Using Information Visualization and Text Mining to Facilitate the Conduction of Systematic Literature Reviews

Sandra Fabbri[1(✉)], Elis Hernandes[1], Andre Di Thommazo[1,2],
Anderson Belgamo[1,2,3], Augusto Zamboni[1], and Cleiton Silva[1,2]

[1]LaPES-Software Engineering Research Lab, Universidade Federal de São Carlos,
São Carlos, SP, Brazil
{sfabbri, elis_hernandes, augusto_zamboni}@dc.ufscar.br
[2]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo,
São Carlos, SP, Brazil
{andredt, anderson, cleitonrqs}@ifsp.edu.br
[3]Methodist University of Piracicaba, Piracicaba, SP, Brazil

Abstract. Systematic Literature Review (SLR or SR) and Systematic Mapping (SM) are scientific literature review techniques that follow well-defined stages, according to a protocol previously elaborated. Besides systematizing the search for relevant studies, the SR predicts the organization and the analysis of the obtained results. However, the SR application is laborious because there are many steps to be followed. Aiming to offer computational support to SR and SM, the StArt (State of the Art through Systematic Review) tool was developed. Besides helping the steps of SR or SM, the StArt tool has implemented visualization and text mining techniques to support the conduction and the reporting of the SR or SM. A comparative analysis was carried out in relation to StArt and other similar tools.

Keywords: Systematic literature review · Systematic mapping · Tool · Visualization

1 Introduction

The Systematic Literature Review process (SR) has its origins in the medical area and its objective, according to Pai et al. [1], is the creation of a complete and impartial summary about a given research topic following well defined and known procedures. Recently, this process is being adapted to the computer science area, particularly in Software Engineering [2]. Some advantages of the SR usage are the coverage, the replicability and the reliability of its process. Besides systematizing the search for relevant studies, the SR predicts the organization and the analysis of the obtained results. However, the SR process is more laborious than the research conducted on an informal basis [2].

A previous activity to the SR should be the Systematic Mapping (SM) which objective, according to Petersen et al. [3], is to build a classification scheme and to structure a software engineering research area. Like a SR, SM is also a laborious activity and its process is similar to the SR process, with many repetitive steps. One of the main differences between SR and SM is that the desired results of SMs are mainly quantitative and the selected studies cannot be read in full. Despite this fact, quantitative data can also aid the summarization that should be provided by a SR.

Thus, considering that there are several steps to be executed and several documents to be managed, computer support can aid the conformance to the SR and SM processes, enabling higher quality in their execution.

Since 2006, the StArt tool [4] has been developed. In 2008 it was completely restructured and a new version is available [5, 6]. This version provides full support to carry out SRs and visualization and text mining are being added for easing data summarization since, in general, the SR outcome generate a lot of data for transforming into knowledge. As mentioned by Burley [7], information visualization is a valuable tool for knowledge integration activities and, in StArt, such views allow the researcher to find, in a simple way, information on the most important events, the evolution of the research topic by the academic community, and so on. This information is very common in SM.

Another important contribution that has been reached with information visualization in StArt is the evaluation of the search strings quality. An important point in this kind of literature reviews is to find and ensure that the search strings bring all the relevant studies on the research topic. The StArt tool provides visualization of all the studies retrieved as well as their references. Hence, it is possible identifying, for example, if a frequently cited reference was or was not retrieved by the search string.

Based on this context, the objective of this paper is to explore the contributions of information visualization for these kinds of literature reviews. Section 2 presents an overview of StArt functionalities and highlights some features that aid the control of SR and SM processes. Section 3 explains the visualization support provided by StArt and how it can be used to enhance the summarization of the investigated topic. Section 4 presents the support of text mining processing. Section 5 presents a comparative analysis of related tools and Sect. 6 presents the conclusions and future work.

2   An Overview of the StArt

Before explaining how information visualization and text mining processing help on identifying important information for SM and SRs, an overview of the main functionalities of StArt is presented below. As mentioned before, the processes of SR and SM have some repetitive steps and require discipline and systematic practice from the researcher. The information must be registered in an organized way, such that the expected results are reached, the process can be replicable and all the information can be packed.

Thus, StArt has been developed for providing automated support to as many steps as possible. Functionalities to ease data summarization were also implemented in the

tool as the possibility to display data through visualization and Excel formatted reports, according to researcher's needs.

