EVALUATION OF REUSE AND MAINTENANCE IN HYPERMEDIA APPLICATIONS FOR EDUCATION: VALIDATION OF METRICS

Emilia Mendes,
Rachel Harrison and
Wendy Hall

e-mail: {mexm95r,rh,wh}@ecs.soton.ac.uk
fax: 0044 1703 592865

This paper reports the results of applying metrics to hypermedia authoring under the SHAPE research project. The aim of SHAPE is to help authors develop high quality large hypermedia applications for education. The quality characteristics considered are the reusability of information, the maintainability of applications and the authoring effort. Although a number of metrics for hypertext systems have been proposed, we believe that many of the measures proposed in the past lack the necessary mathematical and/or empirical justification. The metrics proposed in this paper have been developed using the Goal-Question-Metric approach, and adhere to the representational theory of measurement. We describe the development of the metrics and the results of a quantitative empirical study which compares two different hypermedia authoring systems.

INTRODUCTION

We regard measurement as important for three basic activities:

1. Understanding the development and maintenance processes.
2. Controlling software projects.
3. Improving processes and products.

Measurement can be used to: i) support project planning; ii) determine the strengths and weaknesses of the current processes and products; iii) provide a rationale for adopting/refining techniques; iv) evaluate the quality of specific processes and
products; v) assess the progress of a project during its course; vi) take corrective action based on this assessment; and finally vii) evaluate the impact of such action [Basili et al., 94].

The literature has plenty of examples of projects whose budgets and schedules overran. Software engineers have addressed software engineering problems by continually looking for new techniques and tools to improve process and product, but methodological improvements which lack corresponding empirical validation cannot be considered scientifically valid [Fenton & Pfleeger, 96].

For anyone who has been involved in software engineering it is clear that for a long time there has been little interest in any sort of evaluation to prove the usefulness of a method or tool, as pointed out by Fenton et al.:

“many research findings published can be characterised as “analytical advocacy research”. That is, the authors describe a new concept in considerable detail, derive its potential benefits analytically, and recommend the concept be transferred to practice. Time passes, and other researchers derive similar conclusions from similar analyses ... Yet practitioners often seem unenthused: something important is missing from this picture: rigorous, quantitative experimentation” [Fenton et al., 94].

The representational theory of measurement seeks to formalise our intuition about the way the world works. That is, the data obtained as measures should represent attributes of the entities observed, and manipulation of the data should preserve relationships observed among the entities. Thus, intuition is the starting point for all measurement.

In Section 2 we present a survey of hypertext metrics already proposed and in Section 3 we compare those proposals and offer further discussion. In Section 4 we present the research project SHAPE and describe how we developed metrics applied to hypermedia authoring. Finally, in Section 5 we give our conclusions and comments on future work.

METRICS APPLIED TO SHAPE

The SHAPE project

SHAPE [Mendes & Hall, 97] is an acronym for a research project, carried out at the University of Southampton, and stands for Southampton Hypermedia Authoring Paradigm for Education. The aim of SHAPE is to aid authors in the development of high quality large hypermedia applications for education. For us, the quality characteristics considered are reusability of information, maintainability of applications and authoring effort.

Instead of defining improvements to be applied to an authoring tool and later verifying if they are adequate we decided to use a more consistent and systematic
approach, which is to apply metrics in order to identify how adequate an authoring tool is for the maintainability of applications, information reuse in applications and the level of authoring effort required.

The principles of the metrics we developed are based on Fenton et al.’s framework for software measurement [Fenton & Pfleeger, 96], and on the guidelines from the DESMET project [Kitchenham, 96], [Kitchenham, 93], in the field of software engineering. Both have been extensively used in experiments in the software engineering field [Harrison et al., 95], [Daly, 96], [Briand et al., 96], [Briand et al., 97], [Basili & Rombach, 88], [Basili et al., 94], [McDonell, 91].

We have planned two evaluations for SHAPE. The first was a quantitative evaluation and the second will be both quantitative and qualitative. In Section 4.2 we describe and present the results of the first evaluation.

