Effective Retrieval to Support Learning

A thesis submitted for the degree of
Doctor of Philosophy

Michael C. Harris B.Sc., M.Tech.,
School of Computer Science and Information Technology,
College of Science, Engineering, and Health,
RMIT University,
Melbourne, Victoria, Australia.

December 2010

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Michael C. Harris
School of Computer Science and Information Technology
RMIT University
December 2010

Acknowledgments

Early in my candidature, a stranger at a party asserted that it was impossible to maintain a relationship while completing a PhD; one or the other had to give. I have been incredibly fortunate that Rachel has not only managed to find it in herself not to leave me, she has been a constant source of strength and encouragement throughout the journey.

Another constant has been the incredible patience and generosity of my supervisors, and I express my enormous gratitude to all three of them. I cannot imagine a better research mentor than my first supervisor, Associate Professor James Thom. He was always willing to give feedback, even on early drafts and work that he had read many times before, and his attention to detail is extraordinary. My second supervisor, Dr Falk Scholer, and my former second supervisor, Dr Seyed Mohammad Mehdi (Saied) Tahaghoghi, each provided the perfect complement, and more.

This would have been a much lonelier and more difficult experience without the invaluable support and advice from Dr Lawrence Cavedon and Dr Barb Kelly, who have been through this before, and Anne Gordon and Kath Lynch, who are sharing the journey. Donal Ellis made the sensible decision to take another road, but was always there telling me to DO IT when I wavered. Members of the community of the Habari project distracted me, keeping me sane, and told me to stop messing around and get back to my research, in equal measure. Too many people to mention who are involved in the project helped me solve numerous technical and code issues. Andy Cowling made me laugh and provided much of the soundtrack.

Our investigation of how organisations currently manage educational resources, presented in Chapter 3, was conducted as part of the Reusable Learning Objects project, which was run through the Teaching and Learning Portfolio of RMIT University, and overseen by a Project Reference Group. I thank members of the reference group, particularly Amgad Louka, for their support, and my fellow research officer, Henric Beiers. I could not have completed this work without the participants in my focus groups, interviews, survey, and user experiments, and I am indebted to all of them for taking time out to help me. I would not have found many of my participants without the recruitment help provided by Leanne Trinder.

This work was partially supported by an Australian Postgraduate Award.

And finally, thanks again to Rachel, for proving the stranger wrong.

Credits

Portions of the material in this thesis have previously appeared in the following publications:

Refereed Conferences and Workshops

- Harris, M. C., Thom, J. A. and Scholer, F., How consistent are human judgments of whether an open resource is educational material? In H.T. Shen and A. Bouguettaya, editors, *ADC2010 21st Australasian Database Conference*, Brisbane, Australia, January 2010.

- Harris, M. C., Thom, J. A., Challenges facing the retrieval and the reuse of learning objects. In M. R. Artacho and E. Duval, editors, *Proceedings of the Workshop on Learning Object Repositories as Digital Libraries: Current Challenges*, Alicante, Spain, September 2006.

- Harris, M. and Beiers, H. Barriers to the Reuse of Learning Objects. In P. Kommers and G. Richards, editors, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Montreal, Candada, June–July 2005.

The thesis was written in the `Vim` editor, and typeset using the LaTeX 2_ε document preparation system. Machine learning tasks were carried out using the Weka data mining software. All trademarks are the property of their respective owners.

Note

Unless otherwise stated, all fractional results have been rounded to the displayed number of decimal figures.

Contents

List of Figures

List of Tables

Abstract

To use digital resources to support learning, we need to be able to retrieve them. This thesis introduces a new area of research within information retrieval, the retrieval of educational resources from the Web.

Successful retrieval of educational resources requires an understanding of how the resources being searched are managed, how searchers interact with those resources and the systems that manage them, and the needs of the people searching. As such, we began by investigating how resources are managed and reused in a higher education setting. This investigation involved running four focus groups with 23 participants, 26 interviews and a survey.