As the SM process is a subset of the SR process, StArt was initially planned to support SRs and currently it is being adapted to also support SMs. Figure 1 illustrates the general process of SR, highlighting what is done with (left side) and without (right side) StArt support. As electronic scientific databases do not allow automated search of primary studies, steps 2, 3 and 4 must be executed without the support of the tool. They are: the adjustment of search strings in search engines, which happens while the protocol is being defined and reviewed; the execution of these search strings after the protocol approval; and the exportation of the search result in a BibTex file, respectively. The step numbers used in this figure will be used in the explanation of the StArt functionalities.

The main functionalities of StArt are presented in the screen shot of Fig. 2. At the left side there is the hierarchical directory tree with the SR process phases. At the right side, the information associated to the functionality selected on the left side is presented.

Shortly, the goals of the three phases are:

- Planning Phase, which consists of the protocol filling (Step 1 of Fig. 1);
- Execution Phase, which is composed of Studies Identification (Steps 2, 3, 4, and 5 of Fig. 1), Selection (Steps 6, 7, and 8 of Fig. 1) and Extraction (Step 9 of Fig. 1). In this phase the researcher should identify the studies, select them and extract the relevant information for answering the research question.
- Summarization Phase (Steps 10 and 11 of Fig. 1), which corresponds to the analysis of the data extracted from each accepted study and the elaboration of a final report describing the state of the art. For this phase, StArt provides graphics, spreadsheets and data visualizations, aiming to make the researcher's tasks easier. Such options will be detailed in Sect. 4.

In the next sections, each phase is discussed in detail, exemplifying the support provided by the StArt tool.

2.1 Planning

In this phase StArt supports the SR Protocol elaboration (Step 1 of Fig. 1) according to the attributes suggested by [8]. Some of the attributes are: research question definition (the research question that the review is intended answer); keywords that will be used for searching for studies; search engines (examples: ACM, IEEE, Scopus); criteria for acceptance or rejection of studies; etc. There is a help message for each protocol attribute aiming to guide its filling. The protocol is stored in the tool and can be accessed and modified if necessary. It is worth noting that, to ensure the SR process conformance, the content of the protocol fields are reflected in later steps of the SR process. For example, when a search engine is chosen during the protocol filling, it is added under the Studies Identification of the Execution Phase, as shown in Fig. 3. Similarly, each attribute inserted in the Information Extraction Form Attributes during
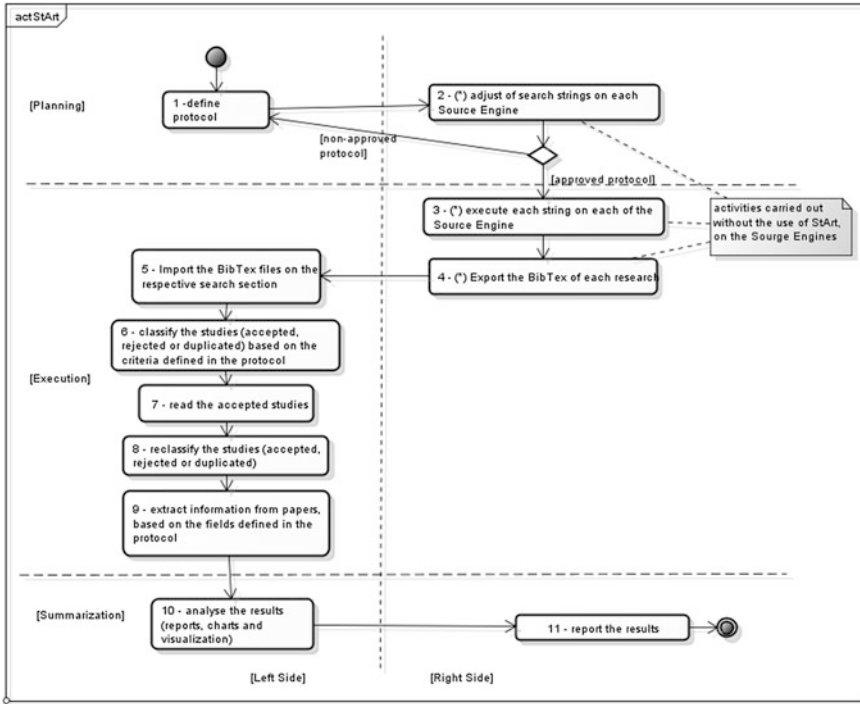
Fig. 1. SLR steps: Left side – actions supported by StArt and Right side – actions not supported by StArt.

the protocol filling becomes a field that must be filled in during the Extraction Step (Step 9 of Fig. 1), as shown in Fig. 4.

