A Visual Text Mining approach for Systematic Reviews

Viviane Malheiros[1] Erika Höhn[1] Roberto Pinho[1] Manoel Mendonça[2]

José Carlos Maldonado[1]

[1]Instituto de Ciências Matemáticas e de Computação - USP
Av. Trabalhador Sãocarlense, 400 - Centro
Caixa Postal: 668 - CEP: 13560-970 - São Carlos - SP - Brazil
{viviane, hohn, rpinho,jcmaldon}@icmc.usp.br

[2]Universidade Salvador – NUPERC
Rua Ponciano de Oliveira, 126 - Rio Vermelho.
CEP 41950-275, Salvador, BA, Brazil
mgmn@unifacs.br

Abstract

The software engineering research community has been adopting systematic reviews as an unbiased and fair way to assess a research topic. Despite encouraging early results, a systematic review process can be time consuming and hard to conduct. Thus, tools that help on its planning or execution are needed. This article suggests the use of Visual Text Mining (VTM) to aid systematic reviews. A feasibility study was conducted comparing the proposed approach with a manual process. We observed that VTM can contribute to Systematic Review and we propose a new strategy called VTM-Based Systematic Review.

1. Introduction

Systematic reviews has recently got the attention of researchers from different knowledge domains. This research methodology provides a rigorous, dependable and auditable review methodology. Its main goal is to build a complete and impartial synthesis identifying, assessing and interpreting every available research related to a specific topic.

The software engineering community is assessing the potential of systematic reviews as a bias free and fair research topic evaluation method [6, 5]. Despite promising results, a systematic review might be hard to perform and also time consuming. Thus, supporting tools might be valuable.

This paper introduces the use of Visual Text Mining (VTM) techniques as supporting tools for systematic reviews. We believe that VTM might significantly improve the systematic review process. In order to verify this assumption, a systematic review was performed both with and without the help of VTM supporting tools. We observed that VTM can contribute to Sistematic Review and we propose a new strategy called VTM-Based Systematic Review.

Section 2 briefly outlines the systematic review process. In Section 3, Visual Text Mining is presented. The devised approach for using VTM techniques during a systematic review is presented in Section 4. Section 5 describes the feasibility study performed. Section 6 analyses the results and proposes a new strategy called VTM-Based Systematic Review. Finally, in Section 7 the conclusions are presented.

2. Systematic reviews

Software engineering professionals and researchers find it difficult to make decisions concerning the adoption of new technologies because there are few objective evidences to confirm adequacy, limitations, advantages, associated costs, and, consequently risk for adopting these technologies [3].

Zelkowitz et al. [16] stress that software managers and professionals seek to improve their development processes by adopting new technologies, but without enough evidence that they will be effective, while other options are ignored despite existing evidences of their probable benefits.

Having evidence about effectiveness, cost, adequacy and

0-7695-2886-4/07 $20.00 © 2007 IEEE
DOI 10.1109/ESEM.2007.21

associated risk for available technologies is essential for software engineering professionals and researchers, when they are selecting technologies, or when they are identifying research issues.

Kitchenham et al. [6] suggest that the goal for an evidence based software engineering is to provide the means by which the best current research evidence can be coupled with practical experience and human values into the decision making process regarding software development and maintenance. One should not expect that a given technology should be always good or bad, but only more adequate in some circumstances or organizations. Moreover, professionals are required to gather evidence from experimental research and to assess it from the perspective of their own environment and needs [3].

A systematic review main goal is to condense evidence for a specific question, identifying, evaluating and interpreting every research resource available regarding that particular topic, question or event. Individual articles, which contribute to a systematic review, are called *primary studies*, while the systematic review itself is a *secondary study*.

Secondary goals for a systematic review are (i) critically assess the quality of chosen primary studies, and (ii) identify and verify source of discrepancies on crossed results from the studies [9]. Systematic reviews may also help in identifying new research topics.

Ideally, every study should start with a systematic review and then be built over collected evidence. Kitchenham [6, 5] gives the following reasons for performing a systematic review: (i) summarize existing evidence regarding a treatment or technology, (ii) identify open issues on current research that should be subject of investigation, (iii) provide a framework to identify/classify new research activities, (iv) verify the support of theoretical hypothesis from experimental evidence, and (v) help in the development of new hypothesis.

The difference between a informal review and a systematic review is that the former are qualitative reviews of available evidence. They are usually written by specialists that apply informal and subjective methods to collect and interpret the selected resources. There is a tendency to select sources that support a specific point of view. Moreover, informal reviews frequently do not clearly specify how the reviews selected resources and assessed their quality [9].

On the other hand, a systematic review requires a throughout search for primary studies where their selection is based on reproducible and clear criteria. Selected sources are critically assessed and condensed according to an explicit and previously established method.

Other differences between informal and systematic reviews are [5]: (i) systematic reviews begin by establishing a protocol that specifies the question concerning the review and how to conduct it, (ii) they are based on a predefined search strategy that seeks to find most of the relevant sources, (iii) search efforts are registered so readers can evaluate its validity and completeness, (iv) they have explicit inclusion and exclusion criteria for evaluating each potential primary study, and (v) the information to be extracted from each study is defined along with the quality criteria by which each study is to be evaluated.

