the review process can result in people too much relying on it, yet it should not be ignored that there is a certain percentage of biased decision made by the machine, which means they are not trustable completely. Another issue is related to the users' prior knowledge regarding the text mining technologies and algorithms. The issues of text mining are important reasons for the immature of SLR tool. It requires too much techniques knowledge, mostly related to machine learning knowledge and too difficult to use.

Besides, text mining is quite often combined with the visual technology because visual technology performs well in presenting the mining results of the unstructured text. **In the later chapters, text mining is used a main technique for supporting the study selection of SLR in SE.**

4.3.4 Other Techniques

There are several other techniques have been applied or discussed to support reviewing papers, or have the potential to be applied (Tsafnat, 2014). Pattern recognition, Natural Language Generation (NLG) technology, coding, graphs using are all discovered from the systematic literature review (Tsafnat, 2014).

4.4 Implications for improving study selection

According to the previous discussion, both text mining technology and information visualization techniques are very useful, equipped with high potential to be applied in SLR tool developments. However, the findings also indicate that they are confronting challenges so far applied in the existing tools and methods. Applying these two technologies or some other technologies in the regular study selection is still the trend, while there are quite many researchers did achieve relatively good results. However, these are not enough under the fact that there is no mature tool or highly adopted tool. Is there any other way to support the study selection process?

As analyzed in challenges of traditional SLR, in practical the time consumed in the study selection process can be divided to two parts: the regular process time consuming and extra wasted time consuming. Almost all the primary studies collected in the SLR section shows the existing results were all targeting in saving time of the regular time consuming. For the second problem that time consuming caused by the immature designed protocol from the planning phase, the ideal solution is to make a 100% perfect protocol, which is impossible in the practical case. Regarding the second issue in traditional study selection process, the accuracy of including the positive papers is also affected by the quality of the inclusion/exclusion criteria; the better quality of the selection criteria, the better primary study results that human experts can produce by the end of the study selection process.

Based on the above analysis, a new idea of optimizing the protocol is proposed; more precisely refers to the inclusion/exclusion criteria. The next chapter will propose a framework of optimizing the decision-making criteria.

5. An iterative framework proposed to optimize inclusion/exclusion criteria.

In the previous SLR results, it has been indicated that the inadequate protocol is a typical problem in SLR process. The inclusion/exclusion criteria affect the study selection quality, which can lead to biased selection results. An iterative framework is proposed in this chapter, aiming at formulating good criteria to make it reliable for making the selection decision. Text mining is selected in the framework since it is known as an effective method in supporting SLR, according the SLR results.

5.1 Framework design

As discussed in chapter 4, in theory refining selection criteria can improve a SLR protocol, which can also improve the study selection results regarding the efficacy and accuracy. An iterative process framework is proposed based on the idea of refining inclusion/exclusion criteria for the later primary study selection. As shown in the figure 11, the framework is an iterative process that contains three core sub-processes. Core sub-processed are filled with bold text in the figure 11; knowledge mining is the only process that supposed to be done automatically by machine and the human experts should conduct the other two processes of knowledge evaluation and optimizing criteria.

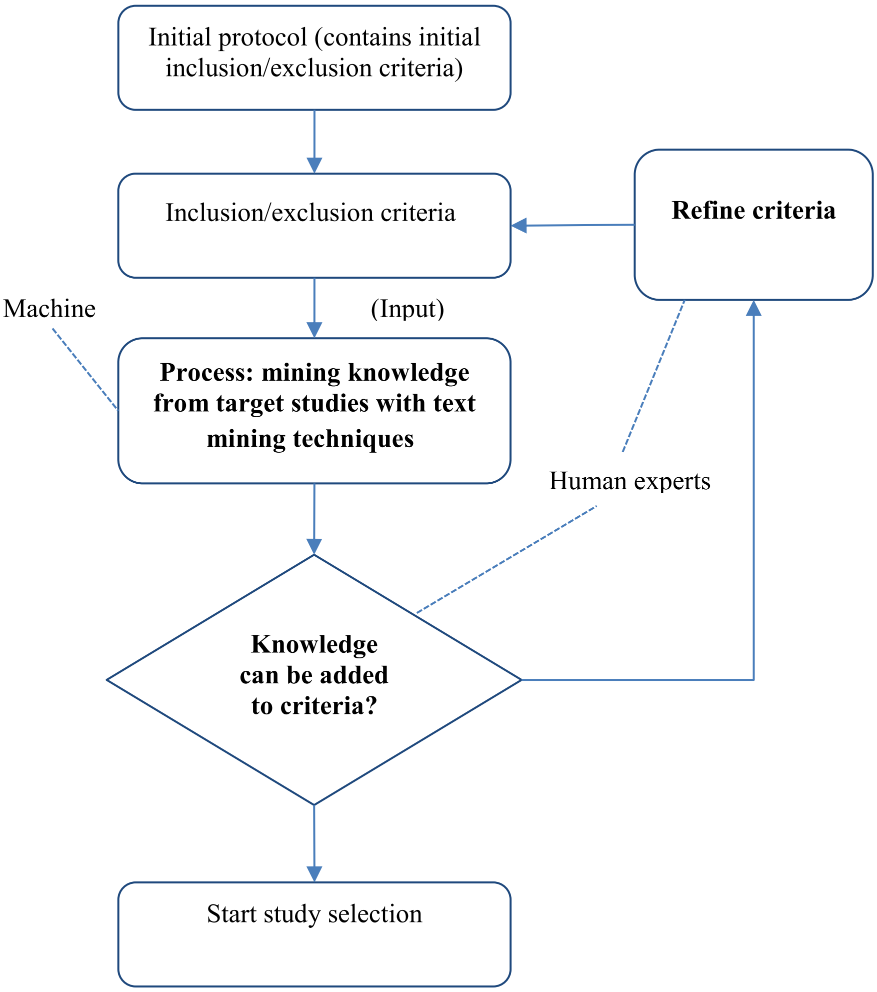

Initial protocol (contains initial inclusion/exclusion criteria)
Inclusion/exclusion criteria
Refine criteria
Machine
(Input)
Process: mining knowledge from target studies with text mining techniques
Human experts
Knowledge can be added to criteria?
Start study selection

Figure 11. Iteratively optimizing inclusion/exclusion criteria.

The optimizing process starts with inputting initial inclusion/exclusion criteria, which is developed in the planning phase in the protocol. The initial criteria influence the core process: mining the knowledge from the candidate studies, and this process should be conducted with the text mining techniques. It requires attention that if the criteria are designed for different rounds of study selection, for each round's the criteria should be refined separately since the information is various in different rounds and targets at various parts of the studies.

Knowledge mining can discover valuable knowledge from the target text, which refers to the target reviewing papers in this framework. Usually the extracted knowledge highly stands for the paper that the knowledge belongs to. Thus, the idea of mining knowledge from the target articles is to give the reviewers a general knowledge of the texts, and understand better of the papers. With the extracted knowledge, the reviewers may refine the selection criteria to be more comprehensively to improve the selection results. Figure 12 describes the steps to conduct text mining inclusion/exclusion criteria.

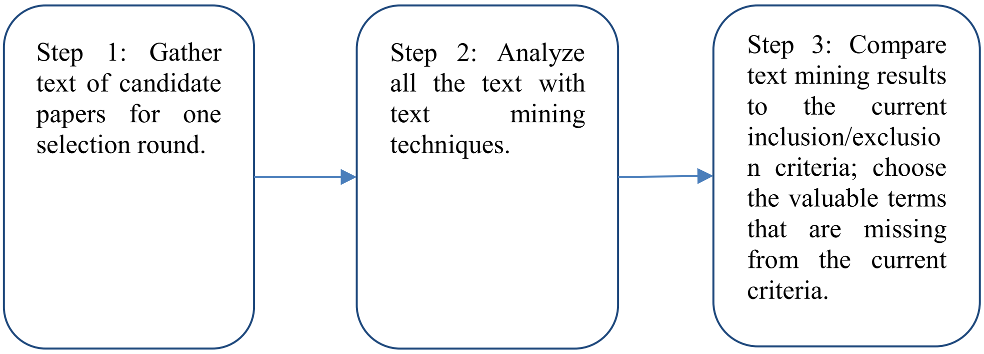

Step 1: Gather text of candidate papers for one selection round.
Step 2: Analyze all the text with text mining techniques.
Step 3: Compare text mining results to the current inclusion/exclusion criteria; choose the valuable terms that are missing from the current criteria.

Figure 12. Steps to conduct text mining with inclusion/exclusion criteria.

When conducting text mining with inclusion/exclusion criteria, a certain part of candidate papers should be gathered for analyzing, the text part can be title, abstracts, keywords, full text. In the step 2, the text mining results can be presented in either quantitative form or descriptive way. Reviewers are supposed to compare the results to the current criteria in order to discover the valuable missing information. The selected terms are defined as "valuable terms". To be noted, the knowledge mining process should be flexible in practical. Text mining is suggested to support the process, not only because it is efficient, but also it is more reliable. But this framework is adaptable to all text-mining techniques or even some other technologies.

After obtained the results of the knowledge mining process, human experts will get involved in the evaluation of the knowledge. If the evaluation result shows that the knowledge is valuable and there is a need to optimize the criteria, the result flows to the process of "optimizing criteria" for human experts. Otherwise when the discovered knowledge is not valuable to the selection criteria, the whole process goes to the end. Once the iterative process ends, reviewers should obtain the optimized selection, which is supposed to be more adequate and more reliable for study selection, while at the mean time study selection process can start.

5.2 Knowledge mining process

The knowledge mining process is an analysis step of discovering knowledge, while the goal is to extract knowledge from a large amount of data. To optimize the decision-making criteria, useful information needs to be discovered from the target texts for study selection. In this knowledge mining process, a text mining technology "automatic term recognition" will be used in discovering the knowledge.

