An approach based on visual text mining to support categorization and classification in the systematic mapping

Conference Paper · April 2010

CITATIONS
21

READS
140

5 authors, including:

Elisa Yumi Nakagawa
University of São Paulo
143 PUBLICATIONS **775** CITATIONS
SEE PROFILE

Daniel Feitosa
University of Groningen
17 PUBLICATIONS **97** CITATIONS
SEE PROFILE

Rosane Minghim
University of São Paulo
115 PUBLICATIONS **1,303** CITATIONS
SEE PROFILE

José Carlos Maldonado
University of São Paulo
305 PUBLICATIONS **3,285** CITATIONS
SEE PROFILE

An Approach Based on Visual Text Mining to Support Categorization and Classification in the Systematic Mapping

Katia Romero Felizardo
ICMC/Universidade de São Paulo
Caixa Postal 668
13560-970 São Carlos, SP, Brazil
katiarf@icmc.usp.br

Elisa Yumi Nakagawa
Depto. de Sistemas de Computação
ICMC/Universidade de São Paulo
Caixa Postal 668
13560-970 São Carlos, SP, Brazil
elisa@icmc.usp.br

Daniel Feitosa
ICMC/Universidade de São Paulo
Caixa Postal 668
13560-970 São Carlos, SP, Brazil
fdaniel@grad.icmc.usp.br

Rosane Minghim
Depto. de Ciências da Computação
ICMC/Universidade de São Paulo
Caixa Postal 668
13560-970 São Carlos, SP, Brazil
rminghim@icmc.usp.br

José Carlos Maldonado
Depto. de Sistemas de Computação
ICMC/Universidade de São Paulo
Caixa Postal 668
13560-970 São Carlos, SP, Brazil
jcmaldon@icmc.usp.br

Context: Systematic mapping provides an overview of a research area to assess the quantity of evidence existing on a topic of interest. In spite of its relevance, the establishment of consistent categories and classification of primary studies in these categories are manually conducted.
Objective: We propose an approach, named SM-VTM (Systematic Mapping based on Visual Text Mining), to support categorization and classification stages in the systematic mapping using Visual Text Mining (VTM), aiming at reducing time and effort required in this process.
Method: We established SM-VTM, selected a VTM tool and conducted a case study comparing results of two systematic mappings: one performed manually and another using our approach.
Results: The results of both systematic mappings were very similar, showing the viability of SM-VTM. Furthermore, since our approach was applied using a tool, reduction of time and effort can be achieved.
Conclusions: The application of VTM seems to be very relevant in the context of systematic mapping.

Systematic Mapping, Information Visualization, Visual Text Mining

1. INTRODUCTION

Evidence-Based Software Engineering (EBSE) has attracted much attention in recent years. It aims at providing knowledge about when, how, and in what context technologies, processes, methods or tools are more appropriate for software engineering practices (Dybå et al., 2005). Systematic review (Kitchenham and Charters, 2007) and systematic mapping (Petersen et al., 2008) have provided mechanisms to identify and aggregate research evidence. While systematic review has been used to provide a complete and fair evaluation of state of evidence related to a topic of interest, systematic mapping (also known as scoping review) is a more open form of systematic review, providing an overview of a research area to assess the quantity of evidence existing on a topic of interest. Considering its objectives, systematic mapping conducts data extraction and analysis through the identification of categories and classification of primary studies in these categories. Considering its relevance, systematic mapping has been recently applied in different domains, such as software testing (Afzal et al., 2008) and requirement specification techniques (Condori-Fernandez et al., 2009). However, the systematic mapping conduction is not a trivial task, since it requires manual effort and domain knowledge by reviewers in order to achieve adequate results. Furthermore, other difficulties are definition of consistent categories and correct classification of the primary studies in these categories (Budgen et al., 2008).

In another perspective, Text Mining is a process to extract patterns and non-trivial knowledge from textual documents (Tan, 1999). Since data represented in a graphical format can be better understood by people (Oliveira and Levkowitz, 2003), the Visual

Text Mining (VTM) research area provides graphical tools that take advantage of people's visual abilities to support the knowledge acquisition process. Visual Text Mining (VTM) is the association of mining algorithms and information visualization techniques, that allow visualization and interactive exploration of data (Oliveira and Levkowitz, 2003).

We believe that the use of VTM in the systematic mapping process, specifically in the data extraction and analysis, could reduce the effort required. We have also observed that there is a lack of work that investigate application of VTM in the context of systematic mapping. Thus, the main objective of this paper is to propose an approach, named SM-VTM (Systematic Mapping based on Visual Text Mining), that applies VTM to support the categorization and classification in the systematic mapping. Results of our case study have shown that the effort to categorization and classification of the primary studies can be reduced using SM-VTM.

The remainder of this paper is organized as follows. In Section 2, the background about systematic mapping process, about VTM and about other related work are presented. Section 3 presents our approach, as well as a supporting tool, named PEx, that implements this approach. In Section 4, we present a case study. In Section 5, we discuss results, lessons learned and limitations of this work. Finally, conclusions and future directions are presented in Section 6.