3. Visual Text Mining

Knowledge Discovery in Databases (KDD) is the process of extracting high-level, potentially useful knowledge, from low-level data [2]. Data Mining (DM) tasks are responsible for the extraction of patterns or models from the data, being usually the most relevant and time consuming part of the whole KDD process.

Visualization techniques are used along with DM to help the KDD process, supporting cluster and outliers detection, classification, pattern finding or rule extraction. Oliveira and Levkowitz [2] group the use of visualization along with data mining techniques in three approaches: (i) Visual Data Exploration (VDE) for Mining – Handling large data sets typically requires filtering, querying, and selecting data, which can be achieved by VDE, (ii) Visualization of Mining Models – Instead of dealing with the original data set, the Visualization of Mining Models is concerned with end results of a mining process. A relevant example of a technique for visualizing a mining model is the one proposed by Wong et al. for inspecting a set of association rules [15], and (iii) Visual Data Mining (VDM). In VDM techniques, visualization and mining are closely coupled. In this context, the visualization might support the user interaction with the mining algorithm and direct it toward a suitable solution to a given task. For example, mining association rules usually require setting thresholds for support and confidence to limit the size of the resulting set of rules. On a integrated approach, the visualization would help the user to set this parameters. VDM techniques also allow the user to better understand the mining process and its results.

Handling text presents a greater challenge for DM as text documents are inherently unstructured and fuzzy. In VDM for text or Visual Text Mining (VTM), text mining algorithms and methods are combined with interactive visualizations to help the user making sense of a collection of documents, without actually reading all of them.

For applying VTM on a systematic review, we chose the Projection Explorer (PEx) tool. PEx is an evolution of the "Text Map Explorer" [10, 7], a tool that has been developed at the University of So Paulo (USP). PEx is a powerful and highly flexible visualization tool that has several text handling facilities, which allows for a VTM exploration of a collection of documents. On a document map, as built by PEx, documents, which are represented by circles, are placed on a 2D map by content, in a way that documents that are similar tend to be close together, while dissimilar documents are placed apart from each other. The tool provides several methods for determining document similarity and placing them on the 2D map.

In our case study, two methods where used for computing document similarities. For the first method, raw ascii versions of articles were preprocessed to build a vector space model (VSM) of documents from selected terms [12]. The following steps were applied over the original articles in order to remove likely irrelevant terms: (i) stop words (common English words) found on a user supplied list where removed, (ii) remaining terms are reduced to their stem, (iii) a frequency count is applied, and (iv) terms with a frequncy count too high or too low, a defined by user supplied thresholds, are also removed. The document $\times$ term matrix is filled with the *term frequency, inverse document frequency measure* measure[11]. Finally, the similarity of two documents is computed as the cosine distance of their vectors [12]. The second method for computing document similarity works directly over the raw ascii files, requiring no further preprocessing and producing a document $\times$ document distance matrix. It is based on the Normalized Compression Distance (NCD) [14], which is an approximation of the conditional Kolmogorov complexity for a pair of documents. The Kolmogorov complexity for a string or document is the size of the smallest algorithm need to output that string or document. Since its size is not computable, compression algorithms are used as an approximation. The conditional complexity for two strings A and B is the size of smallest algorithm needed to output a string B given the string A as input.

For placing documents on the 2D maps, two projection techniques were used. The first employs the Fastmap [4] projection technique followed by a force based scheme [13]. The second technique, called ProjClus [10], splits documents into clusters, projects each cluster individually, and finally builds the final layout based on a projection of cluster centroids.

The selection of projection techniques and methods for computing document similarities did not follow specific guidelines. Multiple methods were used in order to avoid a bias eventually imposed by a technique.

The PEx version which was used includes a method [7] to display labels that identify topics covered by clusters of documents. These labels are built by extracting and filtering association rules. This feature was central for the exploration strategy devised.

4. A Visual Text Mining approach for Systematic Reviews

In order to explore possible contributions from VTM for performing more effective systematic reviews, a specific research topic was chosen and a case study performed. By measuring the time taken and the number of relevant articles selected, the study could point that VTM is a feasible option for improving systematic reviews.

The chosen topic was to evaluate the effectiveness of applying software test processes for improving the quality of the final product, the software itself. It is important to note that the conclusions drawn from the performed systematic reviews themselves are not to be taken as a final statement about the effectiveness of software testing. Some restrictions were introduced on the initial planning in order to closely follow the intended goal of exploring opportunities of using VTM in the process. A example of such restrictions is the option for limiting the search to IEEE resources, which certainly does not account for a wide and complete search.

For the feasibility study, three researchers conducted systematic reviews on the same topic. Researcher A performed a standard systematic review and the others two researchers conducted the Systematic Review with the VTM approach. One of them was a VTM specialist (henceforward called Researcher B), the other one was a software engineering researcher (henceforward called Researcher C).

Researchers B and C applied the VTM tool for the selection of primary sources. The tool enabled them to group, organize and select articles from the collection of articles retrieved by the search engine. Each researcher identified clusters and regions of interest on presented document maps. These regions were identified by searching for relevant terms or by looking at the neighborhood of known relevant articles.

Simultaneously, the Researcher A read all the abstracts from the collection of articles retrieved by the search engine. Finally, selection results were compared aiming at identifying possible benefits introduced by VTM.