The second part of this work is motivated by one of our initial findings; when people look for educational resources, they prefer to search the World Wide Web using a public search engine. This finding suggests users searching for educational resources may be more satisfied with search engine results if only those resources likely to support learning are presented. To provide satisfactory result sets, resources that are unlikely to support learning should not be present. A filter to detect material that is likely to support learning would therefore be useful.

Information retrieval systems are often evaluated using the Cranfield method, which compares system performance with a ground truth provided by human judgments. We propose a method of evaluating systems that filter educational resources based on this method. By demonstrating that judges can agree on which resources are educational, we establish that a single human judge for each resource provides a sufficient ground truth.

Machine learning techniques are commonly used to classify resources. We investigate how machine learning can be used to classify resources retrieved from the Web as likely or unlikely

to support learning. We found that reasonable classification performance can be achieved using text extracted from resources in conjunction with Naïve Bayes, AdaBoost, and Random Forest classifiers. We also found that attributes developed from the structural elements—hyperlinks and headings found in a resource—did not substantially improve classification over simply using the text.

Chapter 1

Introduction

Learning is fundamental to being human. We spend much of our time acquiring knowledge and skills, both informally in our day-to-day lives and in the formal context of education. A variety of resources can be used to help us learn, but from the latter decades of the 20th century more and more of these resources were created in digital form and delivered electronically. This thesis is about finding appropriate digital educational resources to support learning, and introduces a new area of information retrieval research investigating the retrieval of educational resources from the Web.

Educational resources can be retrieved from two types of collections. First, they might be stored in a collection in which all resources are educational. In this case, the problem of retrieving appropriate resources is a topic-based problem. Second, they might be part of a collection of different types of resources. In such a heterogeneous collection, in addition to the topic-based issues, there is a further filtering problem; of the relevant resources retrieved, which are educational or likely to support learning? We can think of the Web as such a heterogeneous collection.

As digital resources have become easier to create [Counts, 2006] and their use in educational settings has dramatically increased [Hannafin and Hill, 2007], universities, corporations, governments, and other institutions have created many collections devoted to educational resources, and significant time and money has been spent in the private and public sectors developing and maintaining systems designed to manage these resources. Many educa-

tional resources are released on the Web as Open Educational Resources (OERs),[1] resources people can use and adapt for non-commercial purposes [Wiley, 2007b].

In 2009, a range of OER initiatives were announced. The production of educational resources to be released online formed a significant part of the $US12 billion American Graduation Initiative announced by the government of the United States of America[2]. In the private sector, the Hewlett Foundation announced that their OER project, having run for seven years, was one of their most successful, and made over $16 million in grants that focussed on the production and management of OERs, and other foundations (Gates, MacArthur, Lumina) funded more than $US13 million OER projects [Hewlett Foundation, 2009]. Many more resources are produced and released under less formal arrangements. With such large numbers of resources being produced, finding appropriate resources is not trivial.

This chapter introduces the research covered in this thesis, describing our research questions in Section 1.1 and giving an outline of how the thesis addresses these questions in Section 1.2.

1.1 Research objectives

Educational resources can be retrieved from two types of collections: homogeneous collections, which contain resources that are all assumed to be educational, and heterogeneous collections, such as the Web, which contain a variety of educational and non-educational resources.

Existing research has focussed on situations in which learning is mediated by an academic or teacher searching an homogeneous collection with a system that queries human-assigned descriptions of resources and attempts to return relevant educational resources. While we contend that applying information retrieval (IR) techniques based on full-text indexing would improve effectiveness, our investigation described in Chapter 3 demonstrates that most people prefer to find educational resources on the Web rather than from institutional repositories. As such, we address the problem of filtering resources retrieved from the Web, so as to return only resources likely to support learning, an area that has not been previously explored.

[1] Refer to Appendix J for a glossary of terms.

[2] http://www.whitehouse.gov/the_press_office/Excerpts-of-the-Presidents-remarks-in-Warren-Michigan-and-fact-sheet-on-the-American-Graduation-Initiative/

In the following, we outline the specific research questions addressed in this thesis.

