Do better IR tools improve the accuracy of engineers’ traceability recovery?

Borg, Markus; Pfahl, Dietmar

Published in:
[Host publication title missing]

DOI:
10.1145/2070821.2070825

2011

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Do Better IR Tools Improve the Accuracy of Engineers’ Traceability Recovery?

Markus Borg
Dept. of Computer Science
Lund University
Lund, Sweden
markus.borg@cs.lth.se

Dietmar Pfahl
Dept. of Computer Science
Lund University
Lund, Sweden
dietmar.pfahl@cs.lth.se

1. INTRODUCTION

Development and maintenance of software often result in information overload. Knowledge workers are in general forced to spend more and more time to extract useful information. Maintaining traceability links between software artifacts is one approach to structure the information space of software development projects. Introducing information taxonomies and manually maintaining links is, however, an approach that does not scale very well. As a result, several researchers have proposed to support traceability recovery with tools based on IR methods, utilizing the fact that artifacts often have textual content in natural language.

Traceability recovery tools are generally evaluated by verifying how many suggested links above a certain similarity threshold are correct compared to a hand-crafted set of correct links. The laboratory model of IR evaluation is applied to calculate the measures recall and precision, sometimes extended by a harmonic mean, the F-measure. Recall and precision are commonly reported using graphs, like the one shown in Figure 4. The “X” in Figure 4 marks the recall and precision values of RETRO [13] and ReqSimile [16] for the first ten links proposed by these tools. For “X”, the recall and precision values of RETRO are 50% respectively 60% larger than those of ReqSimile. RETRO is a well-known traceability recovery tool. ReqSimile is a research tool developed at Lund University for the purpose to support information retrieval in the context of market-driven requirements engineering. However, since the real question is to what extent tools like RETRO and ReqSimile actually help engineers in performing traceability recovery tasks, one may wonder whether it is worthwhile to keep hunting for recall-precision improvements of traceability recovery tools.

To tackle this question, we conducted a controlled experiment with 8 subjects to study how tool support affects the tracing process performed by engineers. The purpose of our experiment was to explore how the output from two traceability recovery tools, RETRO and ReqSimile, impacted the engineers’ accuracy of traceability recovery. Hypothesis: The engineers’ accuracy of traceability recovery supported by RETRO differs by more than $\Delta$ compared to that supported by ReqSimile.

ABSTRACT

Large-scale software development generates an ever-growing amount of information. Multiple research groups have proposed using approaches from the domain of information retrieval (IR) to recover traceability. Several enhancement strategies have been initially explored using the laboratory model of IR evaluation for performance assessment. We conducted a pilot experiment using printed candidate lists from the tools RETRO and ReqSimile to investigate how different quality levels of tool output affect the tracing accuracy of engineers. Statistical testing of equivalence, commonly used in medicine, has been conducted to analyze the data. The low number of subjects in this pilot experiment resulted neither in statistically significant equivalence nor difference. While our results are not conclusive, there are indications that it is worthwhile to investigate further into the actual value of improving tool support for semi-automatic traceability recovery. For example, our pilot experiment showed that the effect size of using RETRO versus ReqSimile is of practical significance regarding precision and F-measure. The interpretation of the effect size regarding recall is less clear. The experiment needs to be replicated with more subjects and on varying tasks to draw firm conclusions.

Categories and Subject Descriptors

D.2.1 [Software Engineering]: Requirements/Specifications—Traceability

Keywords

requirements traceability, information retrieval, controlled experiment, equivalence testing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MALETS ’11 November 12, 2011, Lawrence, Kansas, USA
Copyright 2011 ACM 978-1-4503-1022-2/11/11 ...$10.00.
The alternative hypothesis is that the difference between the engineers’ accuracies is smaller than $\Delta$, which implies that the treatments can be considered equivalent. Accuracy is expressed in terms of recall and precision.

2. RELATED WORK

The last decade, several researchers proposed semi-automatic support to the task of traceability recovery. For example, traceability recovery tools have been developed implementing techniques based on algebraic or probabilistic models [1], data mining [23] and machine learning [19]. Several researchers have expressed the tracing task as an IR problem. The query in such a tool is typically the software artifact you want to link to other artifacts. The answer to a query is normally a ranked list of artifact suggestions, most often sorted by the level of textual similarity. The ranked list is analogous to the output of search engines used on the web. Items in the list can be either relevant or irrelevant for the given task.