2. BACKGROUND AND RELATED WORK

Systematic review has been widely investigated and adopted in EBSE and, more recently, systematic mapping (Budgen et al., 2008) (Petersen et al., 2008) has been indicated when there is a lack of high-quality primary studies (Kitchenham and Charters, 2007), making it possible to obtain an overview of a topic of interest. In general, systematic mapping is conducted by planning, followed by search and screening of primary studies (inclusion/exclusion), similarly to systematic review. However, data extraction and analysis in systematic mapping are conducted in an open form, involving classification of primary studies and categorization of these studies (Budgen et al., 2008). According to Budgen et al. (2008), these stages are not clearly established; however, a more detailed work is present by Petersen et al. (2008). The definition of categories begins with abstract reading (including sections of introduction and conclusion, if necessary) of the selected primary studies. During this reading, keywords and concepts related to the contribution of each primary study are identified. The definition of the categories involves abstractions from individual details of each primary study in order to express a general view of the topic of interest. During the classification of the primary studies into categories, the set of categories can be updated; therefore, categories can be inserted, excluded and merged. At the end, the map is generated representing primary studies classified in the categories. Thus, it is possible to visualize which categories are well covered in terms of number of publications. In spite the relevance of this work, both category identification and primary study classification are not trivial tasks. They are even harder if conducted by non-experts in the domain, such as graduate and undergraduate students (Budgen et al., 2008). Otherwise, specifically for systematic review, a guideline to conduct data extraction and analysis is presented by Kitchenham and Charters (2007).

In another perspective, Knowledge Discovery in Databases (KDD) is the process of extracting high-level, potentially useful knowledge, from low-level data (Keim, 2002). In this context, Data Mining (DM) — that is part of KDD process — has been applied to extract patterns or models from the data. Furthermore, visualization techniques have been combined with DM to help the KDD process. A specific type of combination of visualization and DM techniques is known as Visual Data Mining (VDM) (Keim, 2002), (Oliveira and Levkowitz, 2003). In VDM, visualization supports user interaction with the mining algorithm, directing it towards a suitable solution to a given task. Since text documents are inherently unstructured, Visual Text Mining (VTM), i.e. VDM applied in text, has been focus of specific attention, combining text processin algorithms to interactive visualizations in order to support users making sense of a collection of documents, before deciding which ones to read in detail.

A previous work in our group has explored VTM in EBSE. Malheiros et al. (Malheiros et al., 2007) developed an approach employing VTM to support specifically the selection of primary studies in the systematic review process. The authors compared the reviewers' performances in carrying out the selection of studies just by reading abstracts and their performances when using Malheiros et al.' approach. The attained results have shown that VTM not only sped-up the selection process but also improved its quality, giving support to a more precise selection of relevant studies. Similar to that work, the approach presented here also makes use of VTM techniques and PEx. However, while the former aims at using these techniques to perform the study selection in the process of systematic review, our proposal is focused on using them to specially support the identification of categories and conduction of classification in systematic mapping.

3. SYSTEMATIC MAPPING BASED ON VISUAL TEXT MINING (SM-VTM)

SM-VTM (Systematic Mapping based on Visual Text Mining) is an approach to support categorization

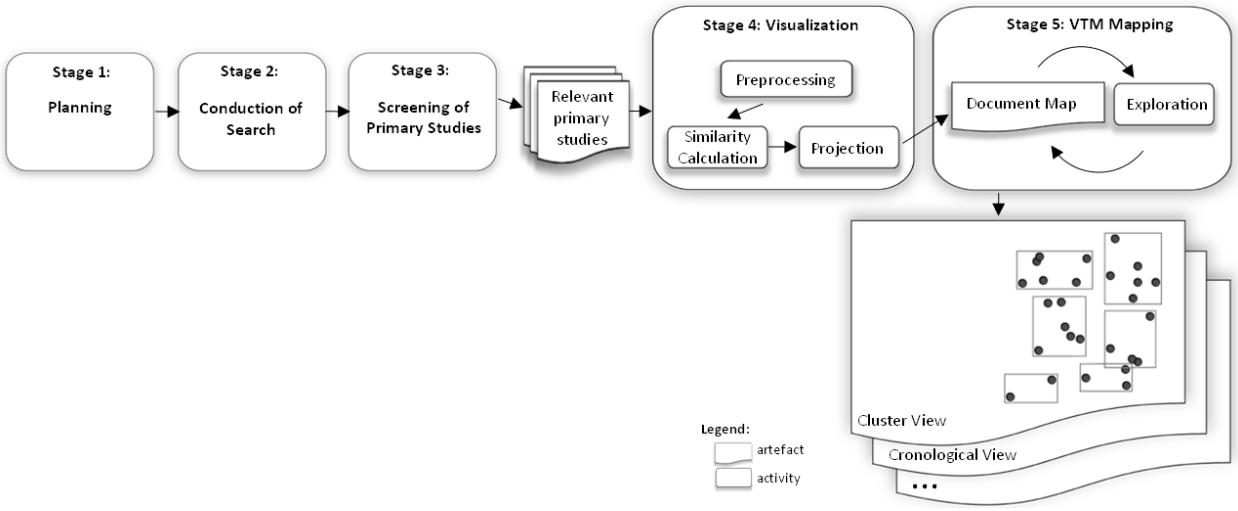


Figure 1: *Systematic mapping based on VTM.*

and classification of primary studies during the systematic mapping process. As illustrated in Figure 1, it is composed by five stages: (i) planning; (ii) conduction of search; (iii) screening of primary studies; (iv) visualization; and (v) VTM mapping. The first three stages are conducted as previously defined by Kitchenham and Charters (2007) and Petersen et al. (2008). In short, in Stage 1, the mapping protocol is defined containing, for instance, the research question, population, source search methods, keywords, paper inclusion and exclusion criteria, and primary study selection process, among others. In Stage 2, primary studies are searched in different sources according to the mapping protocol. In Stage 3, considering the set of primary studies found in the previous stage, the relevant primary studies are selected applying inclusion and exclusion criteria. We have explored application of the VTM Stages 4 and 5, discussed in more details in the following.