5.2.1 Automatic term recognition

In digital studies, terms as the linguistic representation of concepts are the most critical elements (Sager, 1980). Studies indicate that new terms are being created because of the rapid changes in many disciplines, especially in computer science and engineering fields (Frantzi, 2000). It has been indicated that most of the domain specific terms are compound nouns, which means they collocated uninterrupted; statistics shows 85% of the domain specific terms (Nakagawa & Mori, 2002). The rest of the terms are single nouns, which can be used in the compound nouns (Nakagawa & Mori, 2002). In ATR, the first thing to do is extracting the term candidates from the given targeted text. Then the next important step is to calculate a score for each term candidate, while in fact many researchers have proposed various methods related to approximates term hood. The most widely used method is the surface statistic—tf-idf. “Terms are the linguistic representation of concepts, and they are extremely important for digital libraries”.

ATR techniques are mostly based on frequency, since terms usually tend to appear with high frequencies (Frantzi, 2000). There are several methods available for conducting ATR, and many tools have been implemented for those methods. C-Value is selected as the ATR method in this thesis. C-Value has a good performance in measuring multi words terms, and TerMine’s implementation tool is free to use online. C value method was applied for automatic analyzing text in study selection process.

5.2.2 An ATR method: C-Value

C-value approach is introduced by Frantzi in “Automatic recognition of multi-word terms: The C value/NC value method” (2000). C-value is a domain-independent method for multi-word ATR, which aims to improve the extraction of nested terms. It enhances the typical statistical measure of term occurring frequency for extraction, by means of combining linguistic and statistical information. Hence, C-Value method is sensitive to nested terms as a particular type of multi-word terms.

C-Value uses input as a corpus and produces a group of candidate multi-word terms (Frantzi, 2000). These are ordered by their term hood, which is also called C-value. C-value combines the linguistic and statistical analyses to support the ATR process, while it is also a domain-independent method. Statistical analysis refers to assigning the term hood to a term with four different characteristics: (1) the frequency of the occurrence (Candidate term), (2) length, (3) occurrence frequency of term when appears in the longer candidate term, (4) the number of those longer candidate terms. In linguistic analysis, all the candidate terms are enumerated in text by tagging part-of-speech, extracting word sequences and stop list. C-value has a higher precision than the previous common used frequency of occurrence, according to the studies of their comparisons (Frantzi, 2000).

C-Value calculates the long candidates’ string occurring frequency (Frantzi, 2000). TerMine can be used by uploading a text file of all the combined texts or files to TerMine, the analyzing result shows the term recognized by the tool and they are ranked decreasingly by the C-value. The rankings represent the importance of the terms, mainly means the frequency and article representing result. The results can be retrieved with text or table. The score of C-value in a decreasing order, which can be also interpreted that the higher rank of the term, ranks the terms; the more valuable is the term in the total analyzed text. To measure the term hood, C-Value is given as

$$C - value(a) = \begin{cases} log_2|a| \cdot f(a) & a\ is\ not\ nested \\ log_2|a|(f(a) - \frac{1}{P(T_a)}\sum_{b\epsilon T_a} f(b)) & otherwise \end{cases}$$

Where a is the candidate string, f (*) is its frequency of occurrence in the corpus, Ta is the set of extracted candidate terms that contain the candidate string, P(Ta) is the number of these candidate terms.

6. Initial framework evaluation

To evaluate the proposed iterative framework, two study selection experiments were conducted and the results were evaluated with precision, recall and F_1 score, while comparing to traditional study selectin process. The selection process applied the proposed framework will be named “new process” in the later content, to distinguish from the “traditional process”. The first experiment used the SLR conducted in this thesis, and the second one used another existing practical SLR. The results of the first experiment will be discussed in this chapter.

6.1 Corpus

The corpus selected for the experiment 1 is the previous SLR conducted in this thesis: what is the existing research work in supporting SLR process, and how to improve the study selection process. To be included as a related paper, a study needs to fulfill or should be related to RQ1-1 and RQ1-2: how to improve study selection phase of SLR in SE, and what tools and text mining related techniques can be applied to support SLR process. The detailed SLR process is presented in chapter 3.

The original corpus of the selected corpus contains 997 papers. Since the first evaluation is defined as a small experiment, only 200 papers are randomly prepared for conducting the first experiment, while the database source of the 200 papers is only Scopus.

6.2 Experiment

To use the existing SLR from thesis in comparison experiments, the first question comes: what is the correct results for comparing to. As is already known, it is impossible to define absolute correct SLR results because human experts make the protocol and conduct the process, though the quality of the SLR results varies still. Peer review is a method that commonly used in practical to keep the result objective. As in the original study selection of SLR in thesis, the round 1 and round 2 were conducted by both thesis author and thesis supervisor (peer review), and the selection criteria were refined iteratively during the selection, it is reasonable to assume that the original selection results are rather objective and the final used selection criteria are rather adequate. Therefore, the selection results of SLR were considered as the **correct** selection results, while when the author conducting the experiments alone, these correct selection results would be used for comparison.

Therefore, the idea of the first experiment was to experiment traditional selection process and new process on the basis of “raw protocol”, which hadn’t gone through the review conducting phase and the selection criteria hadn’t been revised. The selection results from the existing peer review was used for evaluating the performance of the two methods. As “raw protocol” was used, the “raw” un-refined selection criteria were adopted. The thesis author took the responsibility of undertaking the study selection work. Also, the thesis author was supposed to ignore the final selection criteria from the original SLR and keep the process subjective.

The content of evaluating the prepared papers are: title, keywords, and abstracts, which were screened in different rounds of study selection. The first two rounds of selection were conducted in the experiment 1: round 1 and round 2, while they targeted at different components of the papers. Round 3 was not considered for experimenting because when it reaches round 3, very few papers are left and it would be hard to evaluate and compare the precision and recall. In round 1, titles and keywords were screened together; in round 2, only abstracts are used for study selection. Comparison experiments were conducted between traditional process and new process, for round 1 and round 2 respectively.

The selection criteria were refined during the "peer review" in the original SLR, while the final criteria have richer context. In terms of round 1 and round 2, the only diffidence between old criteria and final criteria is IC1 was refined to include two more related techniques: machine learning and pattern recognition, which was suggested by the thesis supervisor. Therefore, the idea of experiment 1 was to use the old selection criteria without refining in peer review, experimenting the performance of traditional process and new process.

In this thesis, the correct included papers from candidate papers are defined as **positive** papers. Correspondingly, the included papers in the experiment which belongs to the positive papers' group are defined as **true positive (TP)** papers. These two indexes were used in calculating precision and recall of experiment results. Positive papers should be available before conducting the experiment. Since the study selection results from original SLR were considered as correct, the selected papers from the original selection results were referred as positive papers

Table 5 summarizes the data set prepared for experiment 1. The same data set were used in both traditional process and proposed process.

Table 5. Data set prepared for experiment 1.

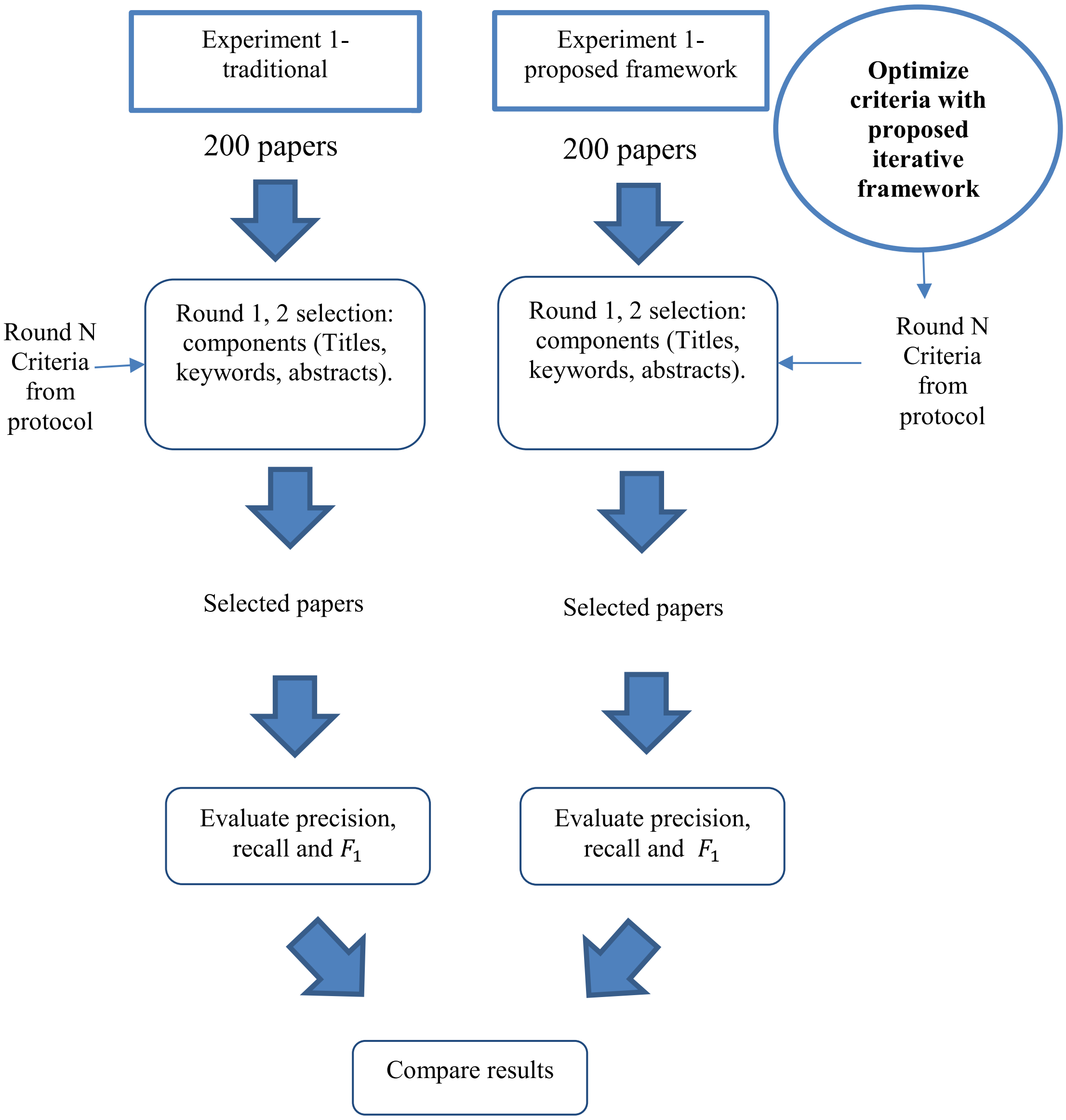
Data Set	**Description**
Sample volume	200
Database	Scopus
Content	Title, keywords, abstracts.
Selection rounds	Round 1 (title & keywords), round 2 (abstracts)
Round 1: TP papers + FN papers	84 (results from peer review in original SLR)
Round 2: TP papers + FN papers	29 (results from peer review in original SLR)

The "raw criteria" from existing SLR is adopted, while comparing to the final criteria of round 1 and round 2 in the original SLR, IC1 has two less related techniques: machine learning and pattern recognition. The criteria of round 1 and round 2 are listed:

- IC1: Fulfills inclusion criteria from the previous rounds AND Is related to SLR/systematic review/mapping study, or is related to SLR/systematic review/mapping study tool/method, or is related to techniques such as text mining/natural language processing/information extraction, or is related to any other possible techniques/application related to the above mentioned terms of techniques, e.g. "neural network").