**Design of the First Evaluation**

For the first evaluation the stated hypotheses were:

H1-0: Microcosm applications are more maintainable and their information more reusable than applications built using a standard WWW environment

H1-L: The use of a link server allows both a better maintainability of applications and reusability of information than embedded ones

H1-G: Generic links allow a better maintainability and reusability of information than buttons

We have chosen to compare Microcosm [Davis et al., 92] to the Web [Berners-Lee et al., 94] because they propose different and almost opposite ways of representing and managing links, and this seems to have a big influence on authoring [Hill et al., 95]. Microcosm is an open environment, characterised by the separation of link structures from the information being linked [Hill et al., 95]. The WWW, on the other hand, provides a simple point-to-point linking model based upon embedded links.

**Procedure**

The survey involved the use of questionnaires that were answered by either Microcosm or Web developers. A survey offers the following advantages [Kitchenham, 96]: i) reaches a lot of users; ii) makes use of existing experience; iii) makes use of standard statistical analysis techniques; and iv) confirms that an effect generalises to many projects/organisations. However, surveys can only confirm association not causality, and are prone to bias.

Both questionnaires had three sections: experience, maintainability and reusability. For each section the questions were proposed with the objective of collecting the necessary data to test the hypotheses.
The experience section was based on a composition of two aspects: Entities and Hypermedia. The maintainability section was also based on a composition of three aspects: Entities, Hypermedia and Maintainability. The reusability section was based on a composition of three aspects: Entities, Hypermedia and Reusability.

In order to prepare both the Maintainability and the Reusability sections we had to consider possible tasks accomplished by authors in the development of hypermedia applications for education.

The Pilot Study

Before sending the questionnaires to both Microcosm and Web authors we carried out a pilot study because it provides an opportunity to learn from mistakes without ruining the main study [Preece et al., 94]. Feedback from colleagues prompted some changes to the questionnaire concerning ambiguous questions, unusual tasks, definitions in the appendix and the number of questions.

The Data Analysis

The survey results were analysed using standard statistical techniques to determine whether the two sets of questionnaires (from Microcosm and Web authors) were from different populations. Results from the Kruskal-Wallis one-way analysis of variance, using a level of significance of 5% are shown in Tables 5-8. For the results presented in Table 9 we used Spearman’s correlation with one-tailed test of significance.

The level of experience of the two groups of users (Microcosm and the Web) was found to be the same, as was the structure of the hyperdocuments [Tab. 4].

Table 4 - Type of structure used by both groups

<table>
<thead>
<tr>
<th>Structure</th>
<th>Microcosm</th>
<th>Percentage</th>
<th>Web</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>01</td>
<td>5,5</td>
<td>01</td>
<td>04</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>12</td>
<td>67</td>
<td>18</td>
<td>64</td>
</tr>
<tr>
<td>Network</td>
<td>04</td>
<td>22</td>
<td>07</td>
<td>25</td>
</tr>
<tr>
<td>No answer</td>
<td>01</td>
<td>5,5</td>
<td>02</td>
<td>07</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>100</td>
<td>28</td>
<td>100</td>
</tr>
</tbody>
</table>

In the questionnaire there were fifteen tasks related to actions in maintenance and reuse. The first nine are common tasks involved in maintenance and reuse and the last six are more unusual tasks, but also important considering maintenance and reuse. They are described below:

1) Finding dangling links within a document that has five links to other documents.
2) Deleting a document, that has five links to other documents, without leaving dangling links.
3) Adding a new paragraph to the beginning of a text document, that has five links to other documents, keeping the links intact.
4) Modifying the source anchor of a link.
5) Modifying the destination of a link.
6) Deleting a link.
7) Checking for dangling links caused by the deletion of a document that had two links.
8) Link ten terms to descriptions defined in a glossary.
9) Copy five documents (each with two links to other documents) to another application, keeping all the links already defined.
10) Finding if a document is part of an island.
11) Moving five documents (each with five links) from one directory to another, keeping their links valid.
12) Moving five documents (each with five links) from machine A to machine B, keeping their links valid, where both machines have the same operating system.
13) Checking for islands caused by the deletion of five links.
14) Linking a word that occurs in five documents (once in each) to a destination document.
15) Copying a document that has two links within your application, keeping all the links already defined.