3.1. Stage 4: Visualization

This stage aims at generating a visual representation of the previously selected primary studies. For this, three activities are conducted:

- **Preprocessing:** This activity is responsible for structuring and clearing data. For this, it receives as input the set of primary studies selected in the previous stage. In our case, it suffices to employ title, abstract and keywords of an article's content. The preprocessing step converts this input into a vectorial representation after counting number of appearance of a set of selected words (known as *bag of words* (Salton et al., 1975)), extracted from the primary studies. Before frequency count, the high number of terms of the bag of words is reduced by removing little representative terms (known as stopwords), such as prepositions, articles and conjunctions. Additionally, remaining terms are reduced to their radical; for instance, *testing* and *tester* are reduced to *test*. For this, Porter's stemming algorithm (Porter, 1980) was used.

  Following that step, a matrix of *documents x terms* is built, that compounds the collection's vector representation, where columns correspond to selected terms and rows correspond to frequency count of those words for each primary studies. To fill the *documents x terms* matrix, a frequency count takes place, in which for each document the Luhn's cut (Luhn, 1958) is applied. This cut aims at eliminating terms that occur less than *n* times (Luhn's lower cut) or more than *m* times (Luhn's upper cut), where *n* and *m* thresholds are defined by users. It makes it possible to eliminate rare terms (that do not discriminate documents) and terms that are too frequent (that are also not representative). For instance, in the software testing domain, the term "testing", because it is very frequent and common int the document set, it is not representative. The frequency of terms represents the importance of each term in the document. Various forms for scaling the count exist. In our case, the matrix of *documents x terms* is filled with the *term frequency-inverse document frequency measure* (Gennari et al., 1989). In this model, the importance of terms is directly proportional to the frequency of their occurrence in each document, and inversely proportional to the frequency of the term in the collection.

- **Similarity Calculation:** Considering the matrix of *documents x terms*, a measure of the degree of similarity degree among the primary studies is calculated. Common similarity measurements, often used to compare documents in the text mining area, are based Euclidean, Manhattan or

Cosine distances. For example, Cosine calculates the similarity between two vectors of *n* dimensions. For this, it is applied the dot product of the vectors divided by the square root of the product of the vector dot products of each vector. This measure shows the similarity between two vectors, considering a scale of 0.0 (entirely dissimilar) to 1.0 (entirely similar).

- **Projection:** In this activity, multidimensional projection techniques are used to place each primary study on a 2D visual map (i.e. a visual representations of the primary studies) by placing the document onto a plane based on their similarity. For this, a number of known projection techniques can be used (Paulovich et al., 2008). As result of the application of projection techniques, a *document map* is generated. A *document map* is a two-dimensional visual representation presenting a set of documents (in our case primary studies) from with the user can start exploring their content and relationships. Each document is mapped to a graphical element on the plane, usually a circle (point), with points' relative positions reflecting similarity relationships between the contents of the documents they represent. Thereby, on the layout, similar documents are meant to be placed close to one another, while dissimilar ones are supposed to be positioned far apart. The user interacts by locating subgroups of highly related documents iteratively until both global and local relationships are understood.

3.2. Stage 5: VTM Mapping

In this stage, categorization and classification activities are conducted, i.e. categories are defined and primary studies are classified (or distributed) into categories. For this, two exploration strategies are applied:

- **clustering**: this strategy applies clustering algorithms on the *document map* and creates clusters of documents (i.e. regions that concentrate similar primary studies). This strategy is taken as an initial the classification process of primary studies; and
- **topic establishment**: this strategy is used to categorize the clusters. In other words, topics that better represent the primary studies in each cluster are established. For this, two terms that have the highest covariance in the bag of words and are related to documents in each cluster are selected. Then, for each remaining (non-selected) term, it is computed the mean of the covariances between this term and those two terms that have highest covariance previously selected. If this is a significant value relative to the highest covariance (i.e. above a defined percentage threshold that can be defined by users), this term is added as a topic of a cluster (Paulovich et al., 2008).

These strategies can be repeatedly applied in order to explore adequately the set of primary studies contained in the *document map.* The number of interactions is defined by the user, using his or her knowledge about the area. At the end, we have two views:

- **cluster view**: this view presents a set of clusters and related topics. Each cluster contains a subset of primary studies and topics are basis to define categories. By analyzing topics of a cluster, a category is established to the subset of primary studies of that cluster. This view makes it possible to identify evidence gaps clearly (i.e. clusters with low concentration of primary studies) and evidence groups (i.e. clusters with high concentration of primary studies); and
- **chronological view**: this view gives a visual representation of the primary studies indicating their publication year. This representation makes it possible to identify how much the topic of interest has been investigated throughout the years.

Based on the information contained in these two views, systematic maps can be built, if necessary. In the next section, we present a tool that can be used to automatize the Visualization stage and ultimately the VTM Mapping stage of our approach.

3.3. Supporting Tool: PEx

The Projection Explorer (PEx)[1] is a flexible visualization tool that has several text handling facilities, which allows for a VTM exploration of a collection of documents (Paulovich et al., 2007). It is an open source tool that has been developed at the University of São Paulo. It implements different projection techniques and methods to determine similarity among documents as well as visualization and exploration tools. In this paper, we discuss only those functionalities that are important in the context of this work.

Figure 2 shows an example of interaction with the *document map*; a subset of documents was selected in order to show their contents. The icons on the right side make it possible to explore the *document map*, i.e. applying clustering algorithms and topic generation.