A protocol was established by the three researchers based on the procedures proposed in [5]. It details the systematic review planning, expliciting items like question formularization, population, sources search methods, keywords, papers inclusion and exclusion criteria definition, primary studies selection process. The full version of the protocol can be found in [8].

At the beginning of the search process, few problems arose, so the initial plan was adjusted. The main adjustment was for limiting the sources of primary studies, as one of the researchers was supposed to review all of them with no supporting tool.

5. Systematic review case study

5.1. Collecting data

The search for articles search was done by Researchers A and C together. A spreadsheet was built for registering the articles found: (i) sequential id, (ii) source – IEEE, ACM etc, (iii) title, (iv) authors, (v) publication date, (vi) publication venue – journal, congress etc, (vii) miscellaneous information, (viii) abstract and (ix) references. For the selection and analysis of articles the following columns were added: (i) full text available, (ii) language – English or Portuguese, (iii) refer to test processes, (iv) refer to evaluation of processes or test methods, (v) refer to measures or metrics for test processes, and (vi) refer to test processes application.

Initially, the following sources were used: (i) IEE, (ii) IEEE journals and conferences, (iii) ACM journals and conferences (iv) Google Scholar, (v) SBQS 2006 (acronym in Portuguese to Brazilian Symposium of Software Quality), (vi) SBES 2005 (acronym in Portuguese to Brazilian Symposium of Software Engineering), and (vii) a test specialist, who also support the systematic review planning. No relevant article were found at SBQS 2006 and SBES 2005.

At this point, the researchers had to deal with a few issues:

1. defining keywords for searching – the words initially selected(process, test, quality, evaluation) returned too many results as they are widely used in different contexts. It was noted that the use of more specific terms yielded better results. Another keyword related issue is that the vocabulary seems not to follow a standard or ontology, as happens with medicine, for instance. For example, the expression "test process" was mentioned by the following variations in English: test process, testing process, test activities, testing activities, test procedure, test model, and test practice. Although these expressions are not semantically identical, they were used in studies that concern test processes. When there is no standard vocabulary, an initial search is required to ensure that keywords selected do not bias the processes. From the selected list of keywords, boolean search queries were built. Figure 1 shows the boolean query for English. A Portuguese version was also built.

2. search engines – search engines have different interfaces (query languages) and different algorithms. For each machine, it was built a specific query string, which required the understanding of their query languages and search strategies. The IEEE Periodicals, IEEE Conference Proceedings, IEE Periodicals and IEE Conference Proceedings search engines supported boolean operations directly, and thus the translation of the queries was straightforward. However, a first glance at the results showed that looking for the terms "test" or "testing" apart from "process" and its alternative was including results that were not actually related to test process, so a new version of the query was built where the presence of the expression "test process" or some variation of it was required. Also, on the new version, the requirement for the terms "evaluation", "validation" or "estimate" was scraped as it was excluding known relevant articles. The first version returned 1038 documents, while the revised version returned 250 documents. The revised version for the query used with the IEEE searches is shown in Figure 2.

However, the search on the ACM Digital Libraries was trickier. As it does not support complex boolean queries, a successive filtering approach was tried. This returned 19.781 documents out of a total of 175.083. Searching for just the main keywords with the query "+software +quality +test process +measurement +evaluation", and thus ignoring synonyms, returned just 218 documents.

$$
\begin{gathered}
(software)\ AND\ (test\ OR\ testing)\ AND \\
(measurement\ OR\ metric\ OR\ measure) \\
(software)\ AND\ (test\ OR\ testing) \\
AND (process\ OR\ activity\ OR\ activities \\
OR\ procedure\ OR\ model\ OR\ practice) \\
(software)\ AND\ (test\ OR\ testing)\ AND\ (quality) \\
AND (assurance\ OR\ guarantee\ OR\ process) \\
(software)\ AND\ (test\ OR\ testing)\ AND \\
(measurement\ OR\ metric\ OR\ measure) \\
AND (evaluation\ OR\ valuation\ OR\ estimate)
\end{gathered}
$$

Figure 1. Boolean expression in English.

$$
\begin{gathered}
(software <in> (metadata))\ AND \\
(test\ process\ OR\ testing\ process\ OR \\
test\ activity\ OR\ testing\ activity\ OR \\
test\ activities\ OR\ testing\ activities \\
OR\ test\ procedure\ OR\ testing\ procedure \\
OR\ test\ procedures\ OR\ testing \\
procedures\ OR\ test\ model\ OR\ testing\ model \\
OR\ test\ models\ OR\ testing\ models\ OR \\
test\ practice\ OR\ testing\ practice\ OR \\
test\ practices\ OR\ testing\ practices \\
<in> (metadata))\ AND\ (measurement \\
OR\ metric\ OR\ measure <in> (metadata)) \\
AND\ (quality <in> (metadata))
\end{gathered}
$$

Your search matched **250** of **1342376** documents.

Figure 2. Revised IEEE search query

At this point, it was decided to limit the number of reviewed articles to 100, as it would be enough to the intended goal of comparing both approaches, despite the fact that it would bias the result of the systematic review itself. The list of included articles was then limited to the first 100 articles returned on IEEE that were in English and with full text available. The spreadsheet with these articles listed is available in [8].

5.2. Manual selection

The manual selection followed the systematic review protocol, with Researcher A reading the 100 abstracts and analyzing them according to the inclusion and exclusion criteria.