For each of these tasks, authors were asked: i) the level of difficulty to accomplish those tasks, on a scale from 1 (very easy) to 5 (very difficult) and ii) the time it would take, in minutes, using 10 different intervals given.

When comparing tasks involving point-to-point links in both Microcosm and the Web we found that in 33% of the answers the medians for the level of difficulty were lower for Microcosm than for the Web and in 46% of the answers the time was shorter.

In 46% of the answers the time spent in both Microcosm and the Web was the same. But Web authors needed to use an auxiliary set of tools in order to accomplish the tasks in a reasonable time and with a low level of difficulty. This was not necessary using Microcosm.

Even with 7 answers where the level of difficulty was higher for Microcosm than for the Web there was no corresponding increase in the time spent to accomplish the tasks. As Microcosm is an open hypermedia system, the author has to edit the linkbase many times in order to maintain links. This task can be considered more difficult than changing links on the Web, but, as shown by the data, there was no overhead on the time spent.

When comparing tasks involving point-to-point links in both Microcosm and the Web we also found 8 answers with a statistically significant difference. Four
showed advantages for the Web and four showed advantages for Microcosm. The medians for tasks involving Microcosm point-to-point links, Web point-to-point links and the corresponding level of significance are presented in [Tab.5]:

Table 5 - Medians for tasks involving point-to-point links in Microcosm and the Web, with corresponding level of significance.

<table>
<thead>
<tr>
<th>Question</th>
<th>Attribute</th>
<th>Median point-to-point Microcosm</th>
<th>Median point-to-point Web</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>02</td>
<td>Time</td>
<td>1</td>
<td>2.5</td>
<td>0.04*</td>
</tr>
<tr>
<td>05</td>
<td>Difficulty</td>
<td>2</td>
<td>1</td>
<td>0.00*</td>
</tr>
<tr>
<td>06</td>
<td>Difficulty</td>
<td>2</td>
<td>1</td>
<td>0.03*</td>
</tr>
<tr>
<td>08</td>
<td>Time</td>
<td>1</td>
<td>3</td>
<td>0.03*</td>
</tr>
<tr>
<td>12</td>
<td>Difficulty</td>
<td>1</td>
<td>2</td>
<td>0.04*</td>
</tr>
<tr>
<td>13</td>
<td>Difficulty</td>
<td>1</td>
<td>2</td>
<td>0.00*</td>
</tr>
<tr>
<td>14</td>
<td>Difficulty</td>
<td>3</td>
<td>1.5</td>
<td>0.03*</td>
</tr>
<tr>
<td>15</td>
<td>Difficulty</td>
<td>2</td>
<td>1</td>
<td>0.00*</td>
</tr>
</tbody>
</table>

*denotes that the result is statistically significant at the 5% level

Questions 5, 6, 14 and 15 represent simple tasks, but for Microcosm authors involve the editing of the linkbase in order to update the information about the links. We understand that this was the reason for a higher level of difficulty using Microcosm. But, even with a higher level of difficulty, no statistically significant differences were found when comparing the time involved in the same tasks.

Questions 2 and 8 showed a statistically significant difference in the time spent in accomplishing the tasks. The time was higher using the Web. Questions 12 and 13 also showed a statistically significant difference in the level of difficulty spent in accomplishing the tasks. Again the level of difficulty was higher using the Web.

Questions 12 and 13 would be easily accomplished (in Microcosm) using generic links for the former question and local links for the latter question. Here we can see that when the applications require the definition of links to be valid within the whole application or within a particular document, the use of point-to-point links on the Web increases both the time involved and the level of difficulty in accomplishing the task.

For 13 questions that were not specifically designed to consider tasks that would be better suited for generic or local links, Microcosm authors were asked to estimate the time and level of difficulty in accomplishing the tasks if the links were either point-to-point or generic.

