Reusability and maintainability in hypermedia applications for education

Emilia Mendes*, Rachel Harrison, Wendy Hall

Department of Electronics and Computer Science, University of Southampton, S017 IBJ, Highfield, Southampton, UK

Abstract

This paper reports the results of applying metrics to hypermedia authoring under the SHAPE (Southampton hypermedia authoring paradigm for education) 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.

In the hypertext field there have been a number of proposals for hypertext metrics, developed mainly in an ad-hoc fashion, contributing to the expression of measures in an ambiguous manner and limiting their use. Unfortunately, many measures proposed in the literature lack the necessary mathematical 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 hypennedia authoring systems. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Metrics; Hypermedia; Hypermedia for education; Maintainability; Reusability

1. Introduction

We regard measurement as important for three basic activities:

- Understanding the development and maintenance processes.
- Controlling software projects.
- 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 [1,2].

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 [3].

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. [4]:

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 [4,36].

The key stages of formal measurement are shown in Fig. 1. This refers to the representation condition [3].

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 our developed metrics applied to hypermedia authoring. Finally, in Section 5, we give our conclusions and comments on future work.

* Corresponding author. Fax: +44 1703 592865; e-mail: [mexm95, rh,-wh]@ecs.soton.ac.uk

0950-5849/98 - see front matter © 1998 Elsevier Science B.V. All rights reserved.
PII: S0950-5849(98)00096-2
2. Survey of hypertext metrics

The application of metrics to hypermedia has already stimulated considerable interest [5–11]. Each paper is briefly described in the following subsections.

2.1. Hypertext metrics by Botafogo et al.

Botafogo et al. [5] defined metrics based on the assumption that giving authors an idea of how complex and connected the hypertext is could help them in the authoring process. They proposed the stratum and compactness metrics, both calculated from the structure of the hypertext. The compactness metric indicates how connected the hypertext is. Its value varies between 0 and 1, corresponding, respectively, to a completely disconnected hypertext and a completely connected hypertext. The stratum metric indicates to what degree the hypertext is organised so that some nodes must be read before the others. Its value also varies between 0 and 1, where 1 corresponds to a linear hypertext and 0 corresponds to no imposed reading order.

2.2. Hypertext metrics by Garzotto et al.

The work of Garzotto et al. [8] provides a framework for design-oriented hypermedia evaluation, using both a set of design objects, based on HDM (hypermedia design model), and a set of fine-grained attributes of these objects that can impact on hypermedia usability. They understand design as the ‘external level’ of a hypermedia application, rather than the design of implementation data structures and code. The external level is comprised of content organisation, behaviour of the various media, functionality offered to the user and layout of the various hypermedia elements. They concentrate on the usability attribute, which they decompose into learnability, comprehensibility, memorability, handling ability and niceness. Learnability, memorability and handling ability are further decomposed into lower level factors: consistency, self-evidence and predictability.

According to Garzotto et al. there are no rigorously defined ‘theoretical’ measures for the criteria. They suppose that one can achieve finer grained attributes by combining the three categories (consistency, self-evidence and predictability) with the three classes of design constituents (structure, behaviour, and presentation) and with the two levels of design (in-the-large and in-the-small), producing eighteen finer-grained attributes in total.

2.3. Hypertext metrics by Yamada et al.

Yamada et al. [11] propose three metrics: (i) an interface shallowness metric; (ii) a downward compactness metric; and (iii) a downward navigability metric. The interface shallowness metric represents the heaviness of the cognitive load on users. It is different from ‘depth’ [5], which is derived from the structural distance. The key idea is that nodes are linked in such a way that they do or do not preserve interface linearity, i.e. links can lead to documents in the same layer or on a different one. The downward compactness metric measures the compactness of links from the root, i.e. it denotes the structural complexity of reaching the nodes from the root. The downward navigability metric measures hypermedia navigability. This assumes that an easily navigable hypermedia system: (1) has a shallow interface layer from the root to the nodes (light cognitive load); and (2) is compact from the root (that is, it is structurally simple to reach the nodes from the root).

2.4. Hypertext metrics by Hatzimanikatis et al.

Hatzimanikatis et al. [10] define a hyperdocument quality model using the factor-criteria-metric model [12]. They focus their work on structure metrics, i.e. metrics that can be computed using only the structure of the hypertext (the
The factors considered are readability and maintainability, and these are decomposed into eight criteria: size, path complexity, tree impurity, modularity, node complexity, coherence, complexity of node contents and simplicity.