Researcher A included the articles that were in compliance to "refers to test processes" criterion and in compliance to at least one of the following criteria: (i) "refers to evaluation of processes or test methods"; (ii) "refers to measures or metrics for test processes"; (iii) "refers to test processes application".

Therefore, from the 100 abstracts, 31 were included to the next review phase after Researcher A spent 3 hours in this study selection.

5.3. Article selection with VTM

The first step executed was a brief explanation about the VTM techniques and PEx resources by Researcher B.

The VTM tool was used to support articles selection. With PEx, the chosen articles (100) were explored considering both their full text and parts of them. The tool allows for creating maps based on one version of the text collection and presenting its users another version of it. For instance, document placement on the map could be based on the full text of articles, but the user could only have access their abstracts. As our objective was to asses the full potential of applying a VTM technique, we chose to give researchers B and C access to the full text.

Using PEx demanded a previous step in order to prepare the collected data to an acceptable format. To be explored, the document files were converted from PDF format to raw ASCII text format. The articles spreadsheet was also processed for exploring using text from parts of the articles (title, abstract, references etc). Resulting files where processed in PEx as described in Section 3.

To explore the collection, three exploration strategies were applied to identify relevant documents. Both Researcher B and C agreed about them before starting their selection (third step):

1. Through the VTM tool boolean search, it is possible to change visual characteristics, like document color or size, according to the occurrence of specific terms. Researchers should freely choose relevant terms, including terms not considered in the original query strings. Researchers should identify clusters and regions of interest on presented document maps considering the concentration of highlighted documents;

2. To identify relevant document clusters and regions of interest on presented document maps, considering the relative positions of documents. Small areas with a high concentration of documents usually indicate a very consistent cluster. Documents closely around a document judged as relevant are probably relevant;

3. To discard irrelevant documents or to detect relevant groups through labeled neighborhood inspection. If requested, the VTM tool can cluster related documents and label them with terms considered more significant to represent each cluster. Terms are chosen by extracting and filtering association rules [7].

Once strategies were outlined, a preliminary document map was generated from the spreadsheet (with title, abstract and author). Researcher C conducted a brief exploration to familiarize with the tool, using a Fastmap projection based on NCD distances(See Section 3), and tried to implement the established strategies. During this acknowledgment step, some important discoveries related to data reliability were done, when applying the second strategy(to consider the relative positions of documents):

- There was a data upload fault, implying in abstracts filled with null values. All this articles were placed together;

- There were couple of articles very close together (Figure 3). Reading their abstracts, which is achieved in PEx by double clicking over a document, two reasons for the overlapping were found: (i) there were similar abstracts for different publications; (ii) there was a selection mistake while filling the spreadsheet for one article.

The same exercise was done with a preliminary document map generated from the full text. The first and third strategies were briefly tested too and Researcher C felt comfortable to conduct the study. The data upload fault and spreadsheet were fixed (Researcher A was alerted) and a completely new map was generated to assure that the first impressions from the exercise would not interfere on explorations.

Both Researcher B and Researcher C worked apart. They were free to select the projection technique to be applied (Fastmap or ProjClus) and to take the time they consider necessary to identify the relevant articles. They should use

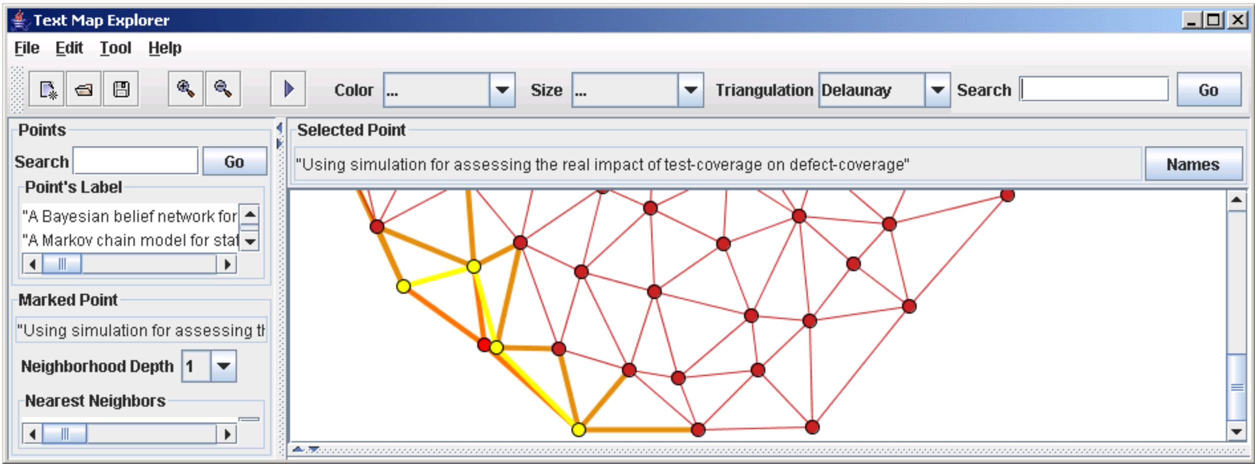

Figure 3. Document map created using FastMap projection, showing a cluster of documents selected for inspection.

the established strategies and the criteria defined in the systematic review plan (see Section 5.2).

Researchers B and C applied the three defined strategies and both of them found the strategies useful on gathering the articles.