When comparing the answers given for generic links to those given for point-to-point links on the Web we found 8 questions (10 answers) with a statistically significant difference. All the 10 answers showed advantages for generic links. The medians for generic links, medians for point-to-point links on the Web and the corresponding level of significance are presented in [Tab. 6]:
Table 6 - Medians for tasks involving generic links and point-to-point links, with corresponding level of significance.

<table>
<thead>
<tr>
<th>Question</th>
<th>Attribute</th>
<th>Median Generic Microcosm</th>
<th>Median point-to-point Web</th>
<th>Level Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>03</td>
<td>Time</td>
<td>1.0</td>
<td>1.5</td>
<td>0.00*</td>
</tr>
<tr>
<td>04</td>
<td>Time</td>
<td>0.5</td>
<td>1.0</td>
<td>0.04*</td>
</tr>
<tr>
<td>05</td>
<td>Difficulty</td>
<td>1.0</td>
<td>1.0</td>
<td>0.00*</td>
</tr>
<tr>
<td>08</td>
<td>Time</td>
<td>1.0</td>
<td>3.0</td>
<td>0.00*</td>
</tr>
<tr>
<td>09</td>
<td>Time</td>
<td>1.0</td>
<td>2.0</td>
<td>0.03*</td>
</tr>
<tr>
<td>10</td>
<td>Time</td>
<td>1.0</td>
<td>2.0</td>
<td>0.07**</td>
</tr>
<tr>
<td>12</td>
<td>Time</td>
<td>2.0</td>
<td>3.0</td>
<td>0.07**</td>
</tr>
<tr>
<td></td>
<td>Difficulty</td>
<td>1.0</td>
<td>2.0</td>
<td>0.00*</td>
</tr>
<tr>
<td>13</td>
<td>Time</td>
<td>1.0</td>
<td>2.0</td>
<td>0.08**</td>
</tr>
<tr>
<td></td>
<td>Difficulty</td>
<td>1.0</td>
<td>2.0</td>
<td>0.00*</td>
</tr>
</tbody>
</table>

*denotes that the result is statistically significant at the 5% level
**denotes that the result is statistically significant at the 10% level

We can see that in 62% of the questions considered, generic links allowed either a shorter time or lower level of difficulty, when compared to accomplishing the same tasks involving point-to-point links on the Web.

The only question (question 13), that compared tasks involving local links to point-to-point links showed a statistically significant difference in favour of local links. The median for local links, median for point-to-point links on the Web and the corresponding level of significance are presented in [Tab. 7]:

Table 7 - Medians for tasks involving local links and point-to-point links, with corresponding level of significance.

<table>
<thead>
<tr>
<th>Question</th>
<th>Attribute</th>
<th>Median Local Microcosm</th>
<th>Median Point-to-point Web</th>
<th>Level Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Time</td>
<td>1</td>
<td>2</td>
<td>0.00*</td>
</tr>
<tr>
<td></td>
<td>Difficulty</td>
<td>1</td>
<td>2</td>
<td>0.08**</td>
</tr>
</tbody>
</table>

*denotes that the result is statistically significant at 5% level
**denotes that the result is statistically significant at 10% level

We found values of Gamma higher than 0.50 not only for the four independent variables presented in table 7, but also for the number of links and the structure of the application. Values for Gamma equal or higher than 0.50 show that there exists an association between the variables compared.
CONCLUSIONS

We have presented our approach to the development of metrics within the SHAPE research project and how they were evaluated.

The metrics were proposed to measure the maintainability and reusability of hypermedia applications for education, so that we could evaluate whether a particular hypermedia application for education was more or less maintainable or reusable than another application. Therefore, the metrics proposed are not restricted to a particular hypermedia system since they can be used to measure the maintainability and reusability of any hypermedia applications.

In order to investigate the metrics proposed we collected the data using applications developed with both Microcosm and the Web.

The data collected showed strong evidence that the link representation, link type, highlighting of anchors, structure of the application and the author’s experience can strongly influence the maintainability of the application and the reusability of information.

We also found some evidence that the number of documents, compactness and stratum can also influence the maintainability of the application and the reusability of information.

The next evaluation will be both quantitative and qualitative and its aim will be to measure the development effort involved in the development of a hypermedia application using both Microcosm and the Web.

REFERENCES