In short, PEx provides functionalities to conduct preprocessing, similarity calculation and projection, creating therefore the *document map*. Furthermore, PEx allows user interaction in order to identify categories and classify primary studies. In more detail, PEx presents features to change the visual attributes of points, such as colour to reflect other document properties[2], such

[1] http://infoserver.lcad.icmc.usp.br/infovis2/PEx

[2] Traditionally, Visualization research needs colours in order to enhance information transfer the data under study. Colour version of the figures in this paper are available in http://infoserver.lcad.icmc.usp.br/infovis2/Ease2010

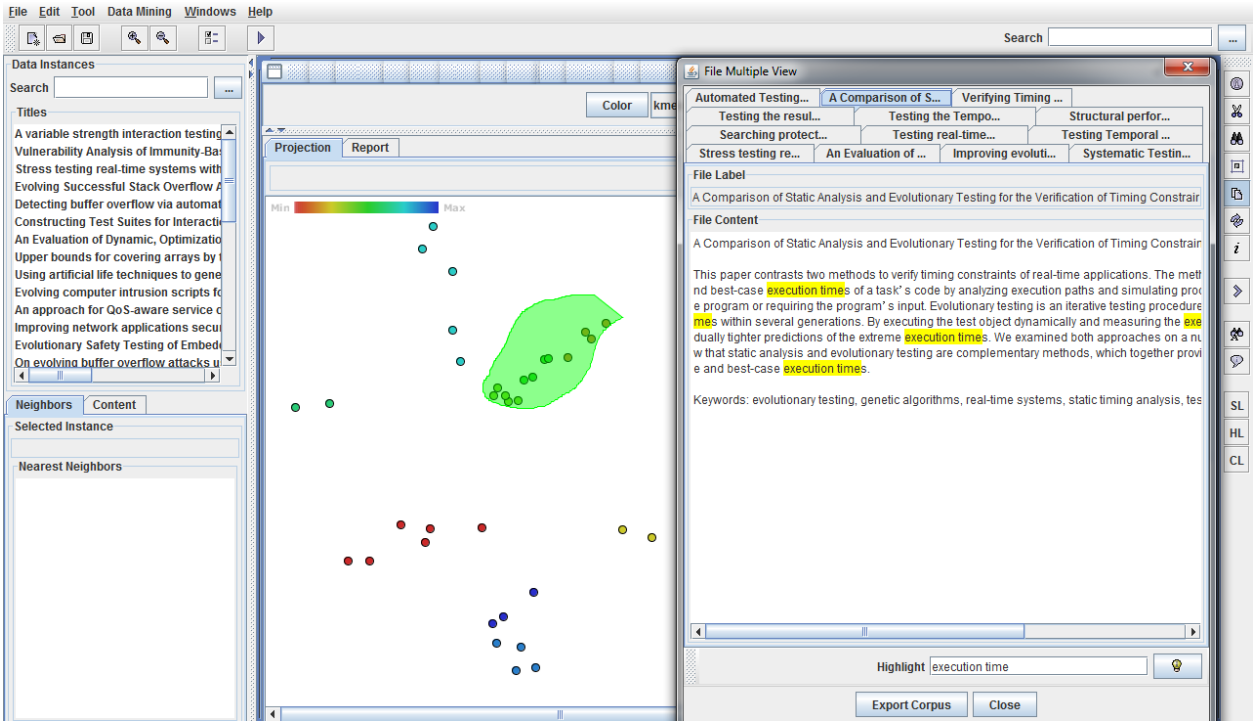


Figure 2: *Projection Explorer: Viewing contents of a subset of documents*

as the publication year of the documents and the occurrence frequency of an expression (words or set of words) in the documents. For instance, in Figure 3, documents that contain a specific expression were coloured based the number of times the expression occurs in the documents. The colour scale varies from red (no occurrence) to blue (many occurrences). In this example, only two documents contain the expression.

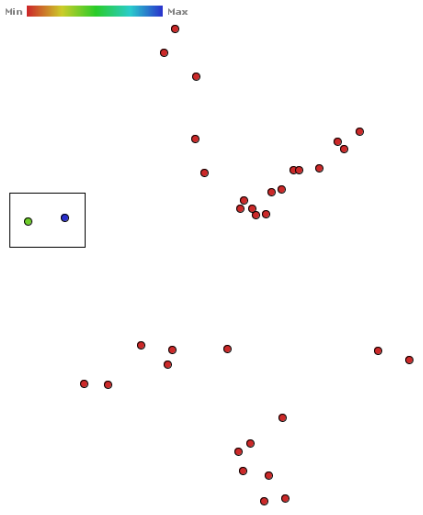


Figure 3: *Documents coloured according to occurrence of a particular expression.*

PEx also generates topics to identify the cluster. Figure 4 illustrates the application of this functionality. For instance, *safety* and *software* are two topics generated for the cluster that contains five primary studies marked as blue points.

In order to illustrate our approach, as well as the use of PEx in this context, we present a case study in the next section.

4. CASE STUDY

In order to demonstrate the use of our approach, a systematic mapping published recently in the literature (Afzal et al., 2008) is used. The systematic mapping conducted by Afzal et al. (2008) mapped Search-Based Software Testing (SBST) area, categorized the primary studies and aimed at identifying which search-based optimization techniques have been applied to non-functional testing. In Figure 5, it is illustrated the result of this systematic mapping. The primary studies identified are distributed between 1996 and 2007. Besides that, the categories established, i.e. the non-functional properties identified were: safety, usability, buffer overflow, quality of service and execution time. It is also possible to observe that the search-based optimization techniques identified were: genetic algorithm (GA), simulated annealing (SA), grammatical evolution (GE), linear genetic programming (LGP), particle swarm optimization (PSO), tabu search (TS), hill climbing (HC) and ant colony (AC).