The labeled neighborhood inspection was specially interesting to discard documents more than to select documents. Figure 4 shows two examples of discarded groups. PEx automatically assigned the labels "voltage:[voltage]¡-[power], [angle]..." and "psnr:[psnr]¡-[encoders], [streaming]...". These labels should refer to the relevant topics on the clustered documents. These labels were generated in both researchers' explorations and both researcher judged that documents in those groups should be discarded. Before discarding the group, one abstract from the group was read by each researcher. Each researcher confirmed that the group should be excluded. To validate the reliability of this kind of procedure and how it could compromise the systematic review after the selection, these excluded document were listed and the result of their manual analysis by Researcher A was observed. Almost all documents were also excluded through the manual analysis.

The VTM tool boolean search function was specially useful to guide the exploration both new terms and the original query string terms. For example, "Web" was took into consideration by a secondary question at the systematic review protocol, so this term was not in the query string. Searching for "Web" at PEx highlighted a specific group of documents, all of them near to each other (see Figure 5). Identifying this group showed to be an interesting reading technique, because it gave the researcher a context of the documents, turning the reading easier than when it is done completely at random. This strategy (search for a term) was also useful to identify groups documents not grouped under a label or not intensively clustered. Some examples of terms that were applied by at least one researcher, besides "Web" are: "test process", "experimental", "empirical", "measurement" and "evaluation". Some combinations of these terms were used too (one for the color attribute and other for the size attribute). The combinations often seemed to be even more effective to point potential interesting regions.

Researcher B took 51 minutes and included 25 articles, while researcher C took 49 minutes and included 27 articles to the next review phase.

A more detailed comparison between results from Researcher B and C and the comparison between VTM results and manual results (Researcher A) is available in Section 6.

6. Data Analysis

The Venn diagram (see Figure 6) illustrates the quantity of included articles by each researcher and their intersections.

For analysing results of this comparative case study, the researches considered as relevant those articles which were chosen by two or more researchers. These articles were taken as the oracle to the subsequent analysis.

There were 40 articles included by at least two researches (considered as the oracle). From this total, 24 articles is the sum of the articles of each intersection (9 + 7 + 4 + 4, see Figure 6). Researchers B and C revisited the 16 articles included exclusively by Researcher A by manual review. Revisiting these 16 articles, it was possible to identify two groups of articles: (i) a group of articles not even analyzed by researchers B or C; (ii) a group of articles read by these researchers, but discarded by their judg-