In 2000, Antoniol et al. did pioneering work on traceability recovery when they used the standard vector space model (VSM) and probabilistic models to suggest links between source code and documentation in natural language [1]. Marcs and Maletic introduced Latent Semantic Indexing (LSI), another vector space approach, to recover traceability in 2003. Their work showed that LSI can achieve good results without the need for stemming, which is fundamental in VSM and the probabilistic models [15]. The same year Spanoudakis et al. used a machine learning approach to establish traceability links. By generating traceability rules from a set of artifacts given by the user, links were derived in the document set. Zhang et al. proposed automatic ontology population for traceability recovery [23]. They developed a text mining system to semantically analyze software documents. Concepts discovered by the system were used to populate an annotation ontology, which was then aligned with a source code ontology to establish traceability links.

Common to those papers is that they have a technical focus and present no or limited evaluations using software engineers solving real tasks. The majority of the published evaluations of traceability tools do not go beyond reporting recall-precision graphs or other measures calculated without human involvement. Exceptions include studies comparing subjects working with tool support to manual control groups. Huffman Hayes et al. developed a traceability recovery tool named RETRO and evaluated it using 30 student subjects [13]. The students were divided into two groups, one working with RETRO and the other working manually. Students working with the tool finished a requirements tracing task faster and with a higher recall than the manual group, the precision however was lower. De Lucia et al. conducted a controlled experiment with 32 students on the usefulness of supported traceability recovery [9]. They found that subjects using their tool completed a task related to tracing various software artifacts faster and more accurately than subjects working manually, i.e. without any support from a dedicated traceability recovery tool. In another study, De Lucia et. al observed 150 students in 17 software development projects and concluded that letting them use IR-based tool support is helpful when maintenance of traceability information is a process requirement [10]. An experiment similar to ours was conducted by Cuddeback et. al., using students and student artifacts [8]. They had 26 subjects

Several researchers proposed ways to obtain better tool output, either by enhancing existing tools implementing standard IR techniques, or by exploring new or combined approaches. Using a thesaurus to deal with synonymy is one proposed enhancement strategy explored by different researchers [18, 12]. Zou et al. investigated term based improvement strategies such as including a part-of-speech tagger to extract key phrases and using a project glossary to weight certain terms higher [24]. Recently, Cland-Huang et al. [6] and Asuncion et al. [2] used a machine learning approach, Latent Dirichlet Allocation, to trace requirements. Furthermore, Chen has done preliminary work on combining IR-methods and text mining in a traceability recovery tool and reported improved results [5].

Even though enhancements lead to better tool outputs in certain cases, their general applicability and the benefit they generate for engineers performing a specific task remain uncertain. Oliveto et al. studied the impact of using four different methods for traceability recovery. In their empirical study, VSM, LSI and the Jensen-Shannon method resulted in almost equivalent results wrt. tracing accuracy [17]. LDA however, while not resulting in better accuracy, was able to capture different features than the others. As far as we know, no studies except Cuddeback et. al. [8], have been published comparing how different quality levels of tool output impact of an engineer in a specific traceability task. If more empirical studies with humans were available, one could conduct a meta-analysis to investigate this matter. Since this is not the case, our approach is instead to compare in an experimental setting the effect of using support tools with differently accurate outputs on traceability tasks performed by humans.

3. EXPERIMENTAL SETUP

This section describes the definition, design and setting of the experiment, following the general guidelines by Wohlin et al. [22]. An overview of our experimental setup is shown in Figure 1.

3.1 Experiment Definition and Context

The goal of the experiment was to study the tool-supported traceability recovery process of engineers, for the purpose of evaluating the impact of traceability recovery tools’ accuracies, with respect to the engineers’ accuracy of traceability recovery, from the perspective of a researcher evaluating whether quality variations between IR tool outputs significantly affect the tracing accuracy of engineers.

3.2 Subjects and Experimental Setting

The experiment was executed at Lund University, Sweden. Eight subjects involved in software engineering research participated in the study. Six subjects were doctoral students, two subjects were senior researchers. Most subjects had industrial experience of software development.

The experiment was conducted in a classroom setting, the subjects worked individually. Each subject was randomly seated and supplied with a laptop with two electronic documents containing the artifacts that were to be traced in PDF format. Each subject also received a printed list per
3.3 Task and Description of the Dataset

It was decided to reuse a publicly available dataset and a task similar to previous tracing experiments to enable comparison to old results. The task, in which traceability recovery was required, was to estimate impact of a change request on the CM-1 dataset. For twelve given requirements, the subjects were asked to identify related requirements on a lower abstraction level. The task was given a realistic scenario involving time pressure, by having the subjects assume the rather unintuitive nature of the dataset, having many unlinked system requirements, the subjects received a hint saying that “Changes to system requirements normally impact zero, one or two design items. Could be more, but more than five would really be exceptional”.