2.2   Execution

This phase of the SR has three steps according to the guidelines proposed in [2, 8]. The first one is Studies Identification (Steps 2–5 of Fig. 1). In this step, the researcher should adjust the search string using the keywords earlier defined in the protocol. After this step, the strings should be applied in each search engine, for example, IEEE, Scopus, ACM, Springer and Web of Science. This action is not supported by the tool and the search results must be imported into StArt. As the studies are being imported into the tool, a score is assigned for each one, depending on the keywords defined in the protocol appear in the title, abstract and keywords list of each study. This score can be used, for example, to establish an order of reading once studies with higher scores should be more relevant to the SR. Also, if the studies with higher scores are not relevant to the research question, it is possible that the strings should be revisited and improved. The string definition is an important point to the success of SRs, and its
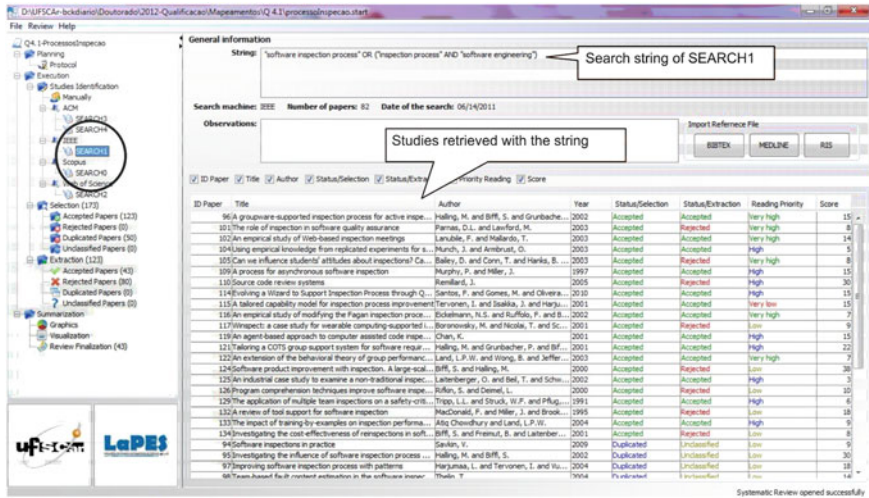
**Fig. 2.** Overview of the StArt tool.

Fig. 3. Search engines defined in the protocol are automatically inserted under Studies Identification.

quality can be accessed through visualization provide by StArt, which is explored and presented in Sect. 4.

The second step is Studies Selection (Step 6 of Fig. 1). In this step, the researcher should use the inclusion and exclusion criteria, defined in the protocol, to classify the studies as accepted or rejected. Duplicated studies are automatically identified by the tool. When the study is accepted, the researcher can attribute to it a relevance level (Very High, High, Low or Very Low).

The third step is Extraction (Steps 7, 8 and 9 of Fig. 2). At this step, the researcher must read the full version of each "Accepted study", elaborate a summary and fill in its Information Extraction Form (Fig. 4B).

Aiming to facilitate this step, it is possible to link the full text file (e.g. PDF files) of each study with its record in the tool.
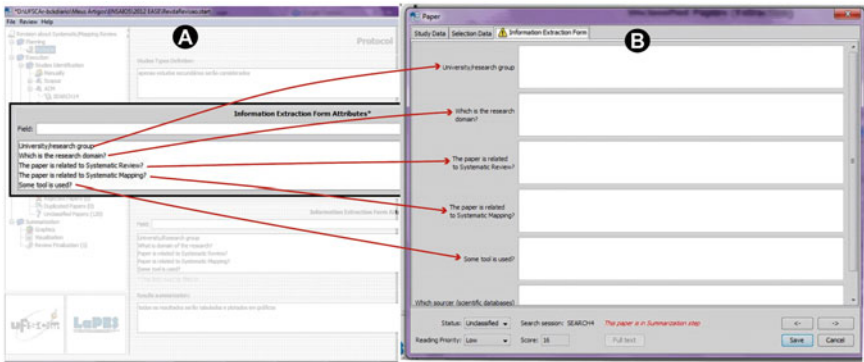
Fig. 4. Relationship between attributes defined in the protocol and the form available during the Extraction step.