- EC1: Fulfill the exclusion criteria from the previous rounds.
- IC2: Fulfills inclusion criteria from the previous rounds AND (Is a SLR or mapping study that used tools or techniques for supporting automating of the process OR is a SLR or mapping study of supporting automatic SLR/mapping study OR is a study about techniques or methods that can be used for supporting automatic SLR or mapping study process), OR is a study that discussed how to support text mining.
- EC2: Fulfill the exclusion criteria from the previous rounds.

The comparison experiments between traditional study selection process and proposed selection process with criteria were conducted separately.


Experiment 1- traditional
200 papers
Experiment 1- proposed framework
200 papers
Optimize criteria with proposed iterative framework
Round N Criteria from protocol
Round 1, 2 selection: components (Titles, keywords, abstracts).
Round 1, 2 selection: components (Titles, keywords, abstracts).
Round N Criteria from protocol
Selected papers
Selected papers
Evaluate precision, recall and F_1
Evaluate precision, recall and F_1
Compare results

Note: for each round, the precision, recall and F_1 will be evaluated and compared.

Figure 13. Experiment 1 process.

Figure 13 describes the process of experiment 1, while the traditional process and new process were applied to selecting relevant papers from the prepared papers. When conducting study selection process for round N, the corresponding selection criteria were used for study selection. The traditional process and the new process are identical except for in the new process, the round N selection criteria were optimized with the proposed iterative framework. When the corresponding criteria are ready to use, the round N study selection process can start; the study selection results were evaluated by precision p, recall q and F_1 score. The equations for the measures are given as:

$$Precision\ p = \frac{TP}{TP\ + NP} = \frac{number\ of\ TP\ papers}{number\ of\ selected\ papers}$$

$$Recall\ \ q = \frac{TP}{TP + FN} = \frac{number\ of\ TP\ papers}{number\ of\ postive\ papers}$$

$$Measure\ F_1 = \frac{2\ \cdot Precision\ \cdot Recall}{Precision + Recall} = \frac{2pq}{p+q}$$

In chapter 3, the evaluation measures of precision, recall and F_1 score are introduced. They are very common measures for assessing how successful a text categorizer is. For this experiment, the higher scores of p, q and F_1, the better results of study selection, as they indicate that the included papers are more valuable and less valuable papers are missed.

A selected tool implemented the C-value analysis: **TerMine**. TerMine is a tool developed by University of Manchester, which supports the automatic recognition of terms. In TerMine demonstrator, the AcroMine acronym recognition and multiword term extraction (C- value) are integrated together. TerMine is a free tool opens to the public and the user could get access to it. TerMine is a term extraction tool developed at the UK National Centre for Text Mining (NaCTeM). It uses a domain-independent method (based on the C-value measure) to extract candidate terms from English text for consideration by the oncologist or terminologist; in addition, it is particularly oriented towards extraction of candidate multiword compound terms.

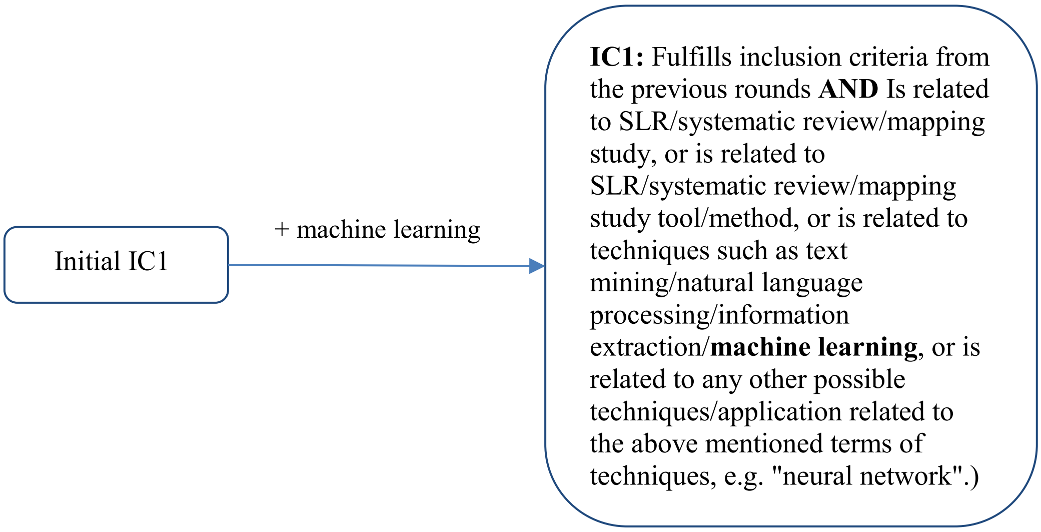
Theoretically, the above initial inclusion/exclusion criteria were optimized in the proposed framework of experiment 1. The results of the comparison experiment are presented and analyzed in the next section.

6.3 Results

Results of traditional process and new process are independent. The two study selection processes are conducted separately with the above data set. In the results, for each round the final selection criteria will be presented after study selection, together with the included papers as TP and FP papers. By comparing the included papers to TP papers, which were prepared in the dataset, the precision, recall and F_1were evaluated for both selection processes. The above measures indicate which selection process produced better results.

Result of traditional process

In the traditional study selection process, including/excluding criteria were revised slightly in inclusion criteria IC1. The inclusion criteria IC1 were refined while machine learning was included as related to text mining techniques (figure 14).



IC1: Fulfills inclusion criteria from the previous rounds **AND** Is related to SLR/systematic review/mapping study, or is related to SLR/systematic review/mapping study tool/method, or is related to techniques such as text mining/natural language processing/information extraction/**machine learning**, or is related to any other possible techniques/application related to the above mentioned terms of techniques, e.g. "neural network".)

Figure 14. Refine selection criteria in the traditional process

The results of traditional study selection are described in table 6. 71 papers and 26 papers were selected after round 1 and round 2 respectively. After round 1, by comparing the 71 papers to the positive papers, which are consisted of TP papers and FN papers, 66 papers from selected papers were found to be TP papers while the rest 5 papers were FP papers. In the round 2 selection results, 20 papers were TP papers and the rest 6 papers were FP papers. Precision, recall and measure F_1 were calculated with the given equations in this chapter.

Table 6. Results of tradition study selection process round 1, round 2

	Round 1	**Round 2**
Initial papers	200	200
Selected papers	71	26
TP papers	66	20
FP papers	5	6
TP papers + FN papers	84	29
Precision p	93%	76.9%
Recall q	78.6%	69.0%
measure F_1	85.2%	72.7%

Result of new process

In the new process, a C-value analysis were conducted for analyzing the content to be reviewed before the study selection process started. In the round 1, TerMine uses titles and keywords as input and produces a list of candidate multi-word terms. Round 2 C-value analysis used similar procedure but with abstracts as input.

In the table 7 and table 8, part of the C-value analysis results is showing, while the actual results have a long list rank of terms up to hundreds. The rankings represent the importance of the terms, which mainly means the frequency and article representing result.

Table 7. Part of the round1 C-value analysis results.

Rank	**Term**	**Score**
1	Systematic literature review	66.047653
2	Software product line	48.911942
3	Systematic literature	45.131866
4	Literature review	43.301723
5	Product line	41.547009
6	Systematic review	33.673912
7	Software product	32.311474
8	Systematic mapping	28.818182
9	Software engineering	28.666666
10	Software development	14.75
22	**Machine learning**	5
28	**Visual text mining**	3.962406
32	**Visual analysis approach**	3.169925
38	**Natural language**	3
82	**Text classification**	2

Table 8. Part of the round 2 C-value analysis results.

Rank	Term	Score
1	Systematic literature review	42.114716
2	Software engineering	34.836735
3	Literature review	29.470589
4	Systematic literature	27.799999
5	Systematic mapping	18.75
6	Elesvier b.v	15.735294
7	Systematic review	14.9
8	Software development	12.941176
9	Research question	11.888889
10	Web application	11.6
65	**Natural language text**	3.169925
114	**Text classification**	2

For each round, the process conductor went through the C-value analysis results from beginning, and compared them to the inclusion/exclusion criteria of this round. When a candidate term is judged to be valuable, it was used for refine the criteria. All the selected valuable terms are marked bold in table 7 and table 8.

In the new process, inclusion/exclusion criteria are revised only in criteria IC1 and IC2. The inclusion criteria IC1 and IC2 are refined with the selected valuable terms from C-value analysis, as presented in figure 15.