The first three stages of our approach (planning, conduction of search and screening of primary studies) are similar to Study 1. Therefore we took the result

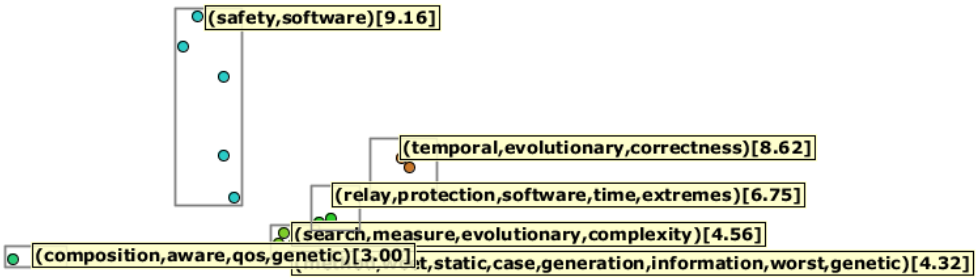


Figure 4: *Most representative topics in each cluster.*

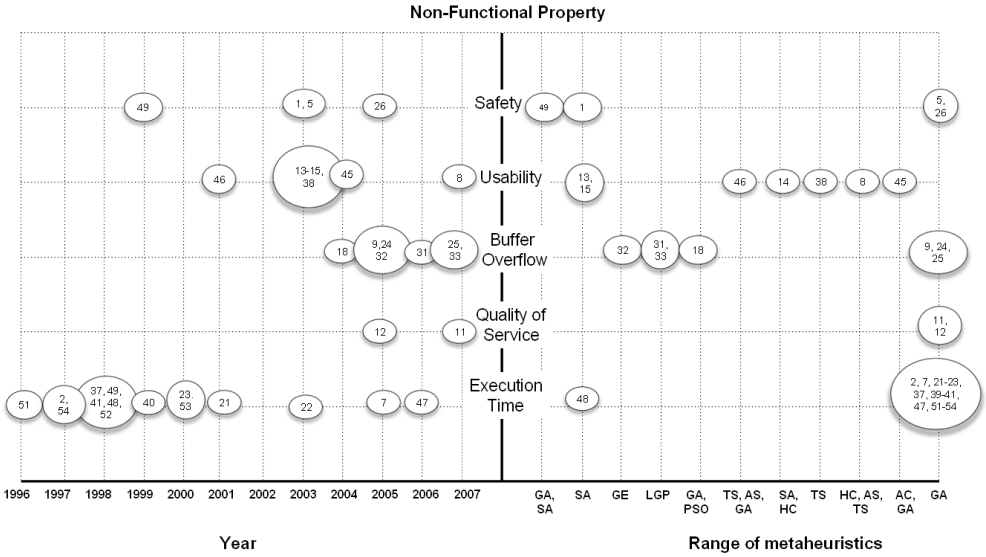


Figure 5: *Systematic mapping on non-functional search-based software testing (Afzal et al., 2008).*

of their third stage (i.e. the same 35 primary studies of Afzal et al. (2008)) to apply in our case study. The other two stages (Stages 4 and 5, visualization and VTM mapping, respectively) that specifically explore VTM in the systematic mapping are presented in more detail in the next sections. In both stages, we used PEx.

It is important to make it clear that the Stages 4 e 5 was conducted apart from the Afzal's work. The results obtained by them were only compared to ours at the end of our study. The study of Afzal et al. (2008) will be named as *Study 1* and ours as *Study 2*.

4.1. Visualization Stage

This stage provides a visual representation of the primary studies. Firstly, we converted each primary study in a document containing the title, abstract and keywords. Following that, PEx conducted the preprocessing that transformed the collection into a *documents x terms* matrix that represents the vector space model of the collection, such as described in Section3.1. Luhns' upper cut for this particular colelction was 100, i.e. terms exceeding 100 occurrences were not considered.

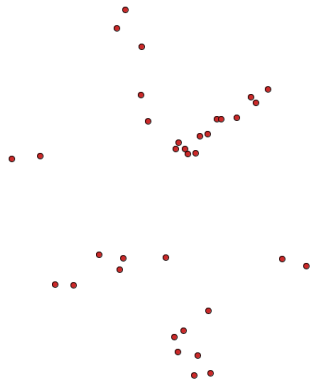


Figure 6: *Document map generated in visualization stage.*

In practice, we provided to PEx a collection of documents as a zipped ASCII file and informed the thresholds of Luhn's cut, the similarity measure and the projection technique. As a result of this stage, PEx generated the *document map* presented in Figure 6. Each point in this map represents a primary study. We can observe regions that concentrate primary studies, indicating similarity among them.

An important point to be emphasized is that the activities undertaken during this stage were completely automated by PEx. Besides that, PEx takes only a few seconds in order to execute these activities and present the *document map*.

4.2. VTM Mapping Stage

During this stage, activities of classification and categorization were carried out. In order to classify the documents (i.e. the primary studies) and establish the clusters, the *k-means algorithm* (MacQueen, 1967) — one of the classical clustering algorithms and also available in PEx — was applied on the *document map*. Using this algorithm, we provided as input the number of clusters that the collection of documents should be classified. We chose initially five clusters, but any other value could be chosen. The result, generated automatically by PEx, can be seen in the Figure 7, that presents the five clusters labeled A to E. Numerical values below each cluster were manually inserted only to illustrate the primary studies contained in each cluster. Comparing this result with the map generated in Study 1 (previously presented in Figure 5), we can observe that cluster A comprises the primary studies related to the *safety* and *quality of service* properties of the Study 1. The combination of clusters B and C coincides with the same primary studies indicated by the Study 1 as studies related to the *execution time*. Finally, clusters D and E combine the primary studies related to the *buffer overflow* and *usability* properties.