Descriptive statistics of CM-1, including two commonly reported text complexity measures, are presented in Table 1. Farbey proposed calculating Gunning Fog Index as a complexity metric for requirement specifications written in English [11]. The second complexity metric reported is the Flesch Reading Ease, previously reported by Wilson et al. for requirement specifications from NASA [21].

3.4 Description of the Tools

RETRO, developed by Huffman Hayes et al., is a tool that supports software development by tracing textual software engineering artifacts [13]. The tool generates RTMs using standard information retrieval techniques. The evolution of RETRO accelerated when NASA analysts working on independent verification and validation projects showed interest in the tool. The version of the software we used implements VSM with features having term frequency-inverse document frequency weights. Similarities are calculated as the cosine of the angle between feature vectors [3]. Stemming is done as a preprocessing step by default. For stop word removal, an external file must be provided, a feature we did not use. We used the RETRO version V.BETA, Release Date February 23, 2006.

ReqSimile, developed by Natt och Dag et al., is a tool with the primary purpose to provide semi-automatic support to requirements management activities that rely on finding semantically similar artifacts [16]. Examples of such activities are traceability recovery and duplicate detection. The tool was intended to support the dynamic nature of market-driven requirements engineering. ReqSimile also implements VSM and cosine similarities. An important difference to RETRO is the feature weighting; terms are weighted as $1 + \log(freq)$ and no inverse document frequencies are considered. Preprocessing steps in the tool include stop word removal and stemming. We used version 1.2 of ReqSimile.

3.5 Experimental Variables

In the context of the proposed experiment, the independent variable was the quality of the tool output given to the

---

Table 1: Statistics of the CM-1 data, calculated using the Text Content Analyzer on UsingEnglish.com.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>High-level Reqs</th>
<th>Low-level Reqs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>235</td>
<td>220</td>
</tr>
<tr>
<td>Words</td>
<td>5 343</td>
<td>17 448</td>
</tr>
<tr>
<td>Words/Items</td>
<td>22.7</td>
<td>79.3</td>
</tr>
<tr>
<td>Avg. word length</td>
<td>5.2</td>
<td>5.1</td>
</tr>
<tr>
<td>Unique words</td>
<td>1 056</td>
<td>2 314</td>
</tr>
<tr>
<td>Gunning Fog Index</td>
<td>7.5</td>
<td>10.9</td>
</tr>
<tr>
<td>Flesch Reading Ease</td>
<td>67.3</td>
<td>59.6</td>
</tr>
</tbody>
</table>

---

Figure 1: Overview of the experimental setup

Figure 2: Histograms showing the link densities of CM-1 (left) and the subset used as the experimental sample (right).
The null hypothesis was formulated as existence of a difference in the outcomes bigger than $\Delta$. $\Delta$ defines the interval where variation is considered to have no practical value. For this pilot study, we decided to set $\Delta$ to 0.05 for both recall and precision. This means that finishing the task with 0.05 better or worse recall and precision does not have a practical value.

The two one-sided test (TOST) is the most basic form of equivalence testing used to compare two treatments. Confidence intervals for the difference between two treatments must be defined. In a TOST analysis, a $(1 - 2\alpha)100\%$ confidence interval is constructed \cite{20}. We selected $\alpha = 0.05$, thus we reject the null hypotheses that the outcomes of the treatments differ by at least $\Delta$, if the $90\%$ confidence interval for the difference is completely confined within the endpoints $-\Delta$ and $+\Delta$. The $90\%$ confidence intervals are calculated as follows:

$$
\text{point}_{\text{estimate}} \text{outcome}_{\text{RETRO}} - \text{point}_{\text{estimate}} \text{outcome}_{\text{ReqSimile}} \pm 2.353 \sqrt{\text{std dev}^2_{\text{RETRO}} + \text{std dev}^2_{\text{ReqSimile}}}
$$

### 4. RESULTS AND DATA ANALYSIS

The experiment had no dropouts and as a result we collected 8 valid answer sheets and debriefing questionnaires. The answer sets were compared to the gold standard available for the datasets and the corresponding values for recall (Re), precision (Pr) and F-measure (F) were calculated. The descriptive statistics for Re, Pr, F, and the number of requirements traced are presented in Table 2. We also calculated the effect sizes using Cohen’s d (cf. last column in Table 2). Results from the questionnaire are shown in Table 3.

Most subjects experienced the task as challenging and did not have enough time to finish. The list of common acronyms provided to assist the subjects, as was done in a previous case study using the CM-1 dataset \cite{13}, was not considered enough to appropriately understand the domain. Generally, the subjects considered the printed candidate link lists as supportive and would prefer having tool support if performing a similar task in the future.