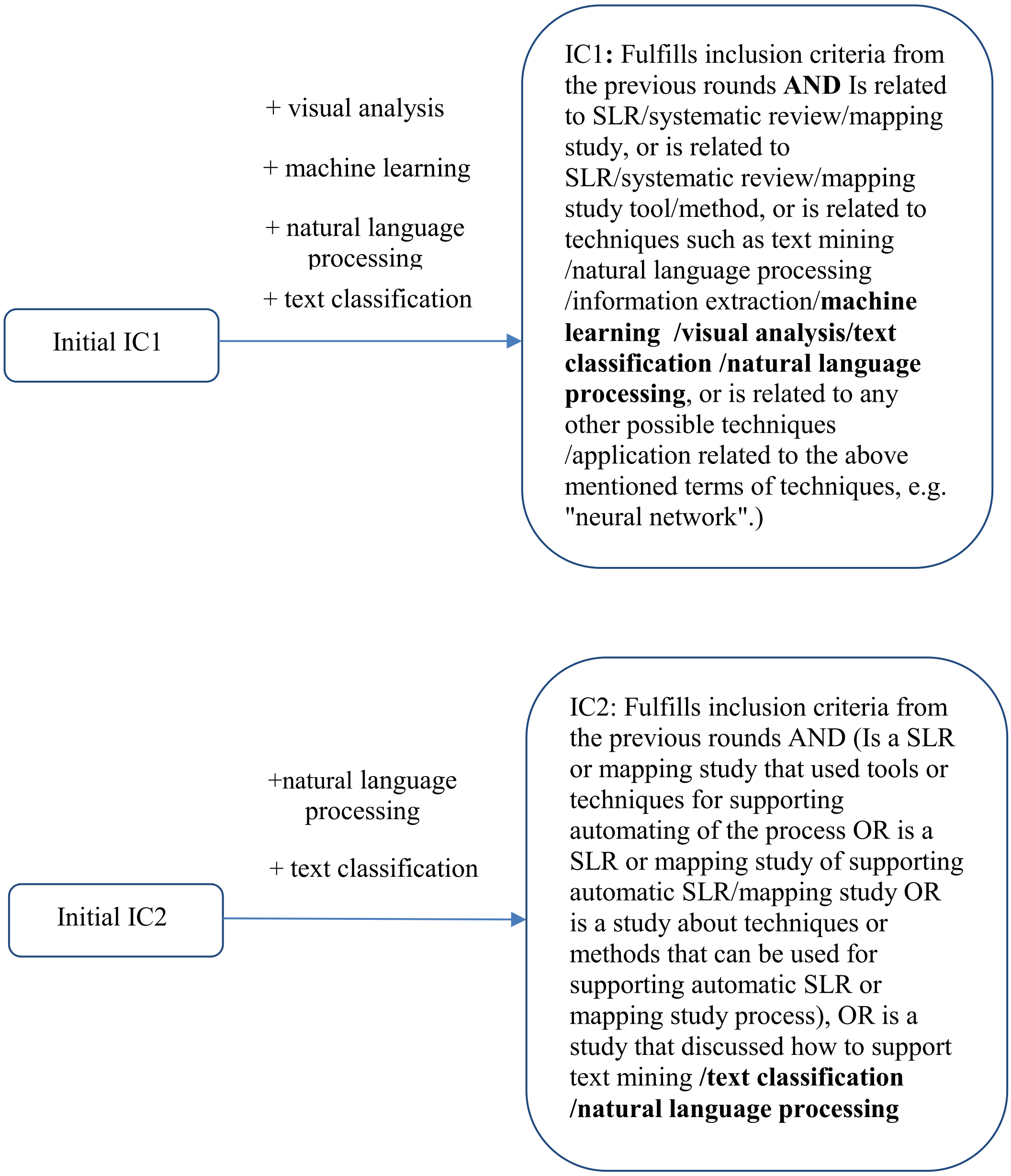

+ visual analysis
+ machine learning
+ natural language processing
+ text classification
Initial IC1
IC1: Fulfills inclusion criteria from the previous rounds AND Is related to SLR/systematic review/mapping study, or is related to SLR/systematic review/mapping study tool/method, or is related to techniques such as text mining /natural language processing /information extraction/machine learning /visual analysis/text classification /natural language processing, or is related to any other possible techniques /application related to the above mentioned terms of techniques, e.g. "neural network".)
+natural language processing
+ text classification
Initial IC2
IC2: Fulfills inclusion criteria from the previous rounds AND (Is a SLR or mapping study that used tools or techniques for supporting automating of the process OR is a SLR or mapping study of supporting automatic SLR/mapping study OR is a study about techniques or methods that can be used for supporting automatic SLR or mapping study process), OR is a study that discussed how to support text mining /text classification /natural language processing

Figure 15. Refine selection criteria in the new process.

The initial selection criteria were refined much more in the new process, comparing to the traditional process. In the round 1, there are five terms are selected as valuable terms for refining the selection criteria, which are machine learning, visual text mining, visual analysis approach, natural language and text classification. In theory, these five terms can be added to the criteria as they remain, but the process conductor chose to leave out "visual text mining" because "text mining" and "visual analysis" should already covered it. In the round 2, only "natural language text" and "text classification" are selected, the process conductor decided to add "natural language processing" and "text classification" as similar to "text mining".

After selection criteria optimization, the refined criteria were used for the actual study selection. Table 9 presents the selection results in the process. The TP papers results were also come from comparing the selected papers to the prepared TP and FN papers. The evaluation measures were calculated.

Table 9. Results of new process round 1, round 2.

	Round 1	**Round 2**
Initial papers	200	200
Selected papers	84	26
TP papers	79	20
FP papers	5	6
TP papers + FN papers	84	29
Precision p	94.0%	75.8%
Recall q	94.0%	86.2%
measure F_1	94.0%	80.7%

The results of new process were compared to traditional process results in table 6 for round 1 and round 2 separately. Based on the above data, the round 1 and round 2 comparisons results are showing in figure 16-a and figure 16-b.

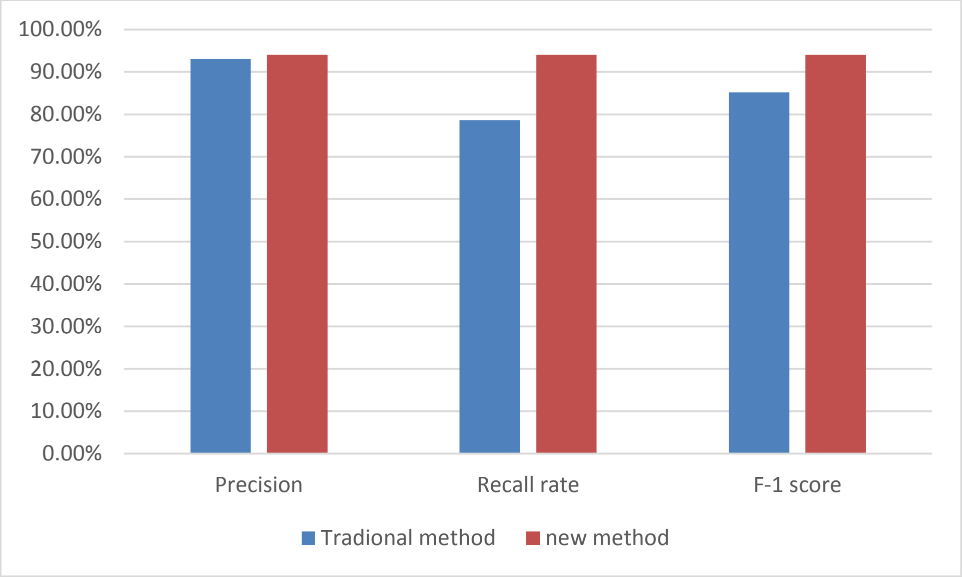


Figure 16-a. Round 1 result comparision between traditional process and new process.

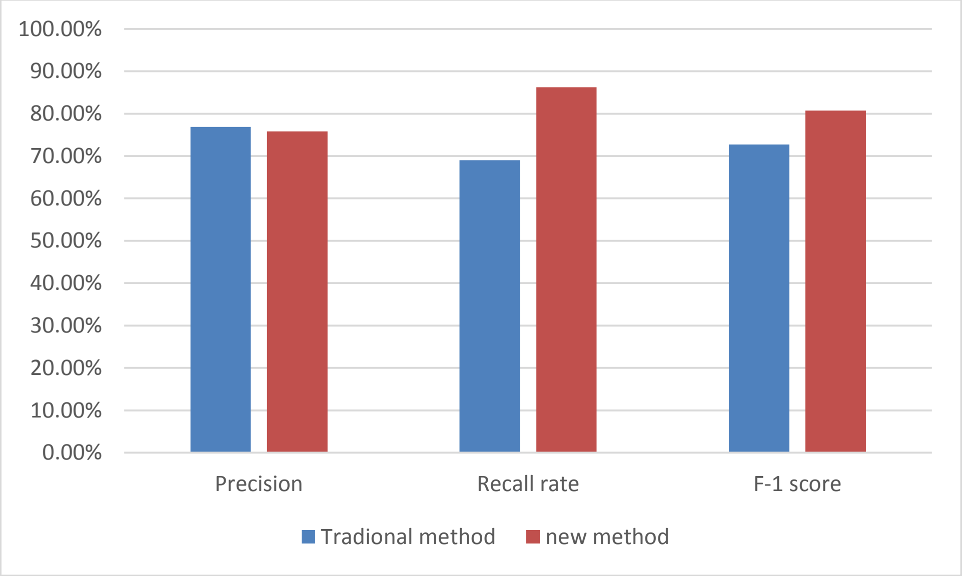


Figure 16-b. Round 2 results comparison between traditional process and new process.

In the round 1, precision does not show big difference between two methods—the new method has a slightly higher rate. However, in terms of recall, the new method has a much higher value (94%), while the traditional method only has 78.6% recall.

Round 2 is a continued experiment after round 1, therefore the selection result from round 1 was used for the initial papers of round 2. Like the round 1, precision value still does not show difference between the two methods, though the traditional method has a slightly higher value. But the new method shows strength in the recall again, with recall of 86.2%, while the traditional method recall only reaches 75.8%.

6.4 Evaluation

This experiment was conducted for evaluating the iterative framework that proposed for improving the study selection process. The traditional study selection process and the new study selection process were evaluated based on the process performance in study selection accuracy and the primary studies' recall.