2.3 Summarization

In this phase (Step 10 of Fig. 2), StArt provides the following facilities:

- Easy access to the information of all studies accepted in Extraction Step. Comments and information extracted in previous steps can be accessed and copied to a text editor added in the tool. After collecting that information, the researcher can transfer this initial version of the summary to a more powerful text editor.
- Generation of charts that support a quantitative SR characterization. For example: the percentage of studies identified by each search engine, the percentage of studies accepted, rejected and duplicated in Extraction step, the times that each inclusion
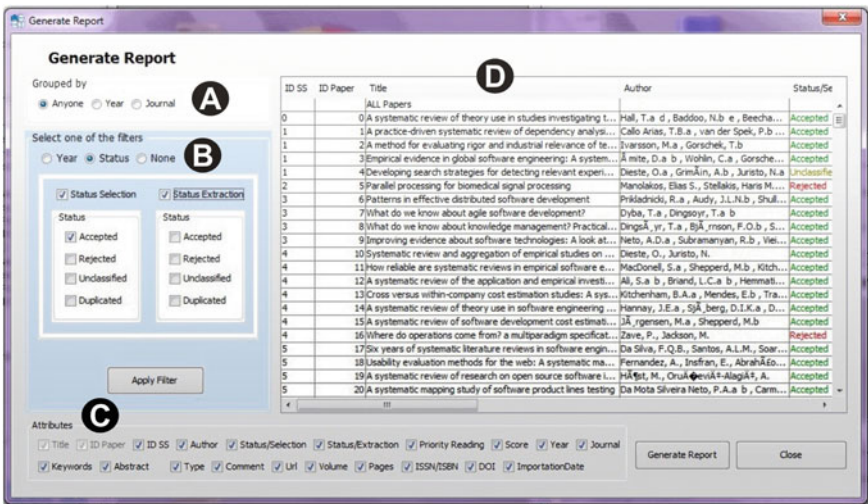
Fig. 5. Options for specifying reports.
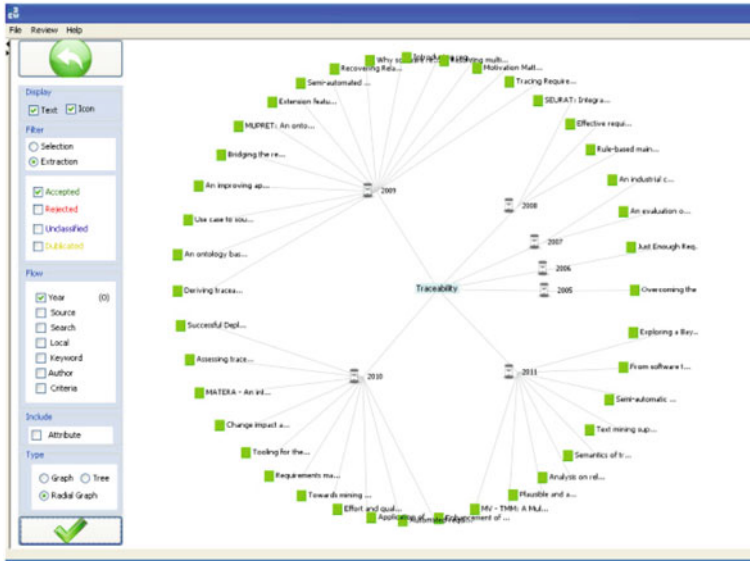
**Fig. 6.** Visualization of publications on "traceability" over time.

and exclusion criterion was used for classifying the studies as accepted or rejected (Fig. 11). In fact, this kind of quantitative data is particularly relevant for Systematic Mappings [3]. In case the researcher choose to do meta-analysis, carry out statistical tests or elaborate other charts, StArt can generate, among other reports, a spreadsheet that allows data manipulation outside the tool. These reports can be generated according to researchers' needs, based on options that allow grouping data in different ways, (Fig. 5A), applying different filters (Fig. 5B) and choosing specific characteristics of the studies (Fig. 5C). Figure 5D shows a preview of the report.

• Deal with a large volume of data to discover features, patterns and hidden trends through visualization. When an SR or SM process is finished, there is a large amount of data related to the research topic that can show trends in the evolution of the topic over time, which is interesting information to explain the state of the art. As mentioned before, the information visualization is a helpful tool for knowledge integration activities.