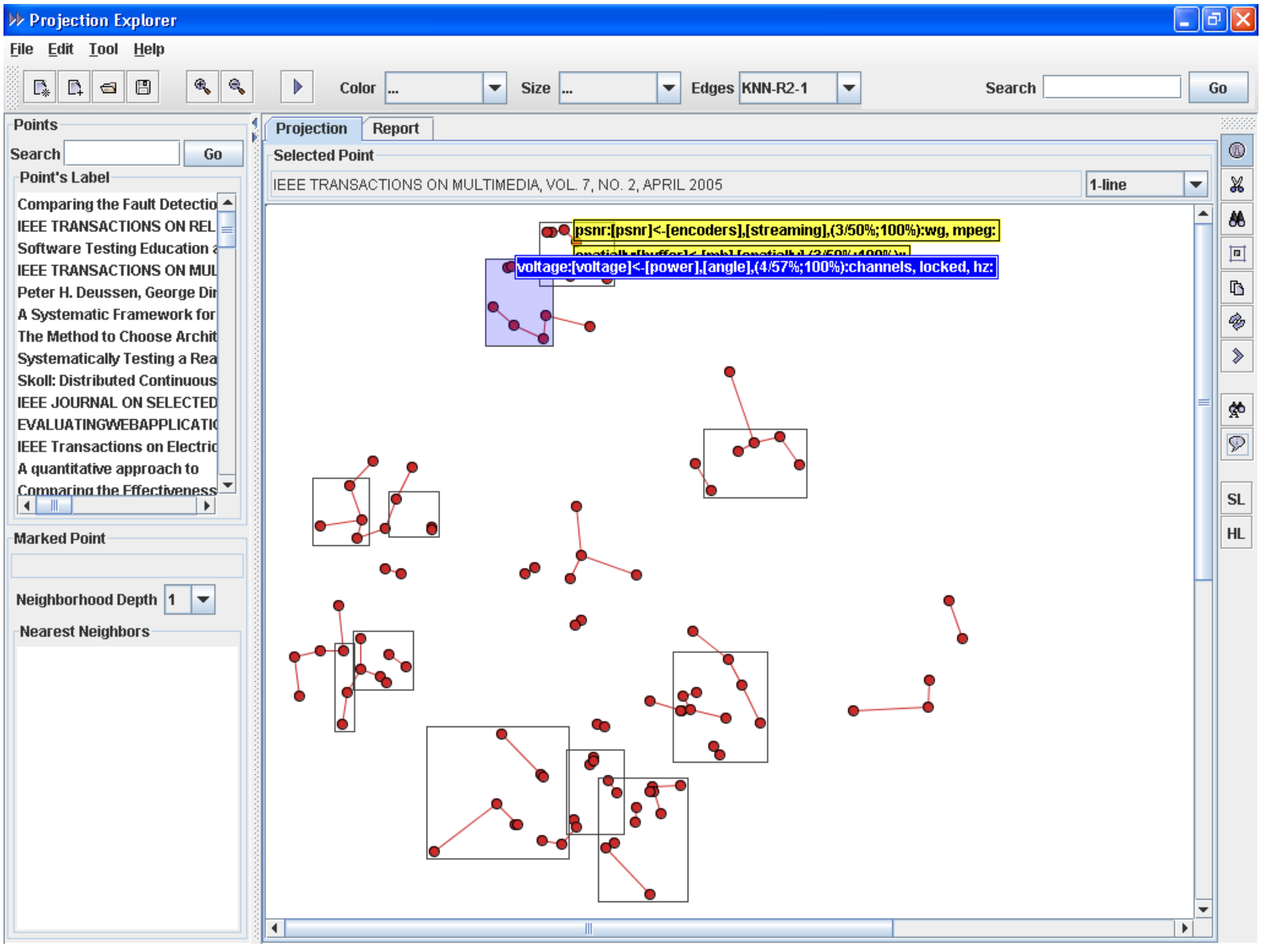

Figure 4. Document map created using ProjClus projection with labeled clusters shown.

ment. From these 16 articles, 11 articles were added to the oracle. Researcher B revisited the 5 articles selected only by Researcher C, including 3 more articles to the oracle. Researcher C revisited the 7 articles selected only by Researcher B, including 2 more articles to the oracle.

Considering the effectiveness of each researcher: (i) Researcher A, using manual review, chose 26 articles (7 + 4 + 4 + 11) of the oracle (40 articles). That is, Researcher A chose 65.00% of the articles included in the oracle, spending 3 hours of review; (ii) Researcher B, applying VTM, chose 22 (9 + 7 + 4 + 2) articles from the oracle, that is, Researcher B chose 55.00% of the total of included articles, spending 49 minutes of review. In similar manner, Researcher C chose 57.50% (9 + 7 + 4 + 3) of the total of included articles, spending 51 minutes of review.

In respect to efficiency of the researches, considering efficiency as $\frac{\textit{chosen and included articles}}{\textit{review time}}$, the results were: Researcher A got 8.67 articles/hour using manual review; Researcher B and Researcher C, applying VTM, got 24.49 articles/hour and 23.53 articles/hour respectively.

Off course that being faster is of no value in a systematic review, if the technique hurts its quality. However, precision, evaluated as the fraction of retrieved articles that were part of the oracle, was not a problem when using VTM. Researcher A had a precision of 83.87% while Researchers B and C had 81.28% and 92% respectively. Researcher B had a lower precision probably due to his lack of knowledge in the domain. Researcher C had a better precision most likely due to her acccess to the full text of articles. If a broader study confirms that precision is not an issue when applying VTM, its higher efficiency allows for the inclusion of more sources. In other words, final results could be more comprehensive, as a researcher would be able to include more document sources. A time limited experiment, where one or more researches would manually evaluate as many articles as possible from a ranked list, while other researches would use the same amount of time exploring the whole collection of articles with VTM could be used to assess this.