In the traditional method, a common problem of the initial selection criteria is lack of required information, which also caused the criteria needs to be revised more often. Criteria revision's influence in the whole study selection varies, which depends on the importance of the new term and the number of previous screened papers that got influenced. The process with proposed framework benefits from the optimized criteria, while the results show the criteria have only be revised once in the experiment.

Comparing to the results of traditional method without refined selection criteria, the new study selection process shows higher accuracy, which indicates the negative papers got excluded in earlier round. For instance, the more negative papers got excluded in the round 1, the less papers that reviewers need to be reviewed in the round 2, which reduces the total time cost in the review conduct phase. The new study selection process results also show higher recall, which mainly reflect the quality of a SLR. The higher recall, the more positive papers are included in the primary studies, and thus the better quality because the objective of SLRs is to collect evidences regarding a certain research problem.

To sum up, the initial evaluation of iterative framework shows that optimizing the selection criteria can improve the selection results in the SLR study selection process, though the precision rate does not make much difference. To further prove the effectiveness of the proposed iterative framework, a validation procedure is conducted with an existing SLR. The validation experiment is presented in the next chapter.

7. Framework validation

A further experimental evaluation is conducted to validate the effectiveness of the proposed optimizing criteria framework. In this experiment, an existing large systematic literature review were used for the experimentation.

7.1 Corpus

The corpus for the experiment 2 is from a completed large systematic literature review (Taušan, 2017), which has the original studies of 5169 papers. This SLR studies how is choreography used in embedded system development domain. Eigth relevant database were searched and 5169 relevant scientific publiscations were found in the initial search. The keywords and seach string were given, as presented in table 10.

Table 10. The given keywords and search string

Keywords and search string	
Part I	choreography
	AND
Part II	embedded OR automotive OR telecom* OR automation OR healthcare OR aerospace OR robotics OR “internet of things" OR "web of objects" OR mobile
Part III	AND
	service-oriented OR soa

Due to limited time and corresponding to experiment 1, 200 papers were selected randomly. The selected data set were used in both traditional study selection process and the new process.

7.2 Experiment

The overall statistics of data set for experiment 2 is shown in the following table 10. It is notable only round 2 of the selected large SLR were used for experimenting, because this SLR’s round 1 uses general information as criteria. The publication was excluded in cases when 1) the written language was not English, 2) it was a duplicate entry, 3) the data were erroneous, 4) it was non-peer reviewed, 5) it was a textbook, 6) it was a MSc or PhD theses, 7) it was a proceedings preface, keynote or panel discussion or 8) it was not from the software engineering field. Therefore, there is no meaning of experimenting round 1 between traditional process and new process.

However, in the round 2, the selection process is based on reading the abstract to include and exclude papers. The TP papers were adopted the from original SLR results. As the SLR involved three reviewers and the process was properly done, it is reasonable to consider the results are reliable. The round 2 selection results are referred as TP papers in this experiment. The thesis author was again responsible for both the traditional and new processes. Table 11 describes the data set for experiment 2.

Table 11. Data set prepared for experiment 2.

Data Set 1	**Description**
Sample volume	200
Database	IEEE, ACM, citeSeer, scopus, SpringerL, ProQuest, Google scholar, Science Direct,
Content	Abstracts.
Selection rounds	Round 2
Round 2: TP papers + FN papers (number)	43 (results from original SLR)

To be noted, there is no initial selection criteria (without refining during selection) available from this SLR, but only the final criteria. The final criteria could not be used for the same reason in the experiment 1, because this experiment was aiming at optimizing the criteria before study selection started. Therefore, based on the research questions and search string, the initial inclusion/exclusion criteria for round 2 study selection were re-formulated in this experiment as following:

- IC2: Fulfills inclusion criteria from the previous rounds AND (Fulfills containing at least two of the follow content categories: (1) choreography, conversation, composition, orchestrate. (2) Embedded system, automotive, telecom, automation, healthcare, aerospace, robotics, Internet of Things, web of objects, mobile. (3) service-oriented or soa.
- EC2: If the paper is about enterprise systems.

Corresponding to the experiment 1, the comparison experiments between two processes were conducted separately by the thesis author. The experiment 2 followed the experiment 1 process defined in figure 13, but applying the new data set and only round 2 were experimented in the study selection process.

7.3 Results

After the comparison experiments conducted, the finalized inclusion/exclusion criteria and selection results were presented and compared between the two study selection processes. Eventually the selection results were evaluated with precision, recall and F_1 score.

Result of traditional process

There is no including/excluding criteria modification during the traditional process. The results of traditional study selection are described in table 12. 156 papers and 33 papers were selected after round 1 and round 2 respectively. After round 2, by comparing the 33 papers to the positive papers, which are consisted of TP papers and FN papers, 27 papers from selected papers were found to be TP papers and the rest 6 papers turned to be FP papers. Precision, recall and measure F_1 were calculated accordingly.

Table 12. Results of tradition selection round 2.

	Round 2
Initial paper	200
After Round 1 selection	156
After Round 2 selection	33
TP papers	27
FP papers	6
TP papers + FN papers	43
Precision p	81.8%
Recall q	63.0%
measure F_1	71.2%

Result of new process

In the new process, a C-value analysis was conducted for analyzing the abstracts of the candidate papers, before the study selection process started. In the round 2, TerMine used abstracts as input and produces a list of candidate multi-word terms.

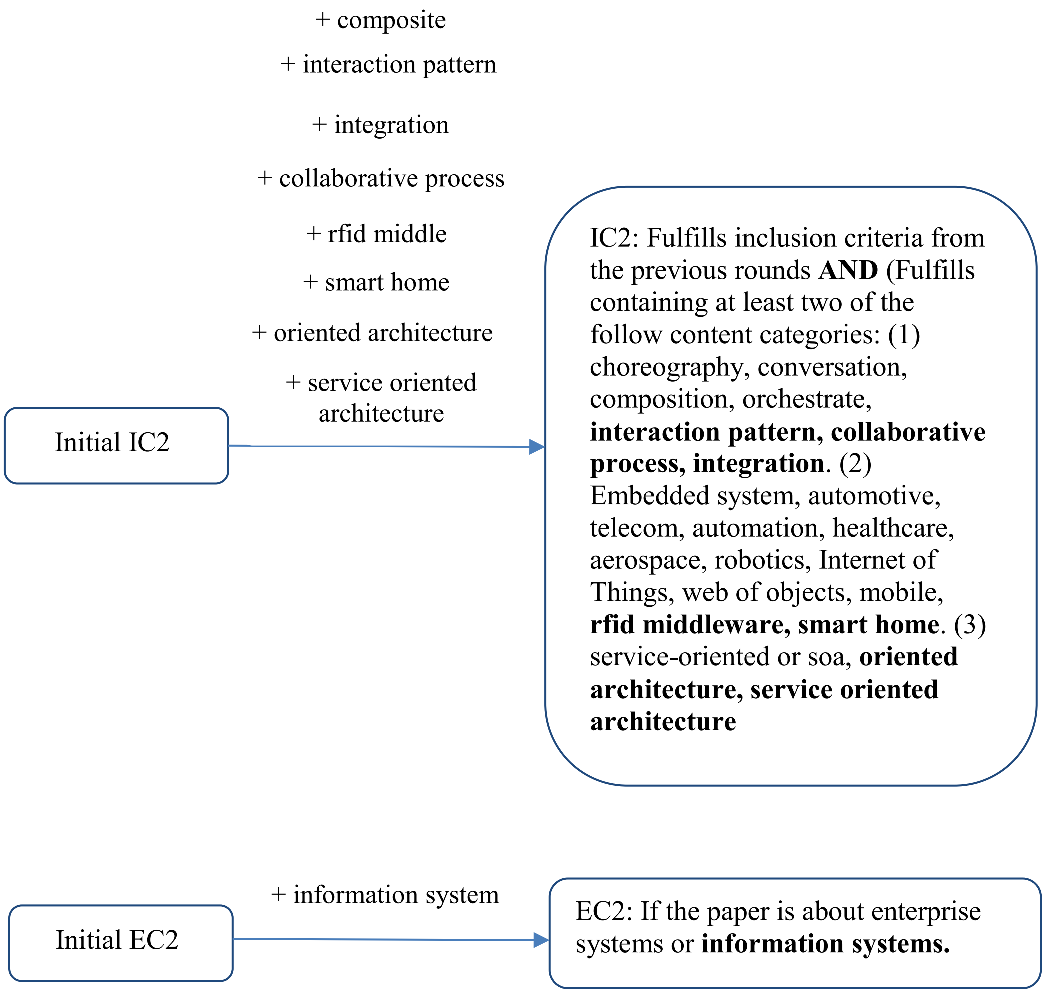
In the table 13, part of the abstracts C-value analysis results is showing. The highest 10 terms are listed, together with 10 terms that selected as valuable terms for criteria refining.

Table 13. Part of the abstracts C-value results.

Rank	Term	Score
1	Wed service	214.403229
2	Business process	50.68919
3	Service composition	44.766666
4	Semantic web	32.878788
5	Semantic web service	29.249763
6	Web service composition	19.01955
7	**Service-oriented architecture**	15
8	**Service oriented architecture**	12.6797
9	B2b e-commerce hub	11.094737
10	Service oriented	10.923077
13	**Oriented architecture**	10.285714
16	**Composite web service**	7.924612
18	**Composite service**	7
22	**Composite web**	6
22	**Information system**	6
35	**Interaction pattern**	4.916667
66	**rfid middleware**	3
132	**Service integration**	2
132	**Smart home**	2
132	**Collaborative process**	2

Before round 2 started, the process conductor went through the C-value analysis results from beginning, and compared them to the inclusion/exclusion criteria of round 2. The selected valuable terms (bold text in table 13) were used for refine the criteria.

In the new process, inclusion/exclusion criteria are revised in both criteria IC2 and EC2. The round 2 selection criteria were refined with the selected valuable terms from C-value analysis (figure 17).