3   Visualization in StArt

Considering the importance of quantitative data for both the SR and SM and the fact that information visualization explores the natural visual ability of humans aiming to facilitate information processing [9], StArt uses visualization to facilitate knowledge management about literature reviews. Using effective visual interfaces, it is possible to
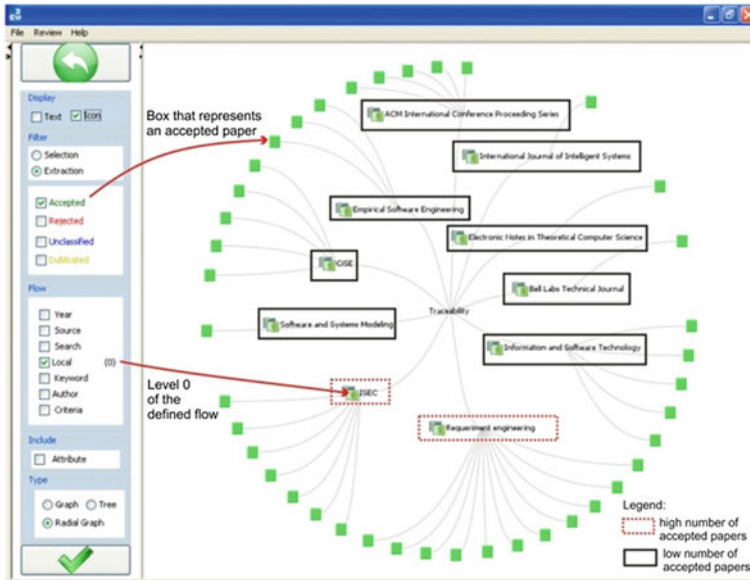
Fig. 7. Visualization of conferences/journals where "traceability" has been published.

quickly manipulate large volumes of data to discover characteristics, patterns and hidden trends.

Based on visualization, for example, it is easier to realize how a specific research topic evolved over time. See Fig. 6 where the researcher's interest was to understand how the topic "traceability" was explored by the academic community, in relation to the question investigated in this example. It is easy to identify that in 2005 and 2006 there was only one published study; in 2007 and 2008 there were few additional studies, but in 2009, 2010 and 2011, the number of studies that mentioned the research topic was more significant than in the previous years.

To build this visualization, the researcher should select the following options (Fig. 6): green rectangle representing an accepted study; part of the study title nearby the rectangle, the publication year as the grouping filter, and the Radial Graph as the visualization technique.

Now, suppose that the researcher would like to identify appropriated places for submitting a study or for publishing results of a literature review. In this case he/she should select almost the same options mentioned before, exchanging year by place. This visualization (Fig. 7), allows identifying the main discussion forums for the topic under investigation. Observe that some places have few studies related to "traceability", while some others have more publications on this topic. Besides, the visualization type was Radial Graph and the studies titles were omitted.

If the researcher wishes to merge both the previous analysis in one graph, it would be better to use a different visualization type. In this case the Tree technique seems
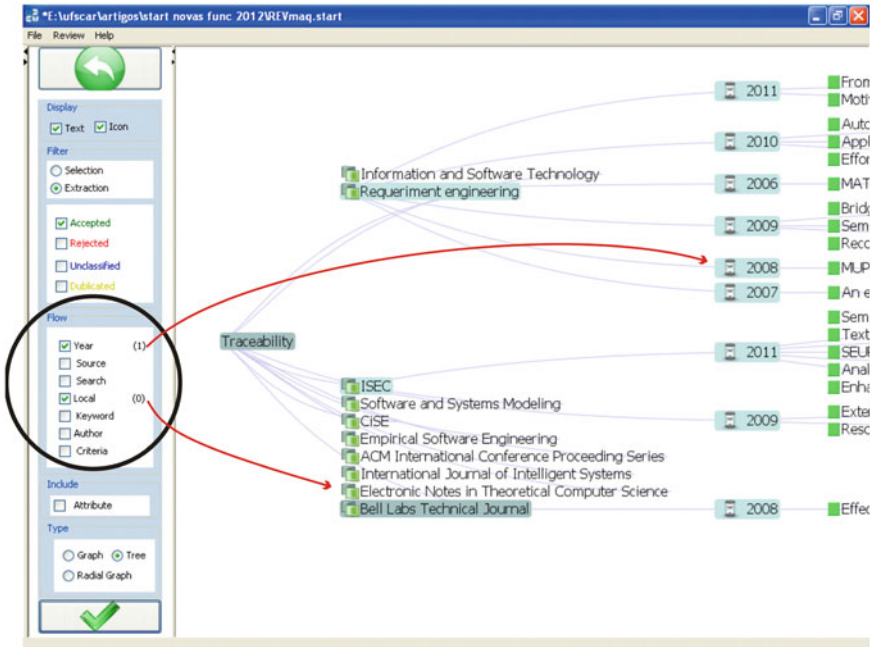
Fig. 8. Visualization of conferences/journals where "traceability" issue has been published according to the year.