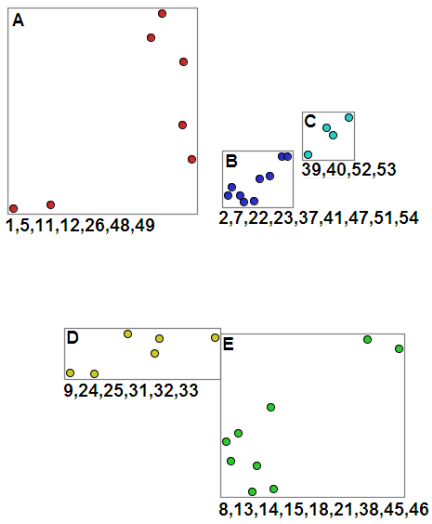
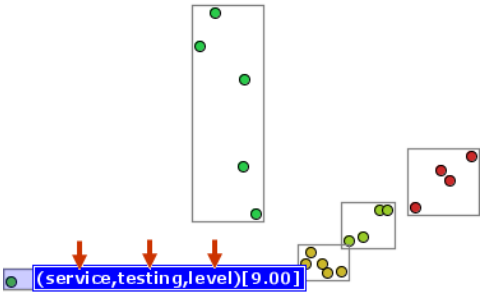
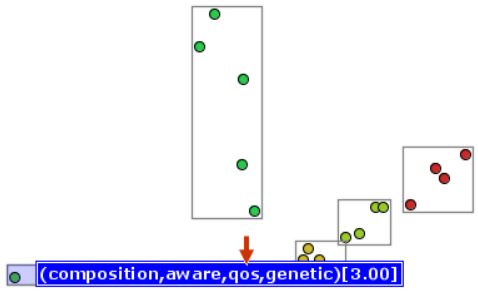


Figure 7: *Document map after application of the clustering algorithm* k-means*, the colour of each point represents the cluster its belong.*

After the generation of clusters, topics were created automatically by PEx, i.e. without user intervention. However, in a first result, the topics were little representative, since general terms, such as *service*, *testing* and *level*, were considered. Figure 8(a) illustrates this situation. Aiming at generating more representative topics, a number of generic terms were inserted in the list of stopwords so that they would not be considered to generate topics. Figure 8(b) illustrates the new topics generated. It is noticed that refined and more representative topics, for example, *qos (quality of service)*, were generated.



(a) Topics initially generated by PEx.



(b) Topics refined, generated after blocking generic terms from the topic extraction process.

Figure 8: *Incremental activity for generation topics in the PEx.*

Besides this first interaction, five other interactions were performed. During the next interactions, the *document map* was explored with 5, 6, 7, 8 and 9 clusters. Additionally 22 terms were manually inserted in the list of stopwords. The terms inserted are: algorithm, algorithms, arrays, attacks, behavior, behavior, computation, computational, computer, data, detection, detector, engineering, level, of, service, system, systems, technique, techniques, test and testing, all of them undiscriminating. This activity required some experience.

At the end, a list of topics was presented for each cluster. Based on these topics, a category can be established for each cluster, i.e. for each subset of primary studies, characterizing one of the main objectives of the systematic mapping. In the next section, we present two views that we propose to the systematic map.

4.3. Results

As main results of application of our approach, two views were generated: *cluster view* and *chronological view*. These views are discussed in more details below and results are compared with Study 1.

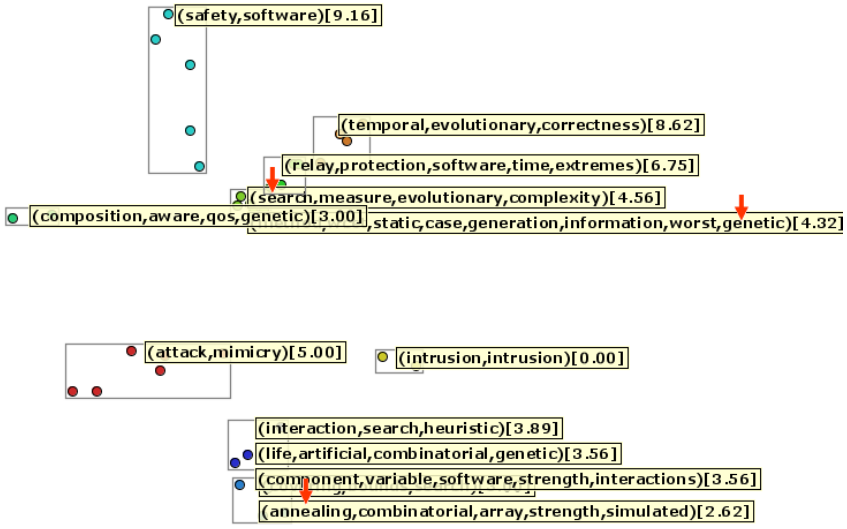


Figure 9: *Cluster view with nine clusters and their respective topics.*

- **Cluster View**

The *cluster view* is the *map document* containing the clusters and their respective topics. It is presented in Figure 9. It is possible to observe that topics, such as *search*, *genetic* and *annealing* (indicated in the figure), are related to the search-based optimization techniques, as also identified in the Study 1.