IC2: Fulfills inclusion criteria from the previous rounds **AND** (Fulfills containing at least two of the follow content categories: (1) choreography, conversation, composition, orchestrate, **interaction pattern, collaborative process, integration**. (2) Embedded system, automotive, telecom, automation, healthcare, aerospace, robotics, Internet of Things, web of objects, mobile, **rfid middleware, smart home**. (3) service-oriented or soa, **oriented architecture, service oriented architecture**

EC2: If the paper is about enterprise systems or **information systems.**

Figure 17. Refine selection criteria in the new process.

In the abstracts' C-value analysis results, there are twelve terms are selected as valuable terms for refining the selection criteria, which are service-oriented architecture, service oriented architecture, oriented architecture, composite web service, composite web service, composite service, composite web, information system, interaction pattern, rfid middleware, service integration, and smart home. Considering there are overlaps in these terms, the process conductor eventually decided to add eight terms to the IC2, as presented in figure 17. Term "information system" was selected to be included in EC2.

The refined criteria were used for the actual study selection. Table 14 presents the selection results in the process. The TP papers results were also come from comparing the selected papers to the prepared TP and FN papers.

Table 14. results of proposed framework selection round 2.

	Round 2
Initial paper	200
After Round 1 selection	156
After Round 2 selection	42
TP papers	34
FP papers	8
TP papers + FN papers	43
Precision p	81.0%
Recall q	79.0%
measure F_1	80.0%

The results of new process were compared to traditional process results in table 12. Based on the above data, the round 2 comparisons results are showing in figure 18.

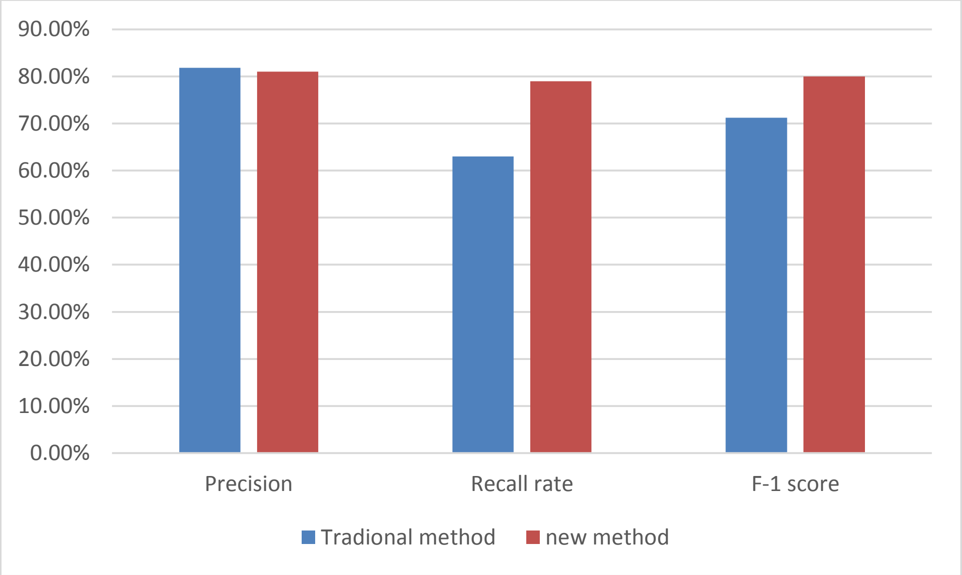


Figure 18. Experiment 2: result comparision between traditional method and new method.

In the experiment 2, the traditional study selection result has 81.8% precision rate and 63.0% recall separately, while the new method result has almost the same precision (81%) but a much higher recall (79.0%) which exceeds 16.0% over the traditional process. Because of the big advantage in recall, the new method has a higher F_1 score (80.0%), which is 8.8% over the score of traditional process (81.2%).

7.4 Evaluation

According to the validation experiment results, the optimizing criteria iterative framework has been validated that it improves the result of the SLR process in the study selection process. The proposed framework has an obvious advantage of recall over the traditional method though the precision of the two processes are quite similar. Thus the F_1 score of the proposed framework also shows obvious advantage.

The experiment 2 further validated the new study selection process shows higher recall, which reflects the quality of the SLR results. The selection accuracy does not make much difference when applying the proposed framework, but that was mainly caused by the very low recall in the traditional method. In the traditional process, only 33 papers are selected in total, with 27 positive papers included, which means it missed 16 positive papers even though the accuracy of including positive papers is relatively high. Therefore, it is not reasonable to compare only measure precision, but the precision and recall combines score F_1 should be used for evaluating the selection results.

To sum up, the second experiment further validated the effectiveness of the proposed iterative framework. The study selection applied with the proposed framework shows a much higher recall than the traditional selection process.

8. Discussion and implications

In this chapter, findings of this study are summarized based on research questions. Contribution and limitation are also discussed in both theory and practical point of view. Additional observations about this study can be found in the last section 8.4.

8.1 Findings of this study

Two research questions are defined to explore how to support SLR in Software Engineering field. The RQ1 is about studying the facts and issues of SLR study selection process; RQ2 is looking for solutions to improve the study selection process of SLR in SE.

RQ1: What are the existing tools and methods in supporting SLR process?

RQ2: How can study selection phase of SLR in SE be improved?

The existing issues of SLR are discovered in different phases including identification of research, primary study selection, study quality assessment and data extraction. Since human experts conduct the traditional SLR process, the main problems in the SLR process can categorized to two types. First critical but typical issue is that in a SLR process, the human experts got involved in a huge amount of manual work, which requires reviewers to go through all the SLR phases. A SLR process period can vary from a few months to several years, depends on the research questions and focus. It is notable that study selection process involves the most biased result from the manual work. Subjectivity from the human experts is also a problem that can lead to biased SLR results.

The SLR results collected the existing evidences of supporting automatic or semi-automatic study selection process. The study types include "review study", "tool", "theory or concept" and "others". The study types distribution analysis shows that the "theory or concept" is the most common research work type. The studies about "tool" to support study selection process have the second most amount of papers, while most of the tools are more about reducing the manual work or fastening the selection process. In the SLR result, there are also several review studies summarized the related work about SLR study selection process supporting. Overall, these review studies have broader research problem than research questions in the SLR of this thesis. The "other" type studies contains potential knowledge for supporting study selection of SLR.

In the recent decades, there are several tools developed for supporting SLR while some of them have received quite lots of attention. SLuRP is a unified SLR program, which uses open source web-enabled database. StArt supports all phases in SLR process, thus users can benefit from the good process management. SLR-TOOL is a free use tool supports all the phases too, similar to StArt. Different from all the previous tools, tool "Revis" uses text mining to support study selection. Apart from the above tools, there are some other tools like PROJECTION EXPLORER are also used for SLR. However, these tools can improve SLR process efficiency and results but none of them are mature enough to be adopted in general. Nowadays SLRs in SE field are still mostly done manually.

Techniques can be applied in SLR study selection process are classified to three groups: visual techniques, text mining and other related techniques. The most common and proved useful techniques are visual techniques and text mining according to the SLR results, and it shows good results when these two techniques are combined. Visual techniques provide a better view present to human experts than reviewing the pure text content. A typical use of visual techniques is studying the relationships of paper relevance by machine learning, which can benefit reviewers to select papers. Text mining related techniques mainly conducts information retrieval, information extraction and data mining. Text mining is known as a very useful knowledge for supporting SLR study selection, nevertheless the existing tool or method are too difficult to use in practical. Revis uses both visual techniques and text mining, but it requires the human experts to understand the techniques. Other techniques are found to have potential to support reviewing phase, such as pattern recognition, coding, graphs using.

An iterative framework is proposed for the SLR study selection, which is aiming at optimizing the inclusion/exclusion criteria. Why optimize the criteria? In the study selection process, there is an intimate correlation between the inclusion/exclusion criteria and primary studies. Existing studies already demonstrated that the criteria decide the quality of the SLR results, though it also varies a lot among difference human experts and their experience. Inadequate protocol from the planning phase could result in biased selection result easily.

The relevant analysis leads to a new idea of optimizing the inclusion/exclusion criteria for study selection process. A simple iterative framework is proposed and evaluated with an experiment conducted between traditional method and new method with the proposed framework. Considering the small sample used in the first evaluating experiment, and also the subjective reasons, a validation experiment with a practical SLR sample was conducted for a further evaluation. The validation experiment adopted an existing SLR from a PhD student's research work, with the all the search questions, background, raw data and SLR results provided. In both experiments' evaluation results, the comparison of accuracy and recall between the two methods shows that the new method with optimized criteria has apparent advantage in recall, though the selecting accuracy does not make big difference between the two methods. When combining the precision, and recall to F_1 score, the proposed framework has a much better performance. The good recall can be interpreted that more positive papers are included in the primary studies with less mistaken left out papers, which influences the primary studies quality. It has been concluded that the proposed iterative framework improves the study selection process of SLR in SE field.

8.2 Contribution

The main contribution of this study is the iterative framework proposed to improve the study selection process of SLR. The original framework is designed to improve the SLR protocol, more specifically the inclusion/exclusion criteria, which finally improve the selection process. The extended framework with knowledge discovering method TerMine has been evaluated with two experiments and it has been proved that the proposed framework does support the study selection process by improving the accuracy of including the positive studies.

The new method for study selection brings fairly big benefit to researchers, especially to those who are new to a research field or topic. Typically, when human expert lack of required knowledge to conduct the study selection, it is quite difficult to formulate good inclusion/exclusion criteria. However, these will result in either the researchers need to spend plenty of time repeating the tedious study selection work, which happens usually because of the criteria revision, or the reviewers limited knowledge lead to a biased result. The proposed new method facilitates reviewers gaining information before starting the actual study selection work.

In addition, this thesis presents a complete systematic literature review of study selection process of SLR, which contains the available results of supporting study selection process; the confronting problems and the possible techniques. Despite of the fact that there are several SLR or mapping study that have done the review of supporting SLR process, none of them was entirely focus on the study selection process. Therefore, the SLR results from this thesis can benefit people who will be doing the related researching in the future.

8.3 Limitation

The two experiments are based on rather small data sets due to the limited time of study author. 200 articles were chosen as the volume based on some existing studies, which shows that the 200 articles are capable to support producing reasonable study selection results under reasonable time.