better, as shown in the screenshot of Fig. 8. The researcher can expand the levels according to their need.

A double click on a selected study shows information (like authors, abstracts, etc) about it.

In addition to the features described above, visualization is also used to show the relationship among the studies recovered in literature review. This information allows evaluating the set of studies and enhancing the search for them. This resource is better explained in next section.

4   Text Mining in StArt

According to [10], the growing number of publications combined with increasingly cross-disciplinary sources makes it challenging to follow emerging research topics and identify key studies. It is even harder to begin exploring a new field without a starting set of references.

During the conduction of literature reviews many studies are retrieved from various search engines through search strings. Hence, the researcher must be careful not to leave out any studies that may be relevant. According to [11], the usual problem of systematic reviews is that the more inclusive the search strategy, the more irrelevant
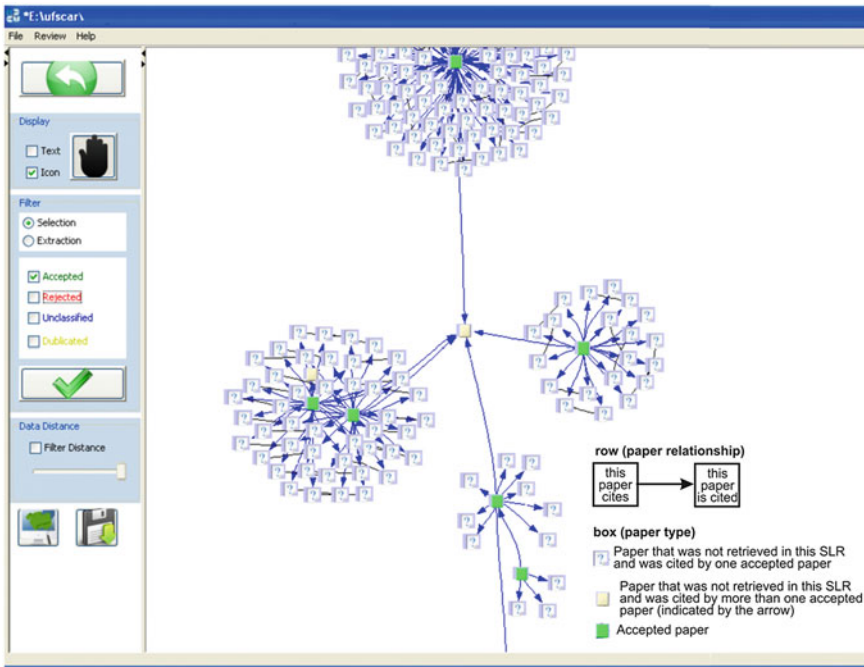
Fig. 9. Relationship among the studies uploaded into StArt and their references.

studies will be retrieved; the more precise and specific the search strategy, the more relevant studies will be missed.

In order to help minimizing this problem, StArt provides support to identify the references of each study retrieved by the search strings. This support allows knowing if there are studies not retrieved, but referenced.

As the search engines generally do not provide the list of references from each study, this information is obtained by reading and extracting the references of the PDF files of the retrieved studies. Every time a PDF file is linked to a study, StArt searches the references in the PDF file. Aiming to identify information like authors, publication place and title, regular expressions are used to identify the bibliographic reference template that was used (APA, Harvard, IEEE, etc.). To determine which study is related to another one, the similarity between the titles of the studies is calculated using the text mining algorithm proposed by [12]. The result of this process is shown through visualization as presented in Fig. 9. The study in the centre of the figure was not retrieved in the literature review, but is referenced by five studies that were retrieved.