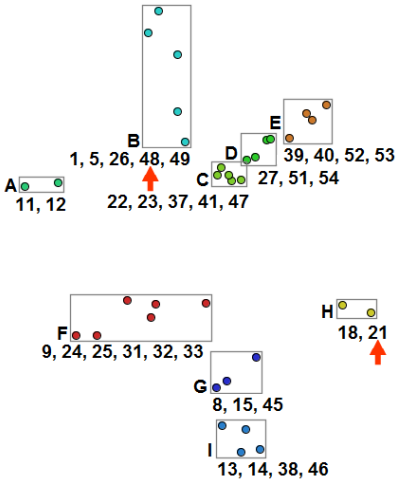


Figure 10: *Cluster view with nine clusters, without the topics.*

To ease the comparison between our results and those of Study 1, the *cluster view* is presented again in Figure 10, however, without the topics. Again, the numerical values below each cluster were manually inserted only to illustrate the primary studies in each cluster. It is possible to observe that the cluster A grouped the primary studies 11 and 12, exactly the same studies classified in the Study 1 as *quality of service*. Cluster B grouped the primary studies classified in the Study 1 as *safety*. The union of the clusters C, D and E corresponds to those classified as *execution time*. The union of clusters F and H corresponds to those classified as *buffer overflow*. Finally, the union of clusters G and I corresponds to primary studies in the Study 1 classified as *usability*. Only the studies numbered as 21 and 48 were not classified in the same way in the studies 1 and 2. In other words, the studies 21 and 48 were classified by us as *buffer overflow* and *safety*, respectively, but both had been classified as *execution time* in Study 1. To facilitate the comparative reading, the scenario discussed above is shown in Table 1 and in Figure 11. This figure was manually edited to show how the categories identified in the Study 1 corresponds to clusters in our *cluster view*.

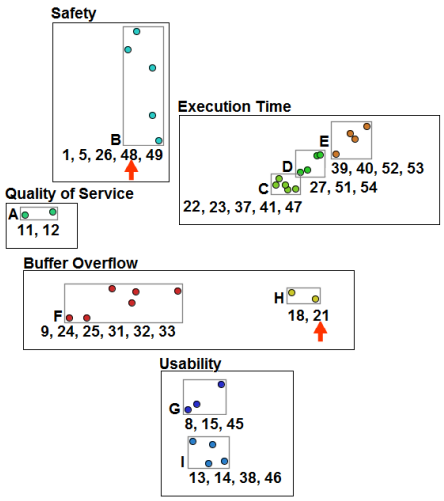


Figure 11: *Cluster view from the perspective of the results of the Study 1.*

In order to verify how the categories (i.e. the non-functional properties) established in Study 1 match with our results, we used *cluster view* to present the occurrence frequency of the terms that correspond to the categories. The *cluster view* was thus coloured, ranging from red (with no occurrence of the term) to blue (the largest number of occurrences of the term). For instance,

Table 1: *Comparative results between studies 1 and 2*

Non-functional property	Study 1: Primary Studies	Study 2: Primary Studies	Study 2: Clusters
Safety	1,5,26,49	1,5,26,48,49	B
Quality of Service	11,12	11,12	A
Execution Time	2,7,21,22,23,37,39,40,41,47,48,51,52,53,54	2,7,22,23,37,39,40,41,47,51,52,53,54	C, D, E
Buffer Overflow	9,18,24,25,31,32,33	9,18,21,24,25,31,32,33	F, H
Usability	8,13,14,15,38,45,46	8,3,14,15,38,45,46	G, I

as mentioned earlier, clusters C, D and E are related to the *execution time* property. We selected the term *execution time* and the result is shown in Figure 12. The primary studies coloured differently from red are really concentrated in the clusters C, D and E. Two studies in cluster B contain also the term, but with low occurrence. Selection of other terms has also pointed to similar results.

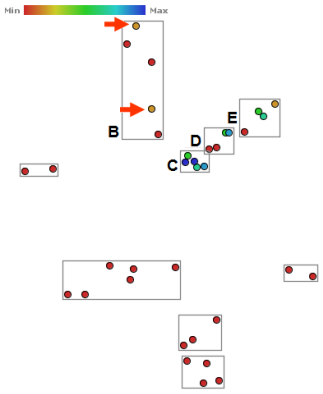


Figure 12: *Primary studies coloured according to the occurrence frequency of the term execution time.*

- ***Chronological View***

Figure 13 presents the *chronological view*. To generate this view, the points contained in the *cluster view* were coloured automatically by PEx to represent the publication year of each primary study. It is possible to observe that the lowest number of primary studies is concentrated in the year 2000 (red points). This view is very useful to identify concentration of primary studies on the topic of interest throughout the years.

5. DISCUSSION

The use of VTM in the systematic mapping process has provided us with feedback about how we can benefit with ability of automating categorization and classification activities in this process. In this section, we discuss some issues related to it, what also includes lessons learned and limitations we have encountered.

The use of VTM in systematic mapping has shown to be very useful in two main perspectives: (i) we have proposed a different point of view to represent a systematic map: *cluster view* e a *chronological view*. Besides providing static information, for instance,

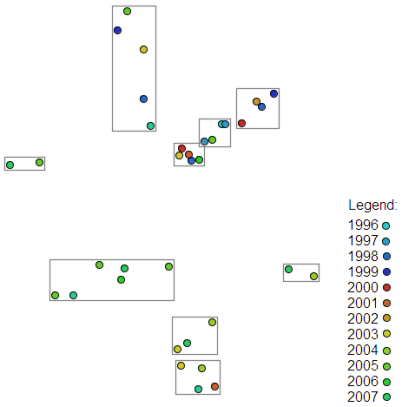


Figure 13: *Chronological view representing the publication year of the primary studies.*

primary studies classified in clusters, these views alow user interaction, since they are supported by an automated tool. Thus, users can explore these views, getting information to build other visual representations of a systematic map, as those presented by Afzal et al. (2008); and (ii) we believe there is enough evidence for the reduction of effort and time in order to conduct categorization and classification activities in systematic mapping if compared with manual conduction. In a few interactions with the tool, none of them requiring actual reading of the documents, we have achieved similar results to a completely manual approach.

A good cohesion of the clusters, i.e. clusters that have a group of documents with high similarity in their contents, is assured by the use of consolidated VTM techniques, such as LSP. In addition, in our approach, categorization of the papers is easier if compared with the manual approach, since an initial suggestion of the terms is provided by PEx. It is important to highlight that there are evidences, as already published in (Lopes et al., 2007) and (Eler et al., 2009), that the topics generated by PEx “translate” indeed the real content of the documents contained in the clusters.