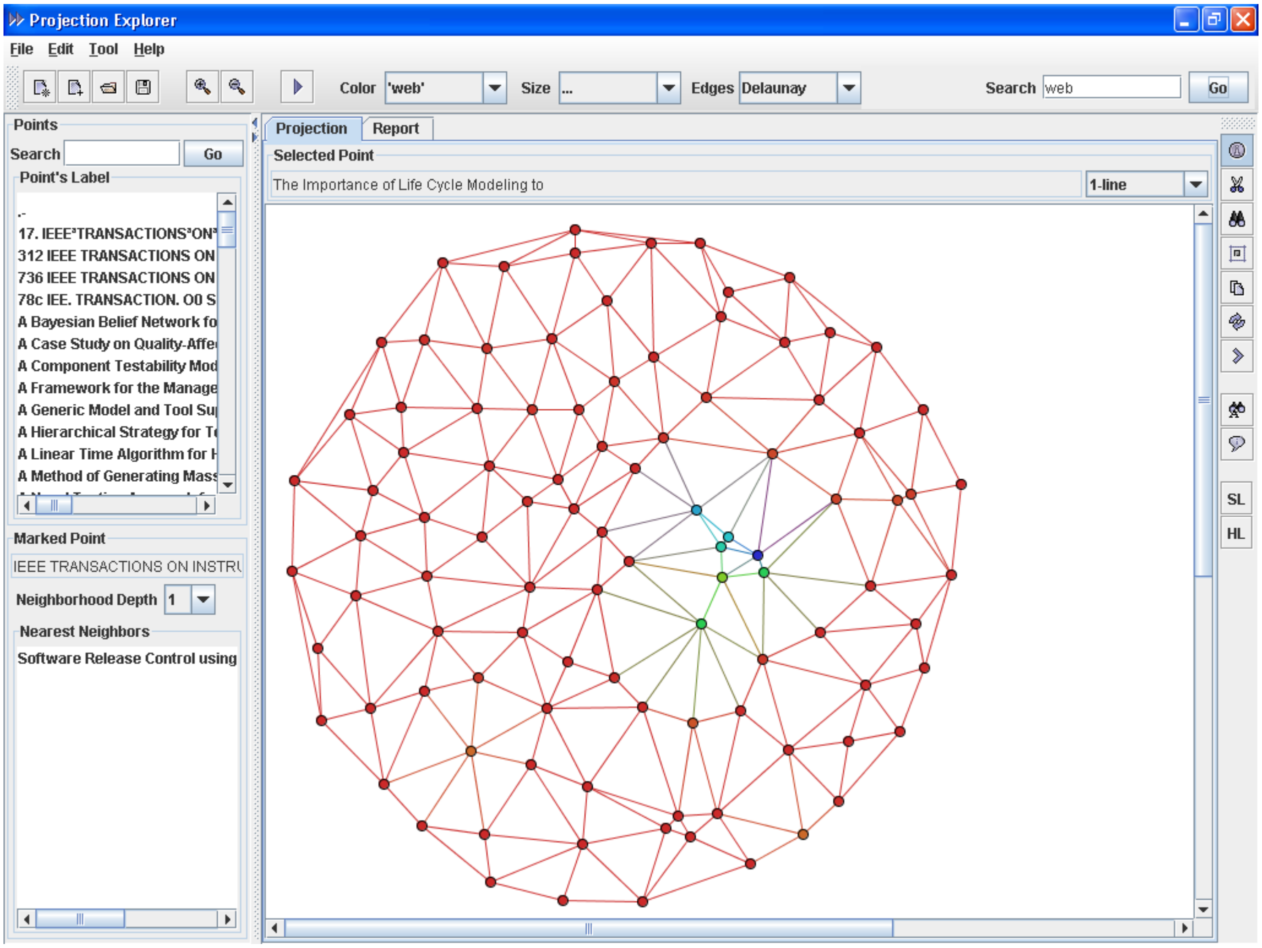

Figure 5. Document map created using FastMap projection. Documents colored according to search for term “Web”.

6.1. Threat of Validity

In this study, we noticed that article evaluation is subjective, either using manual review or applying VTM, possibly due to the absence of a unique ontology in the abstract and full text elaboration. The influence of the researcher judgment could be noticed, for example, by the articles that were considered relevant by Researcher B and C (after selecting in the map and reading the text), but were discarded by Researcher A, who read them all. When specific expressions, defined as criteria (like “testing process”), are not explicitly found, researchers have decided on selecting an article or not based on their own interpretation.

A particular case, is the intersection of the articles selected by Researcher B and C, and not selected by Researcher A. This occurrence can be generated by the subjective interpretation. But we saw that it also could be generated because Researchers B and C had access to the full text of articles, although the full text was not read for every article considered by them. We identified that selecting articles with PEx and considering their full text in cases of doubt improved the selection, as it was observed that at least one relevant article was discarded on the manual review, but included when using the VTM tool. In that example, the abstract poorly reflected the article’s actual content.

6.2. Enhanced VTM-Based Systematic Review

After performing the case study and analysing its results, the group discussed an enhanced strategy that could increase the the benefits of applying VTM for systematic reviews.

The proposed steps are:

1. Selection of one relevant source from the protocol list of sources;

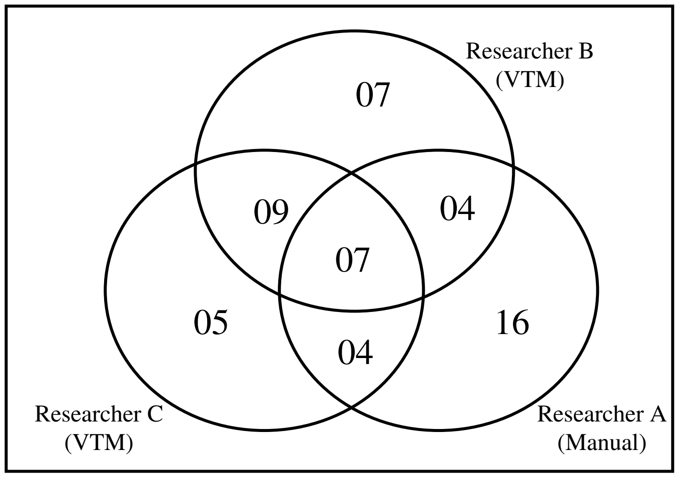

Figure 6. Amount of articles judged relevant by each of the researches.

2. Extraction of N articles according to the article relevance order as given by the search engine from the source selected at Step 1;

3. Manual analysis of the N retrieved articles to identify landmarks (reference articles), according to the protocol inclusion and exclusion criteria;

4. Automatic extraction of articles from all searches engines, following the systematic review protocol;

5. VTM techniques appliance to gather the articles that will be included for the next systematic review phases. At this step, all strategies pointed at Section 5.3 should be applied and, additionally, the landmarks identified in Step 3 should be included. These landmarks should be specially highlight on visualizations, thus providing additional guidence for the user while she/he explores the complete document collection.

A new investigation will be conducted by the authors to validate this enhanced approach and to address the identified threats of validity (Section 6.1).

7. Conclusion

This article shows that VTM could make the systematic review process more effective. During the selection of studies, VTM was valuable for data cleansing. It was easier to spot problems as the erroneous assignment of abstracts to articles or articles that were not imported correctly. Similar articles and almost identical articles were also easily identified.

The use of visualization allowed for more information to be processed at once. It also makes it feasible to consider the full text of articles for selecting the ones that are to be included in the systematic review.