There are possible subjective factors that might affect the results. However, the assumption of keeping the experiment as objective as possible has eliminated the impacts. Further experiments in the future may need for the future study, especially with a large set of studies for conducting selection. The two experiments may have very different time consuming and accuracy, mainly because of the experiments conductor's knowledge of the two tasks and also the complexity of the experiment. The data set of experiment 1 turns to be much simpler and has more clear inclusion criteria, besides, it is an area that the human expert quite familiar with. However, the data set from the validation experiment are more complicated and the selection experiment involves much more individual opinions.

Besides, in the experiments, the method used for optimizing the including/excluding criteria is quite simple. TerMine is an online free tool for conducting C-Value analysis, which discusses valuable terms from a large set of the unstructured text. However, it is important to notice that emphasis of this thesis is evaluating the original iterative framework for improving selection process. It is free to choose any tools that would perform well in discovering the unstructured text. In the future work, the proposed framework can be extended with different tools to test the performance.

8.4 Additional observations

TerMine results quality shows different effects in the evaluation experiments. In the experiment 1, TerMine C-Value analysis was conducted separately for the round 1 and round 2, of which the contents are titles & keywords and abstracts respectively. The content of round 2 is much more than that in round 1, therefore the C-value analysis result of round 2 has a much longer list of terms with C-value. Besides, the result of round 2 turns to be better in the terms ranking, which provides valuable information even in the very beginning of the ranking list, which also means TerMine C-value analysis performs better in the second-round selection with the abstracts.

Different from experiment 1, experiment 2 only involved one round TerMine C-Value analysis, which was targeted in the abstracts parts of the studies. The terms ranking result turns to be good as the second round of the experiment 1 that valuable terms appeared in the very beginning of the list. Therefore, based on the experiment 1 and experiment 2, it obviously that abstracts perform better in the TerMine analysis.

In the experiments, terMine was selected as the main tool that used for optimizing the including/excluding criteria, which is quite simple to use. However, the tool is not our emphasis. It is used for demonstrating that the proposed concept of optimizing criteria performs well in improving the SLR's study selection process.

Topic Modeling is another Toolkit applying machine learning to analyze large amount of unlabeled text. The "topic" concept here is consisted by a cluster of words, which occur frequent at the same time. By using the contextual clues, Topic Modeling connects words that have similar meanings and also distinguishes them by the multiple meanings of the words usages.

Also, it should be noted that the two experiments for evaluation and validation are conducted under an assumption: the human expert is objective enough. The human expert who conducted the study selection was trying to keep the result, as much objective as possible, but still it is a fact that human cannot act like a robot and human manual work always involves in the subjective factors. Leaving out the individual knowledge and the previous memory of study selection the same or related the study.

Unbiased results and less time consumed are the two indicators of study selection process performance. The initial idea to evaluate the proposed iterative framework was to use results' accuracy and measure the time consumed in the study selection process. However, the time was very difficult to track and record due to limited experiment environment, thus the time indicator was removed eventually. For the future research, time indicator is a good research orientation.

9. Conclusions

The main task of this thesis can be summarized to two parts: find out the existing research work in support SLR study selection process, and explore how to support the study selection process of SLR in software engineering field. Based on the achievements contributed by other researchers, a new iterative framework was proposed to support the study selection process.

Typically, there are four phases in SLR process: identification of research, primary study selection, study quality assessment and data extraction. The SLR process is expensive due to the study selection phase which consumes both human resource and time. More or less, for each phase there will be some existing problems when conducting SLR in practice. Among all the issues, the most critical one is that the results can be biased by manual work or applying other methods in state-of-the-art. As SLR method gets more widely used, how to improve the SLR process has gained increasingly attention. Besides, in last decade there are increasingly papers have been published, which are related in "tools", "theory/concept" and "review study". Also, the SLR result shows that "theory/concept" is the most popular form of the research result. Currently visual technique and text mining is utilized in state-of-the-art, while machine learning tends to be relevant.

By analyzing the study selection result, it can be concluded that there is a very close relationship between the inclusion criteria and exclusion criteria. Other research results also indicate that inadequate protocol can lead to biased selection results (low precision and low recall). Hence, an iterative framework which aims at optimizing the inclusion/exclusion criteria was proposed. The key idea is adding the missing keywords or revising the keywords according to the valuable information retrieved by automation techniques. Text mining is used for retrieving useful information among a huge text data. All the criteria optimizing work should be done before the actual study selection work. Then based on the useful retrieved information, human experts are easy to conduct a criteria-revising process. The process is simple and easy because the retrieved information is very intensively.

The framework is evaluated with a real parallel experiment which compares traditional method and new proposed method. By evaluation, the iterative framework draw a better performance which included more positive papers, with better combined score of precision and recall. To further verify the validity of our framework, a larger sample from a practical SLR case was used, while the second experiment followed the same experiment procedure as the previous one. The validation result further proves that the iterative framework does improve the results of SLR in SE.

The further research is possible to focus on the time indicator in the study selection process. Besides, exploring other methods to automatically optimize the SLR protocol is also a future research orientation.

References

Adeva, J. G., Atxa, J. M., Carrillo, M. U., & Zengotitabengoab, E. A. (2014). *Automatic text classification to support systematic reviews in medicine.* Expert Systems with Applications 41, pp. 1498–1508.

Ananiadou, S., & McNaught, J. (2006). *Text Mining for Biology and Biomedicine.* Boston/London: Artech House. Computational Linguistics.

Ananiadou, S., & Procter, R. (2007). *Supporting Systematic Reviews using Text Mining.* Social Science Computer Review, Volume 27 Number 4.

Bowes, D. (2012). *SLuRp - A tool to help large complex systematic literature reviews deliver calid and rigorous results.* 2nd International Workshop on Evidential Assessment of Software Technologies, EAST 2012, September 22, 2012 - September 22, 2012, Lund, Sweden, Association for Computing Machinery.

Baccianella, S., Esuli, A., & Sebastiani, F. (2013). *Variable-constraint classification and quantification of radiology reports under the ACR Index.* Expert Systems with Applications 40, pp. 3441–3449.

Brereton, P., Kitchenham, B., Budgen, D., Turner, M., & Khalil, M. (2007). *Lessons from applying the systematic literature review process within the software engineering domain.* Journal of Systems and Software, pp. 571–583.

Barrett, N., Weber, J. J., & Thai, V. (2012). *Automated Clinical Coding Using Semantic Atoms and Topology.*Computer-Nased Medical Systems.

Carver, J. C., Hassler, E., Hernandes, E., & Kraft, N. A. (2013). *Identifying Barriers to the Systematic Literature Review Process.* In Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on, pp. 203-212.

Cruzes, D., Basili, V., Shull, F., & Jino, M. (2007). *Automated Information Extraction from Empirical Software Engineering Literature: Is that possible?* First International Symposium on Empirical Software Engineering and Measurement.

Felizardo, K. R. (2012). *A Systematic Mapping on the use of Visual Data Mining to Support the Conduct of Systematic Literature Reviews.* Journal of software, vol. 7, NO, 2.

Felizardo, K. R., & Salleh, N. (2011). *Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews.* 2011 International Symposium on Empirical Software Engineering and Measurement, DOI 10.1109/ESEM.

Felizardo, K. R., Souza, S., & Maldonado, J. C. (2013). *The use of visual text mining to Support the Study Selection Activity in Systematic Literature Reviews: A Replication Study.* Third International Workshop on Replication in Empirical Software Engineering Research.

Fernández-Sáez, A. M., & Bocco, M. C. (2010). *SLR-TOOL: A Tool for Performing Systematic Literature Reviews.* ICSOFT 2010 - 5th International Conference on Software and Data Technologies.

Frantzi, K., Ananiadou, S. & Mima, H. (2000). *Automatic recognition of multi-word terms.* International Journal of Digital Libraries 3(2), pp. 117-132.

Hamad, Z, & Saim, N. (2014). *Systematic literature review (SLR) automation: a systematic literature review.* Journal of Theoretical and Applied Information Technology, Vol. 59 No.3.

Hearst, M. (1999). *Untangling Text Data Mining.* Proceeding ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 3-10.

Hernandes, E. (2012). *Using GQM and TAM to evaluate StArt-a tool that supports Systematic Review.* CLEI Electronic Journal 15.

Hevner, A.R., March, S.T., Park, J, & Ram, S. (2004). Design science in information systems research. MIS Quarterly Vol. 28 No. 1, pp. 75-105.

Kitchenham, B. A. (1997). *Evaluating software engineering methods and tools, part 7: planning feature analysis evaluation.* ACM SIGSOFT Software Engineering Notes, 22(4), 21-24. American Psychological Association.

Kitchenham, B. A., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). *Systematic literature reviews in software engineering–a systematic literature review.* Information and software technology, Vol. 51, no. 1, pp. 7-15.

Kitchenham, B. A., & Brereton, P. (2013). *A systematic review of systematic review process research in software engineering.* Information and Software Technology 55, pp. 2049–2075.

Kitchenham, B. A., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering.* Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University.

Kitchenham, B. A., Dyba, T., & Jorgensen, M. (2004). *Evidence based software engineering. ICSE 2004.* Proceedings. 26th International Conference on Software Engineering pp. 273-281.

Kitchenham, B. A., & Jones, L. (1997). *Evaluating SW Eng.methods and tools, part 8: analysing a feature analysis evaluation.* ACM SIGSOFT Software Engineering Notes, 22(5), pp. 10-12.

Kitchenham, B., Linkman, S., & Law, D. (1997). *DESMET: a methodology for evaluating software engineering methods and tools.* Computing & Control Engineering Journal, 8(3), pp. 120-126.

Lee, A. S. & Baskerville, R. L. (2003). *Generalizing Generalizability in Information Systems Research.* Information Systems Research 14, 3, pp. 221-243.

Malheiros, V. (2007). *A Visual Text Mining approach for Systematic Reviews.* Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on.

Marshall, C., & Brereton, P. (2013). *Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study.* 2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement.

Nakagawa, H., & Mori, T. (2002). *A Simple but Powerful Automatic Term Extraction Method.* Proceeding COMPUTERM '02 COLING-02 on COMPUTERM 2002: second international workshop on computational terminology - Volume 14.