Table 4 characterizes the tool outputs of RETRO and ReqSimile as well as the tracing results provided by the subjects participating in the experiment. The upper part of the table shows the data for the treatment with RETRO, the lower part that for the treatment with ReqSimile. Each row in the table provides the following data: the ID of the high-level requirement (Req. ID), the number of low-level requirements traced are presented in Table 2. We also calculated the effect sizes using Cohen’s d (cf. last column in Table 2). Results from the questionnaire are shown in Table 3.
Figure 4: Recall-Precision graph for RETRO and ReqSimile for requirements tracing (our sample). The 'X'-symbols mark candidate link lists of length 10. Overall accuracy of answer sets returned by subjects is presented as circles, the diameter represents the relative number of links in the answer set. For a picture where also tool output is presented with relative sizes, see Figure 5.

All subjects apparently proceeded in the order the requirements were presented to them. Since subjects A and E investigated only three and four high-level requirements respectively, they clearly focused on quality rather than coverage. However, the precision of their tracing results does not reflect this focus. The mean recall for subjects supported by RETRO was higher than for subjects supported by ReqSimile, and also the mean precision. The standard deviations were however high, as expected when using few subjects. Not surprisingly, subjects reporting more links in their answer set reached higher recall values.

The debriefing questionnaire was also used to let subjects briefly describe their tracing strategies. Most subjects expressed focus on the top of the candidate lists. One subject reported the strategy of investigating the top 10 suggestions. Two subjects reported comparing similarity values and investigating candidate links until the first "big drop". Two subjects investigated links on the candidate lists until several in a row were clearly incorrect. Only one subject explicitly reported considering links after position 10. This subject investigated the first ten links, then every second until position 20, then every third until the 30th suggestion. This proved to be a time-consuming approach and the resulting answer set was the smallest in the experiment. The strategies explained by the subjects are in line with our expectation that presenting more than 10 candidate links per requirement adds little value.

As Figure 4 shows, a naive strategy of just picking the first one or two candidate links returned by the tools would in most cases result in better accuracy than the subjects achieved. Also, there is a trend that subjects supported by RETRO handed in more accurate answer sets. Pairwise comparison of subjects ordered according to accuracy, i.e. B to E, A to F, C to G, D to H, indicates that the better accuracy of RETRO actually spills over to the subjects' tracing result.

Figure 5 shows relative sizes of answer sets returned by both human subjects and the tools, presenting how the number of tool suggestions grows linearly. The majority of human answer sets contained between one or two links per requirement, comparable to tools generating one or two candidate links.

The 90% confidence intervals of the differences between RETRO and ReqSimile are presented in Figure 6. Since none of the 90% confidence intervals of recall, precision, and F-measure are covered by the interval of equivalence, there is no statistically significant equivalence of the engineers' ac-
Errors of practical significance. For recall, the effect size is too small respectively. Most researchers would consider them as being effect sizes for precision and F-measure are high and medium sizes greater or equal than 0.8 are considered large [7]. The effect sizes greater or equal 0.5 are considered to be of medium size, while effect sizes greater or equal than 0.8 are considered large [7]. The effect sizes for precision and F-measure are high and medium respectively. Most researchers would consider them as being of practical significance. For recall, the effect size is too small to say anything conclusive.

5. THREATS TO VALIDITY

The entire experiment was done during one session, lower-

6. DISCUSSION AND FUTURE WORK

The results of the pilot experiment are inconclusive. The low number of subjects did not allow us to collect strong empirical evidence. The equivalence test (TOST) did not reject similarity and the difference test (two-sided T-test) did not reject similarity. Thus, in this experiment, neither the evidence against equality nor difference was strong enough to reject either null hypothesis.

Although not statistically significant, we could see a trend that subjects supported with the better tool performed more
accurately. Somewhat surprisingly, the precision of the subjects was not higher than that of the results the tools produced. For our specific task, with a time pressure, just using the tool output would generally be better than letting our subjects solve the task, using the tool output as support. One could argue that our subjects actually made the results worse. One direction of future research could be to explore under which circumstances this is the case.

Our experiment is in line with the finding of Cuddeback et al. [8], stating that subjects seem to balance recall and precision. We observed this trend in a very different experimental setup. Foremost, our task was to trace a subset of artifacts under time pressure using printed candidate link lists as support, as opposed to vet a complete RTM without time pressure using a tool. Other differences include: types of subjects and artifacts, and a sparser golden standard RTM.

Is it meaningful to study a tracing task with subjects that are not very knowledgeable in the domain? Embedded software development in the space industry is not easy to simulate in a classroom setting. Yet there is a need to understand the return on investment of improved accuracy of IR tools in support of traceability recovery. Controlled experiments can be one step forward. Our ambition is to replicate this pilot experiment using a larger number of student subjects, to explore whether any statistically significant results appear.