They defined metrics for: (i) path complexity; (ii) tree impurity; (iii) modularity; and (iv) individual node complexity. Path complexity is the number of different paths or cycles that can be found in a hypertext graph. The path complexity of a linear hyperdocument is minimal. Tree impurity is the extent to which a graph deviates from being a tree. The modularity metric measures whether the modules are selfcontained and independent. The individual node complexity is the complexity that a single node imposes on the overall hypertext structure.

3. Discussion and comparison

In the previous section four proposed hypertext metrics were presented. However, little corresponding empirical validation of these metrics has been published.

Table 1 compares the four proposed hypertext metrics considering the four questions that should be asked when validating a measure [15]:

<table>
<thead>
<tr>
<th>Questions</th>
<th>Metrics Proposals</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2)</td>
<td>Yes</td>
</tr>
<tr>
<td>(3)</td>
<td>No</td>
</tr>
<tr>
<td>(4)</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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 define 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 analysis of the metrics data collected will provide improvement feedback to SHAPE, since we understand that the authoring paradigm in SHAPE represents the improvement of authoring through measurement, analysis and feedback [19].

We are applying to our work the same principles already used in various experiments in the software engineering field [1,2,15,20–26]. Our work is based on Fenton et al.’s framework for software measurement [3], and on the guidelines from the Desmet project [27,28].

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

4.2. Design of the first evaluation

For the first evaluation the stated hypothesis was:

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

Using the data collected we also wanted to evaluate if:

- The use of a link server allows both a better maintainability of applications and reusability of information than embedded ones.
- Generic links allow a better maintainability and reusability of information than the equivalent set of point-to-point links.

In the Microcosm model, a link associates a particular source selection with its destination and can be specific (point-to-point), local or generic. A local link can be followed from any occurrence of the source selection in a particular document [29]. In the standard implementation of Microcosm local link anchors are not highlighted. A generic link can be followed from any occurrence of that source selection in any document [29]. In the standard implementation of Microcosm they are also not highlighted.

We have chosen to compare Microcosm [29] to the Web [30] because they propose different and almost opposite ways of representing and managing links, and this seems to have a big influence on authoring [31]. Microcosm is an open environment, characterised by the separation of link structures from the information being linked [31]. The WWW, on the other hand, provides a simple point-to-point linking model based upon embedded links.

4.2.1. Procedure

The survey involved the use of questionnaires that were answered by either Microcosm or Web developers. A survey offers the following advantages [27]:

- Reaches a lot of users.
- Makes use of existing experience.
- Makes use of standard statistical analysis techniques.
- Confirms that an effect generalises to many projects/organisations.

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.

4.2.2. 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 [32]. Feedback from colleagues prompted some changes to the questionnaire concerning ambiguous questions, unusual tasks, definitions in the appendix and the number of questions.

4.2.3. Fenton and Pfleeger’s framework

The conceptual framework proposed from Fenton and Pfleeger [3] can be applied to the diverse software-measurement activities that contribute to an organisation’s software practices. The practices can be not only the development and maintenance activities but also any experiments and case studies performed in order to investigate new techniques and tools. It is based on three principles:

- Classifying the entities to be examined.
- Determining relevant measurement goals.
- Identifying the level of maturity that an organisation has reached.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Classification of products, processes and resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>Attributes</td>
</tr>
<tr>
<td>Products</td>
<td>External</td>
</tr>
<tr>
<td>Application</td>
<td>application size, connectivity, structure (topology), compactness [5,33]</td>
</tr>
<tr>
<td>Tool</td>
<td>link representation, type of link, highlighting of anchors</td>
</tr>
<tr>
<td>Processes</td>
<td>External</td>
</tr>
<tr>
<td>Maintenance</td>
<td>effort, difficulty.</td>
</tr>
<tr>
<td>Reuse</td>
<td>effort, difficulty.</td>
</tr>
</tbody>
</table>
4.2.3.1. Classifying the entities to be examined  Entities can be processes, products or resources. Processes are collections of software-related activities, products are any results from a process activity and resources are entities required by a process activity. For each entity measured it is important to identify its internal and external attributes. Internal attributes can be directly measured, whereas external attributes can only be indirectly measured. The classification of entities applied to the quantitative evaluation for SHAPE is presented in Table 2.

4.2.3.2. Determining relevant measurement goals  The relevant measurement goals were determined using the goal-question-metric (GQM) approach [1,2], which is based upon the assumption that any measurement must be defined in a top down fashion. The result of applying the GQM approach is a model that has three levels: (i) the conceptual level — goal; (ii) the operational level — question; and (iii) the quantitative level — metric. The goal is refined into several questions and each question is then refined into metrics, either objective or subjective. SHAPE’S corresponding GQM is presented in Table 3.