This functionality is especially useful during the execution of pilot literature reviews, which should be conducted for adjusting the protocol and the search strings, as suggested by [8]. If there are studies not found but referenced many times, the

Fig. 10. List of similar studies in relation to a selected study – the similarity grade is highlighted.

researcher should verify, for example, if the keywords of these studies should be considered in the protocol and search strings. If so, a new search applying these new keywords must be performed aiming to find relevant studies that were missed.

Start also offers the functionality for detecting which of the studies imported into the tool are similar. The similarity is calculated based on the abstracts through Vector Processing Model [13]. The result of this processing is shown in a table as presented in Fig. 10. This table provides a list of similar studies and their respective similarity grade in relation to a study previously selected.

This list of similar studies can be used, for example: (i) to define the next study to be analyzed; (ii) to facilitate comparison between similar studies and (iii) to make the inclusion and exclusion of studies easier – studies with a high level of similarity to an excluded study tend to be also excluded.

Other researches use text mining in the context of SR or SM, but it is not available in tools that support the whole SR or SM processes.

Malheiros et al. [14] proposed the use of a visualization tool, named PEx, to support the first step of studies selection. PEx has a module that processes the abstract of the primary studies, eliminates stopwords, calculates the terms frequency and, based on this result, displays clusters of studies to facilitate their analysis.

Felizardo [15] continued the previous research and presented the VTM (Visual Text Mining) tool which supports studies selection. As proposed by Malheiros [14], the result of text mining processes is shown by different visualization techniques which help applying the inclusion and exclusion criteria previously inserted in VTM tool.

It is important to notice that the focus of these studies is the studies selection step. On the other hand, in Start, visualization and text mining are currently being used to support the search string definition and the SR or SM Summarization phase.

5   Related Tools

In the literature, there are some tools to support the management of bibliographic references, which are commonly used by researchers to aid in the SLR process. The purpose and the coverage of these tools are different and they are not related to the SLR process proposed by Kitchenham [2], except for SLR Tool [16].

Only SLR Tool [16] focuses on Systematic Literature Review. However, its installation requires the availability of a specific database management system and a pre-configuration of the environment, which can restrict its use, mainly by researchers of other research areas such as Medicine and Nursing, who are also supporters of the SLR process.

Another characteristic of SLR Tool is that it only works with the English and the Spanish versions of the Windows operating system. StArt, on the other hand, can be easily installed, since it has a simplified installer and it can also run with a copy of the executable on the researchers' machines. Table 1 presents related tools that were analyzed and shows some of their features.

Table 1. Features of the StArt and the tools found in the literature.

	Free	Protocol definition	Bibliographic reference manager	Citation export	Attributes customization	Automatic classification of papers	Similar papers identification
JabRef	Yes	No	Yes	Yes	Yes	No	No
EndNote	No	No	Yes	Yes	Yes	No	No
ProCite	No	No	Yes	No	No	No	No
Reference manager	No	No	Yes	No	No	No	No
RefWorks	No	No	Yes	Yes	No	No	No
BibEdt	Yes	No	Yes	No	No	No	No
Zotero	Yes	No	Yes	Yes	Yes	No	No
Biblioscape	No	No	Yes	No	No	No	No
Bookends	Yes	No	Yes	Yes	Yes	No	No
Library master	No	No	Yes	Yes	Yes	No	No
Mendeley	No	No	Yes	Yes	Yes	No	No
Mekentosj	No	No	Yes	Yes	No	No	No
SLR tool	Yes	Yes	Yes	Yes	No	No	No
StArt	Yes	Yes	Yes	No	Yes	Yes	Yes

6   Conclusions

This paper explored the use of visualization for making easier the interpretation of data provided by Systematic Literature Review and Systematic Mapping. This visualization is available in StArt, which also supports the steps of SR and SM processes. As these processes are laborious, posses many repetitive steps and require that all information is packed, the availability of computational support is relevant.

Although there are some tools that have been used by researches to aid the conduction of literature reviews, most of them are reference manager. Some examples are JabRef (jabref.sourceforge.net), EndNote (www.endnote.com), ProCite (www.procite.com), Reference Manager (www.refman.com), RefWorks (www.refworks.com) and Zotero (www.zotero.org). Only SLR tool [16] focuses on SR process [8]. However, it works only on the English or Spanish versions of the Windows operating system.