Our approach can be automated by VTM tools, such as PEx, thus, several task, for instance, stemming, similarity calculation and projection, are automatically conducted by the tool. In spite of promising results of SM-VTM, it must still be used in different topics and domains of interest in order to attain further evidence. Furthermore, the use of SM-VTM requires some experience and knowledge in the use of text

mining and visualization tools. It is mainly required knowledge that users understand the output of projection techniques and clustering algorithms and learn how to handle word exclusion for topic determination.

We believe, however, that the evidence that a lot less examination of actual text content is necessary to finish the final mapping is strong suggestion of the future of visual mining for EBSE. Tailoring the tool for this particular purpose will improve user abilities for fast systematic mapping further.

6. CONCLUSION AND FUTURE WORK

Systematic mapping has lately received much attention in SE community; however, in spite of its relevance, in general, it has been manually conducted. Furthermore, category identification and primary study classification are important activities, impacting directly to the quality of results. In this perspective, the main contribution of this paper is to present SM-VTM, a VTM-based approach that supports categorization and classification activities in the systematic mapping. Results of our case study indicate that VTM is an important additional element, since it can contribute considerably with categorization and classification of primary studies. Effort reduction to conduct systematic mapping can be achieved, since our approach is automated using a supporting tool.

We intend to apply SM-VTM in different topics and domains of interest. For instance, we are currently analyzing a case study that aims at identifying an overview about how software engineering activities have been used to develop embedded systems, including robotic systems (Feitosa, 2009). Another research line is to investigate other types of exploration on the *document map*, aiming at getting useful information contained in this map.

Acknowledgments:
This work is supported by Brazilian funding agencies (FAPESP, CNPq and CAPES) and the INCT-SEC Project (Processes: 573963/2008-8 and 08/57870-9).

7. REFERENCES

Afzal, W.; Torkar, R. and Feldt, R. (2008) *A Systematic Mapping Study on Non-Functional Search-Based Software Testing*. In Proc. of SEKE'09, San Francisco, USA, pp.1-3.

Budgen, D.; Turner, M.; Brereton, P. and Kitchenham, B. (2008) *Using Mapping Studies in Software Engineering*. In Proc. of PPIG'08, Lancaster University, UK, pp.195-204.

Condori-Fernandez, N.; Daneva, M.; Sikkel, K.; Wieringa, R.; Dieste, O. and Pastor, O. (2009) *A systematic mapping study on empirical evaluation of software requirements specifications techniques*. In Proc. of ESEM'09, Washington, USA, pp.502-505.

Dybå, T.; Kitchenham, B. and Jørgensen, M. (2005) *Evidence-based Software Engineering for Practitioners*. IEEE Software, 22(1), pp.58-65.

Eler, D. M.; Paulovich, F. V.; Oliveira, M. C. F.; Minghim, R. (2009) Topic-Based Coordination for Visual Analysis of Evolving Document Collections. In Proc. of IV'09, Barcelona, pp.149-155.

Feitosa, D. (2009) *Software Engineering for Robotic System Development: a Systematic Mapping*. Undergraduate dissertation, University of São Paulo, São Carlos, Brazil (in Portuguese).

Gennari, J.; Langley, P. and Fisher, D. *Models of incremental concept formation*. (1989) Artificial Intelligence, 1-3(40), pp.11-61.

Lopes, A. A.; Pinho, R.; Paulovich, F. V.; Minghim, R. (2007) Visual text mining using association rules. Computer and Graphics, 31(3), pp.316-326.

Keim, D. A. (2002) *Information visualization and visual data mining*. IEEE Transactions on Visualization and Computer Graphics, 1(8), pp.1-8.

Kitchenham, B. and Charters, S. (2007) *Guidelines for performing Systematic Literature Reviews in Software Engineering*. Technical Report EBSE 2007-001, Keele University and Durham University.

Luhn, H. P. (1958) *The automatic creation of literature abstracts*. IBM Journal of Research and Development, 2(2), pp.159-165.

MacQueen, J. B. (1967) *Some methods for classification and analysis of multivariate observations*. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp.281-297.

Malheiros, V.; Höhn, E. N.; Pinho, R.; Mendonça, M. and Maldonado, J. C. (2007) *A visual text mining approach for systematic reviews*. In Proc. of ESEM'07, Washington, USA, pp.245-254.

Oliveira, M. C. F and Levkowitz, H. (2003) *From visual data exploration to visual data mining: a survey*. IEEE Transactions Visualization and Computer Graphics, 9(3), pp.378-394.

Paulovich, F. V.; Nonato, L. G.; Minghim, R. and Levkowitz, H. (2008) *Least Squares Projection: a Fast High Precision Multidimensional Projection Technique and its Application to Document Mapping*. IEEE Transactions on Visualization and Computer Graphics, 4(3), pp.364-375.

Paulovich, F.; Oliveira, M. C. F. and Minghim, R. (2007) *The projection explorer: A exible tool for projection-based multidimensional visualization text map explorer: a tool to create and explore document maps*. In Proc. of SIBGRAPI'07, Belo Horizonte, Brazil, pp.27-36.

Petersen, K.; Feldt, R.; Shahid, M. and Mattsson, M. (2008) *Systematic Mapping Studies in Software Engineering*. In Proc. of EASE'08, Italy, pp.1-10.

Porter, M. F. (1980) *An algorithm for suffix striping*. Program, 14(3), pp.130-137.

Salton, G.; Wong, A. and Yang, C. S. (1975) *A vector space model for automatic indexing*. Communications of the ACM, 11(18), pp.613-620.

Tan, A. (1999) *Text mining: the state of the art and the challenges*. In Proc. of PAKDD'08, Beijing, pp.65-70.