1.1.1 How are educational resources currently used and managed?

It should be clear that the management of educational resources, their use and reuse, and the interaction between the individuals and systems involved, presents many complexities. In recognition of that, we chose to begin our work with qualitative research, to develop an understanding of those complexities. This understanding is then used to inform our work in the following chapters.

To ensure the effectiveness of retrieval to support learning it is useful to understand the context in which people seek and use educational resources. In situations where resources are to be retrieved from a homogeneous collection of educational resources, the broad context is usually within an educational institution. We address two aspects of this question; how organisations currently manage educational resources, and how academics and teachers use and reuse educational resources.

We first focus on how educational institutions manage resources and the opinions and requirements of individuals within those educational organisations. This includes the high level organisational and cultural issues around the management and reuse of digital resources for teaching and learning. We do this by talking to people who have been involved in projects that attempt to manage and reuse learning material, and people who might be affected by the introduction of a software system to facilitate the management and reuse of educational resources.

Next, we investigate specific instances in which people working in an educational institution have successfully discovered and reused learning material. We explore the circumstances in which the reuse occurred, how the resources that were reused were found, and how the individuals involved want to be able to find resources to reuse in the future.

1.1.2 How should systems that filter educational resources be evaluated?

After completing our investigation into the management and reuse of educational resources in institutional settings, we focus on the retrieval of resources to support learning from the Web. The Web is a heterogeneous collection of resources, not all of which are likely to support learning. Filtering educational resources from heterogeneous collections has not been the

subject of any research before this work, and as such we investigate how systems that filter educational resources should be evaluated. As our starting point, we propose an evaluation methodology based on the Cranfield method, which is widely used in the field of information retrieval. The Cranfield method uses test collections to simulate real-world conditions, by taking a set of resources that are representative of the domain under investigation and a set of topics that are representative of the information needs of users, and drawing a relevance relationship between the resources and topics.

An important aspect to consider when using the Cranfield method is what makes an appropriate test collection, and whether an existing test collection can be adopted or a new test collection should be constructed. As the filtering of educational resources has not been investigated before, there is no existing collection of resources with associated judgments of which resources are likely to support learning. Therefore, we investigate how a labelled test collection should be constructed. Specifically, we investigate how many judges should be used; the effect of displaying the query used to retrieve the resource; the relationship between the confidence of judges, the ease of judging a resource, and the expertise of judges on the judgment outcome; the impact of the educational depth of resources; and the attitudes of judges to the overall judging process.

As most test collections used in information retrieval experiments use a single judge to assign a relevance relationship between a topic and a resource, we use multiple judgments to investigate whether the level of agreement about whether a resource is likely to support learning can be used to justify the use of a single judgment. Additionally, as the assessment that judges make about resources is constant—whether they are likely to support learning—as opposed to the assessments made between a resource and a variety of topics for most test collections, we investigate the effect on judgments of displaying the query used to retrieve the resource. We are also interested in how judges make assessments of resources, and therefore we investigate the relationship between the confidence and ease with which judges assign a particular label, the expertise of judges, and the judgment outcome. We explore the possibility that the educational depth of resources impacts the ability of classification models to accurately classify resources. Finally, we examine the attitudes of judges to the overall judging process.

1.1.3 What methods can be used to effectively filter educational resources?

A common approach to filtering problems is to use machine learning techniques to classify resources. In machine learning classification, a set of labelled training data is used as input to a classification or induction algorithm, which produces a machine learning model. This model can then be used to classify test data that is also labelled, with performance measured based on how accurately the model classifies the test data. In the case of the classification of web resources as educational or not educational, each item or instance in the training and test data is a set of attribute-value pairs that represents a web resource. The values of these attributes are programatically extracted from features of individual resources. Examples of such features include the terms in the text, the number of headings, and the ratio of link text to other text.

Using a small development collection of web resources, we explore features that can be extracted from resources and represented as attributes to be used for the effective classification of resources as educational or not educational. We also investigate a variety of commonly used classification algorithms, and assess which algorithms perform best for this classification task.