4.2.3.3. Identifying the level of maturity that your organisation has reached  The level of maturity is based on the capability maturity model (CMM) [34], which provides a framework for organising the evolutionary steps into five maturity levels that lay successive foundations to continuous process improvement. These five maturity levels define an ordinal scale for measuring the maturity of an organisation’s software process and for evaluating its software process capability. Each maturity level comprises a set of process goals that, when satisfied, stabilise an important component of the software process.

The level of maturity within the hyperinedia application development community that is considered for SHAPE is either level 1 or 2.

4.2.4. 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 we used the Kruskal–Wallis one-way analysis of variance, with a level of significance of 5% and 10%. This test is extremely useful for deciding whether \( k \) independent samples are from different populations or whether they represent merely chance variations among random samples from the same population. To identify the correlation between the independent and dependent variables we used Gamma as a measure of correlation, with a level of significance of 10%. Gamma gives in a single number a summary measure of the existence, strength, and direction of the relationship [35].

Both groups shared similar experiences and levels of involvement in the development of the applications. This was confirmed statistically. The applications developed by both the Web authors and the Microcosm authors shared similar compactness, stratum, size, connectivity and structure of the applications. This was also confirmed using statistical tests. Both groups made use of various guidelines for the development of their applications.

4.2.5. The results  Table 4 shows the frequency of use of different architectures (sequential, hierarchical or network). The most frequently used architecture was found to be the hierarchical.

There was a statistically significant difference at the 5% level between the number of tools used by Web authors and Microcosm authors. Web authors used a higher number of tools, such as an HTML editor, an application generator and a software to organise and manage the HTML files. Microcosm authors, on the other hand, mentioned tools such as the link editor, the document management system and a word processor.

We measured the two dependent variables [effort (time) and level of difficulty] using a questionnaire with 15 tasks. For each of these tasks, authors were asked: (i) the level of difficulty to accomplish the task, 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. Thirteen tasks

<table>
<thead>
<tr>
<th>Structure</th>
<th>Microcosm percentage (%)</th>
<th>Web percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>5.5</td>
<td>04</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td>Network</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>No answer</td>
<td>5.5</td>
<td>07</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
were based on usual activities concerning maintenance and reuse. As we did not want to bias the evaluation, only two questions were developed where the tasks involved might be more effectively accomplished using generic or local links. These were questions 12 and 13 respectively. The tasks are described below:

Finding dangling links within a document that has five links to other documents. Deleting a document, that has five links to other documents, without leaving dangling links.
Adding a new paragraph to the beginning of a text document, that has five links to other documents, keeping the links intact.
Modifying the source anchor of a link.
Modifying the destination of a link.
Deleting a link.
Checking for dangling links caused by the deletion of a document that had two links.
Link ten terms to descriptions defined in a glossary.
Copy five documents (each with two links to other documents) to another application, keeping all the links already defined.
Finding if a document is part of an island.

Moving five documents (each with five links) from one directory to another, keeping their links valid.
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.
Checking for islands caused by the deletion of five links.
Linking a word that occurs in five documents (once in each) to a destination document.
Copying a document that has two links within your application, keeping all the links already defined.

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.

### Table 5

<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 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.

### Table 6

<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.
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 Table 5:

<table>
<thead>
<tr>
<th>Quest subject</th>
<th>Median local Microcosm</th>
<th>Median point-to-point Web</th>
<th>Level significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1</td>
<td>2</td>
<td>0.00*</td>
</tr>
<tr>
<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.

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 Table 6:

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 Table 7:

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.

Authors were asked, on a scale from 0 (not helpful) to 5 (very helpful), about the influence of the highlighting of anchors on maintainability/reusability. The results are presented in Table 9:

The medians for Microcosm are higher than the medians for the Web and this is probably caused by the fact that links on the Web are always highlighted, which is not the case in Microcosm.

The questionnaire did not consider reuse of links, since it does not make sense to reuse point-to-point links and these are the only types of links available on the Web. But the reuse of links is an important issue in the reuse process as a whole and we think that one of the advantages of Microcosm is that it allows the reuse of local and generic links, as
they are stored separately from the documents, in linkbases. Any linkbase can be ‘plugged in’ to any application, and this leads to a high level of flexibility for authors.

5. 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 inmaintainability 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.

Our next evaluation will be a quantitative and qualitative evaluation and will measure the authoring effort involved in developing a hypermedia application for education. The evaluation will consist of developing the same application using both Microcosm and the Web. The application will be designed using the principles from cognitive flexibility theory [7] and the authors will be undergraduate computer science students studying human–computer interaction.

References