The researcher can also be less strict when determining keywords, and add more synonyms, for example. The increased amount of results can be easily absorbed by the VTM tool. Cluster exploration may help in discarding irrelevant articles. Very tight clusters placed away for other documents can be entirely discarded after the analysis of a few members.

If there are known relevant articles on the collection, perhaps cited by an expert, or chosen by the VTM-based strategy proposed here, the researcher can begin the exploration by its neighborhood. This neighborhood analysis may be also useful to direct the research when he/she decides to actually read some of the abstracts. Articles in the same neighborhood are likely to deal with the same topic. If the researcher reads all abstracts from selected articles in a neighborhood before moving to the next, he/she will probably minimize the need to change context from article to article, and thus minimize his/her cognitive load.

Secondary question can be supported by highlighting specific terms on the document map as supported by VTM tool. For example, by searching for the term “web”, it was easy to locate all articles that had that term, which, by the way, where clustered together (see Figure 5). That neighborhood was scanned to find any article that eventually has that topic but does not have the term “web”.

As the main advantage for the VTM tool is to enable the inclusion of more articles for initial consideration, the strategy of limiting the collection of articles to 100 might have limited our study. A broader experiment was suggested with a time limit imposed on both manual reviewers and to those using the VTM technique. This experiment should include more researchers and more articles. Qualitative evaluation of results could be restricted to consider those articles that were analyzed by all researchers, while quantitative results could be appropriately scaled. Additionally, according to the VTM specialist, the chosen tools work better with more articles.

The choice of the VTM tool, its configuration and the required data preprocessing was not trivial requiring the intervention of the VTM specialist.

It could be argued that the systematic review would not be based on reproducible and clear criteria when using VTM techniques. However, it is important to note that selecting articles after reading their abstracts is also based on a subjective evaluation by the reader. Moreover, as the use of VTM allows for more studies to be included in the initial collection, it is possible to remove artificial restrictions that may have been imposed, as the restriction to search only certain libraries. For instance, even if the researcher decides to read only the abstracts from articles that are on the neighborhood of known relevant articles, this restriction is based on content instead on a criterion chosen at random.

Future developments may include:

1. Apply Visual Data Mining as a decision support tool when analyzing the set of selected articles. Visualization could also be used for presenting final results;

2. Consider other VTM tools for selecting articles according to the number of articles available, their source, structure etc. For instance, a tool that explicitly maps citation relations among articles could be useful [1];

3. Verify if the use of the VTM approach has different impact for different knowledge domains;

4. Validate the proposed enhanced VTM-Based systematic review process (see Section 6.2) with a time-limited experiment.

5. Write a reference guide for applying VTM for systematic reviews, including possible contributions and issues that need attention.

References

[1] C. Chen, L. Thomas, J. Cole, and C. Chennawasin. Representing the semantics of virtual spaces. *Multimedia, IEEE*, 6(2):54–63, 1999.

[2] M. C. F. de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *Visualization and Computer Graphics, IEEE Transactions on*, 9(3):378–394, 2003.

[3] T. Dyba, B. Kitchenham, and M. Jorgensen. Evidence-based software engineering for practitioners. *Software, IEEE*, 22(1):58–65, 2005.

[4] C. Faloutsos and K. Lin. Fastmap: A fast algorithm for indexing, datamining and visualization of traditional and multimedia databases. In *ACM SIGMOD International Conference on Management of Data*, pages 163–174, San Jose-CA, USA, 1995. ACM Press: New York.

[5] B. Kitchenham. Procedures for Performing Systematic Reviews. *Keele University, UK, Technical Report TR/SE-0401-ISSN*, pages 1353–7776, 2004.

[6] B. Kitchenham, T. Dyba, and M. Jorgensen. Evidence-based software engineering. *Software Engineering, 2004. ICSE 2004. Proceedings. 26th International Conference on*, pages 273–281, 2004.

[7] A. A. Lopes, R. Pinho, F. Paulovic, and R. Minghim. Visual text mining using association rules. *Computers & Graphics Special Issue on Visual Analytics*, 2007.

[8] V. Malheiros, E. Höhn, R. Pinho, J. C. Maldonado, and M. M. Neto. Support material for the “a visual text mining approach for systematic reviews” article. `http://www.icmc.usp.br/~rpinho/malheirosetal2007`, 2007.

[9] M. Pai, M. McCulloch, J. Gorman, N. Pai, W. Enanoria, G. Kennedy, P. Tharyan, and J. Colford Jr. Clinical Research Methods. *National Medical Journal Of India*, 17(2), 2004.

[10] F. V. Paulovich and R. Minghim. Text map explorer: a tool to create and explore document maps. In *Information Visualization (IV06)*, pages 245–251, London, 2006. IEEE Computer Society Press.

[11] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24(5):513–523, 1988.

[12] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[13] E. Tejada, R. Minghim, and L. Nonato. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, 2(4):218–231, 2003.

[14] G. P. Telles, R. Minghim, and F. Paulovich. Normalized compression distances for visual analysis of document. *Computers & Graphics Special Issue on Visual Analytics*, 2007.

[15] P. Wong, P. Whitney, and J. Thomas. Visualizing association rules for text mining. *Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on*, pages 120–123, 1999.

[16] M. Zelkowitz, D. Wallace, and D. Binkley. Experimental validation of new software technology. *Software Engineering And Knowledge Engineering*, pages 229–263, 2003.