As StArt is closely associated to the SR and SM processes, it provides many facilities that make easier the conduction of these types of reviews. Some characteristics that differentiate it from the other tools are the score, which is calculated automatically and can give insights on the paper relevance; different types of data visualization that can aid to map the research area; extraction of the references of the studies gathered in the review, that allows evaluating the adequacy of search strings and improving the quality of the whole activity; and other facilities that make the conduction of the process more manageable.

Considering the importance of packing the SRs or SMs data, StArt saves all data in a ".start" file which allows conducting a review in sessions and sharing a review with another researcher. In addition, as StArt provides a simple text editor for writing an initial summary of the state of the art, this summary is also packed. StArt is being continuously evolved and tested. The tool was also evaluated from the perspective of its usefulness and ease of use, according to the TAM model, which found that the tool is useful to users and can be easily used by researchers [6].

As future work, it is planned to continue the development of StArt emphasizing the analysis related to Systematic Mappings. This objective has already initiated with the addition of visualization, but there are other features that can enhance its support for SM. Besides, it is planned some experimental studies that aim to establish a strategy to improve search strings based on the references of the collected studies and also to explore the tool as a support to conduct meta reviews.

References

1. Pai, M., McCulloch, M., Gorman, J.D., Pai, N., Enanoria, W., Kennedy, G., Tharyan, P., Colford Jr, J.M.: Clinical research methods - systematic reviews and meta-analyses: an illustrated, step-by-step guide. Natl Med. J. India **17**, 89–95 (2004)

2. Kitchenham, B.A.: Procedures for performing systematic reviews. Software Engineering Group, Keele University, Keele, Tech. Rep. TR/SE 0401 (2004)
3. Petersen, K., et al.: Systematic mapping studies in software engineering. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, Bari, Italy (2008)
4. Montebelo, R.P., et al.: SRAT (Systematic Review Automatic Tool) Uma Ferramenta Computacional de Apoio à Revisão Sistemática. In: V Experimental Software Engineering Latin American Workshop, ICMC-São Carlos (2007)
5. Zamboni, A.B., Thommazo, A.D., Hernandes, E.C.M., Fabbri, S.C.P.F.: StArt Uma Ferramenta Computacional de Apoio à Revisão Sistemática. In: Brazilian Conference on Software: Theory and Practice - Tools session, UFBA (2010)
6. Hernandes, E.C.M., Zamboni, A.B., Thommazo, A.D., Fabbri, S.C.P.F.: Avaliação da ferramenta StArt utilizando o modelo TAM e o paradigma GQM. In: X Experimental Software Engineering Latin American Workshop, ICMC-São Carlos (2010)
7. Burley, D.: Information visualization as a knowledge integration tool. J. Knowl. Manage. Pract. **11** (2010)
8. Kitchenham, B.A.: Guidelines for performing systematic literature reviews in software. Software Engineering Group, Keele University, Keele, University of Durham, Durham, Tech. Rep. EBSE-2007-01 (2007)
9. Gershon, N., Eick, S.G., Card, S.: Information visualization interactions. ACM Interact. **5**, 9–15 (1998) (ACM Press)
10. Dunne, C., Shneiderman, B., Gove, R., Klavans, J., Dorr, B.: Rapid understanding of scientific paper collections: integrating statistics, text analytics, and visualization. JASIST: J. Am. Soc. Inf. Sci. Technol. **63**, 2351–2369 (2012)
11. Boell, S.K., Cezec-Kecmanovic, D.: Are systematic reviews better, less biased and of higher quality? In: Proceedings of the European Conference on Information Systems, Helsinki, Finland (2010)
12. Salton, G.: Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, Reading (1989)
13. Salton, G., Allan, J.: Text retrieval using the vector processing model. In: Symposium on Document Analysis and Information Retrieval. University of Nevada, Las Vegas (1994)
14. Malheiros, V., Höhn, E., Pinho, R., Mendonça, M., Maldonado, J.C.: A visual text mining approach for systematic reviews. In: International Symposium on Empirical Software Engineering and Measurement, ESEM, pp. 245–254 (2007)
15. Felizardo, K.R., et al.: Using visual text mining to support the study selection activity in systematic literature reviews. In: International Symposium on Empirical Software Engineering and Measurement, ESEM, pp. 77–86 (2011)
16. Fernández-Sáez, A.M., Genero, M., Romero, F.P.: SLR-Tool: a tool for performing systematic literature reviews. In: Proceedings of the JISBD, pp. 329–332 (2010)