Nunamaker, J. F., Chen, M., & Purdin, T. D. (1990). Systems development in information systems research. Journal of Management Information Systems, pp. 89-106.

Offermann, P., Levina, O., Schonherr, M., & Bub, U. (2009). Outline of a design science research process. Proceeding of the 4th international conference on design science research in information systems and technology article No. 7.

O'Mara-Eves, A., Thomas, J., & McNaught, J. (2015). *Using text mining for study identification in systematic reviews: a systematic review of current approaches.* O'Mara-Eves et al. Systematic Reviews 2015, 4:5.

Poelmans, J., & Kuznetsov, S. (2013). *Formal Concept Analysis in Knowledge Processing: a Survey on Models and Techniques.*

Sager, J. C., Dungworth, D., & McDonald, P. F. (1980). *English Special Languages: principles and practice inscience and technology.* Oscar Brandstetter Verlag KG, Wiesbaden, 1980.

Silva, F. Q., Santos, A. L., Soares, S., Franca, A. C., Monteiro, C. V., & Maciel, F. F. (2011). *Six years of systematic literature reviews in software engineering: An updated tertiary study.* Information and Software Technology 53, pp. 899–913.

Sun, Y., Yang, Y., Zhang, H., Zhang, W., & Wang, Q. (2012). *Towards evidence-based ontology for supporting systematic literature review.* In Proceedings of the 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012), pp. 171-175.

Tomassetti, F. (2011). *Linked data approach for selection process automation in systematic reviews.* Evaluation & Assessment in Software Engineering (EASE 2011), 15th Annual Conference on.

Torres, J. A., Cruzes, D. S., & Nascimento, L. (2012). *Automatic Results Identification in Software Engineering Papers. Is it Possible?* In Proceedings of the 12th International Conference on Computational Science and Its Applications (ICCSA 2012), pp. 108-112.

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). *Systematic review automation technologies.* Tsafnat et al. Systematic Reviews.

Thomas, J., McNaughtb, J., & Ananiadoub, S. (2011). *Applications of text mining within systematic reviews.* Research Synthesis Methods.

Taušan, N., Markkula, J., Kuvaja, P., & Oivo, M (2017). *Choreography in Embedded Systems Domain: A Systematic Literature Review*. Information and Software Technology.

Wohlin, C. (2014). *Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering.* Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering.

Zubidy, A. A., & Carver, J. C. (2014). *Review of Systematic Literature Review Tools.* University of Alabama technical report.

Appendix A. Primary studies

[P01] Ananiadou, S., & Procter, R. (2007). *Supporting Systematic Reviews using Text Mining.* Social Science Computer Review, Vol. 27, No. 4.

[P02] Thomas, J., McNaughtb, J., & Ananiadoub, S. (2011). *Applications of text mining within systematic reviews.* Research Synthesis Methods.

[P03] Felizardo, K. R. (2010). *An Approach Based on Visual Text Mining to Support Categorization and Classification in the Systematic Mapping.* International Conference on Evaluation and Assessment in Software Engineering (EASE).

[P04] Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). *Systematic review automation technologies.* Tsafnat et al. Systematic Reviews.

[P05] Baccianella, S., Esuli, A., & Sebastiani, F. (2013). *Variable-constraint classification and quantification of radiology reports under the ACR Index.* Expert Systems with Applications 40, pp. 3441–3449.

[P06] Barrett, N., Weber, J. J., & Thai, V. (2012). *Automated Clinical Coding Using Semantic Atoms and Topology.*Computer-Nased Medical Systems.

[P07] Felizardo, K., Andery, G., & Paulovich, F. (2012). *A visual analysis approach to validate the selection review of primary studies in systematic reviews.* Information and Software Technology 54, pp. 1079–1091.

[P08] Bowes, D. (2012). *SLuRp - A tool to help large complex systematic literature reviews deliver calid and rigorous results.* 2nd International Workshop on Evidential Assessment of Software Technologies.

[P09] Felizardo, K. R. (2012). *A Systematic Mapping on the use of Visual Data Mining to Support the Conduct of Systematic Literature Reviews.* JOURNAL OF SOFTWARE, VOL. 7, NO. 2, FEBRUARY 2012.

[P10] Felizardo, K. R., & Salleh, N. (2011). *Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews.* 2011 International Symposium on Empirical Software Engineering and Measurement, DOI 10.1109/ESEM.2011.16.

[P11] Felizardo, K. R., Souza, S., & Maldonado, J. C. (2013). *The use of visual text mining to Support the Study Selection Activity in Systematic Literature Reviews: A Replication Study.* Third International Workshop on Replication in Empirical Software Engineering Research.

[P12] Fernández-Sáez, A. M., & Bocco, M. C. (2010). *SLR-TOOL: A Tool for Performing Systematic Literature Reviews.* ICSOFT 2010 - 5th International Conference on Software and Data Technologies.

[P13] Adeva, J. J., & Atxa, J. M. (2014). *Automatic text classification to support systematic reviews in medicine.* Expert Systems with Applications 41, pp. 1498–1508.

[P14] Hamad, Z., & Saim, N. (2014). *Systematic literature review (SLR) automation: a systematic literature review.* Journal of Theoretical and Applied Information Technology, Vol. 59 No.3.

[P15] Kitchenham, B. A., & Brereton, P. (2013). *A systematic review of systematic review process research in software engineering.* Information and Software Technology 55, pp. 2049–2075.

[P16] Marshall, C., & Brereton. P. (2013). *Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study.* 2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement.

[P17] Marshall, C., & Brereton, P. (2014). *Tools to Support Systematic Reviews in Software Engineering: A Feature Analysis.* EASE '14.

[P18] Poelmans, J., & Kuznetsov, S. (2013). *Formal Concept Analysis in Knowledge Processing: a Survey on Models and Techniques.*

[P19] Zubidy, A. A., & Carver, J. (2014). *Review of Systematic Literature Review Tools.* UNIVERSITY OF ALABAMA TECHNICAL REPORT SERG-2014-03.

[P20] O'Mara-Eves, A., Thomas, J., & McNaught, J. (2015). *Using text mining for study identification in systematic reviews: a systematic review of current approaches.* O'Mara-Eves et al. Systematic Reviews 2015, 4:5.

[P21] Tomassetti, F. (2011). *Linked data approach for selection process automation in systematic reviews.* Evaluation & Assessment in Software Engineering (EASE 2011), 15th Annual Conference on.

[P22] Felizardo, K. R. (2011). *Analysing the Use of Graphs to Represent the Results of Systematic Reviews in Software Engineering.* Software Engineering (SBES), 2011 25th Brazilian Symposium on.

[P23] Malheiros, V. (2007). *A Visual Text Mining approach for Systematic Reviews.* Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on.

[P24] Fabbri, S. (2012). *Managing Literature reviews information through visualization.* 14th International Conference on Enterprise Information Systems, ICEIS 2012, June 28, 2012 - July 1, 2012, Wroclaw, Poland, International Conference on Enterprise Information.

[P25] Hernandes, E. (2012). *Using GQM and TAM to evaluate StArt-a tool that supports Systematic Review.* CLEI Electronic Journal 15.

[P26] Hernandes, E. M. (2013). *Experimental studies in software inspection process: A systematic mapping.* 15th International Conference on Enterprise Information Systems, ICEIS 2013, July 4, 2013 - July 7, 2013, Angers, France, SciTePress.

[P27] Cruzes, D., Mendonça, M., Basili, V., Shull, F., & Jino, M. (2007). *Using context distance measurement to analyze results across studies.* In Proceeedings of the First International Symposium on Empirical Software Engineering and Measurement, pp. 235-244.

[P28] Cruzes, D., Mendonça, M., Basili, V., Shull, F., & Jino, M. (2007). *Automated information extraction from empirical software engineering literature: is that possible?* In Proceedings of the First International Symposium on Empirical Software Engineering and Measurement, pp. 491-493.

[P29] Torres, J. A., Cruzes, D. S., & Nascimento, L. (2012). *Automatic Results Identification in Software Engineering Papers. Is it Possible?* In Proceedings of the 12th International Conference on Computational Science and Its Applications (ICCSA 2012), pp. 108-112.

[P30] Sun, Y., Yang, Y., Zhang, H., Zhang, W., & Wang, Q. (2012). *Towards evidence-based ontology for supporting systematic literature review.* In Proceedings of the 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012), pp. 171-175.

[P31] Kitchenham, B. A., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., & Ebrahimi, T. (2012). *Systematic literature reviews in software engineering – A tertiary study.* Information and Software Technology 52, pp. 792–805.

Appendix B. Study Quality assessment

Article ID	Q1	Q2	Q3	Q4	Total Score	Detail
P01	Y	Y	Y	P	3.5	
P02	Y	P	Y	Y	3.5	
P03	Y	Y	Y	Y	4	
P04	-	-	-	-	-	Review study
P05	Y	P	Y	Y	3.5	
P06	Y	Y	P	Y	3.5	
P07	Y	Y	Y	Y	4	
P08	Y	Y	Y	Y	4	
P09	-	-	-	-		Review study
P10	Y	Y	Y	Y	4	
P11	Y	Y	Y	Y	4	
P12	Y	Y	Y	Y	4	
P13	Y	Y	Y	P	3.5	
P14	-	-	-	-	-	Review study
P15	-	-	-	-	-	Review of SLR process
P16	-	-	-	-	-	Review study
P17	-	-	-	-	-	Review study
P18	Y	P	P	P	2.5	
P19	-	-	-	-	-	Review study
P20	-	-	-	-	-	Review study
P21	Y	Y	Y	Y	4	
P22	Y	Y	Y	Y	4	
P23	Y	Y	Y	P	3.5	
P24	Y	Y	P	P	3	

P25	Y	N	Y	P	2.5	
P26	Y	Y	Y	Y	4	
P27	Y	Y	Y	Y	4	
P28	P	N	P	P	1.5	
P29	Y	P	Y	Y	3.5	
P30	Y	Y	N	P	2.5	
P31	Y	P	Y	P	3	