To avoid bias, the final assessment of the effectiveness of machine learning models should be made using data that are independent of the data used in the development of the models. For this reason, we investigate whether the attributes developed from the extracted features and the chosen classification algorithms are capable of developing effective classification models using a larger, independent collection of resources.

1.2 Thesis organisation

To address our research questions concerning the effective retrieval of educational resources to support learning, the thesis is organised as follows.

In Chapter 2 we present a review of the literature relevant to our research. We begin by describing other work investigating the use of digital resources in educational settings, including how the recent increase in the use of digital resources is viewed by those working in educational institutions. We introduce the concept of learning objects, which grew out of an effort to create a theoretical framework in which to consider educational resources designed for

reuse, and discuss related work in user search goals and the retrieval of educational resources. This is followed by a description of methodologies for the evaluation of information retrieval systems, focussing on the test collections used in such evaluation and the measures that are used to quantify the performance of systems. We end the chapter with a discussion of machine learning, particularly its use in information retrieval settings. We describe the evaluation measures and methods regularly used in machine learning, relating them to those used in information retrieval, and discuss methods used in the comparison of machine learning classifiers.

In Chapter 3 we describe our qualitative work exploring the management and reuse of educational resources, specifically addressing our first research question. We first outline the research methodology that we employ and motivate our methodological decisions. We then describe our data collection and discuss the results. We begin with two focus groups where participants had some awareness or involvement with projects to manage educational resources, and two focus groups where participants did not. This was followed by a series of interviews with people recommended as experts in the field. Results from the focus groups and interviews were then used to conduct a survey of university staff. Finally, we conducted a second series of interviews with people who had successfully reused educational resources.

In Chapter 4 we propose a methodology for the evaluation of systems that filter educational resources, addressing our second research question. We start by describing the elements of a system for building suitable test collections. An important part of such systems is the judges who will label the resources in the collection, and we describe methods for measuring the amount of agreement between judges about how a resource should be labelled. We then present an experiment to investigate the level of agreement between judges when classifying resources as educational or not educational, with the aim of establishing how many judges should label each resource. In information retrieval evaluation, judges label resources according to whether or not they are relevant to a particular query, whereas we investigate the labelling of resources based on whether or not the resources are likely to support learning. In this case, the query used to retrieve the resource is not central to the task, and so we also investigate the effect of query visibility on judgments of whether or not a resource is likely to support learning.

In Chapter 5 we address our third research question, proposing methods by which educa-

tional resources can be filtered. We begin by describing the construction of a test collection to be used for the development and tuning of machine classifiers for educational resources. Focussing on text that can be extracted from resources, we introduce a cross-section of commonly used machine learning induction algorithms and assess their appropriateness for the task of classifying resources as educational or not educational. Having identified a subset of induction algorithms that perform well in this domain, we investigate alternative methods for turning the text extracted from resources into sets of attributes to be used as input for these algorithms. Finally, we investigate attributes derived from other features of resources that might be useful in classification, specifically hyperlinks and headings.

In Chapter 6 we draw together the work presented in previous chapters, continuing the exploration of our second and third research questions. We run a further user experiment using the methodology we proposed for developing appropriate test collections to construct a larger, independent collection of resources. To make the judging process easier, we provide judges with an artificial context in which to make their assessments. Specifically, we ask that judges assess resources as likely to support the learning of a student in high school, likely to support learning in another context, or unlikely to support learning. To gain insight into the judging process, we ask judges exploratory questions after each judgment and at the completion of all judgments. From this labelled collection, we again programatically extract attributes based on the features we have previously explored, and assess the performance of a shortlist of classification models.

The thesis concludes in Chapter 7, in which we summarise the contributions made by this research, along with the limitations and directions for future research. Major contributions include the finding that educators would prefer to use a public search engine to find educational resources, the description of a methodology for the evaluation of systems that filter educational resources, the finding that judges demonstrate a high level of agreement in whether resources are likely or unlikely to support learning, and that terms extracted from the text of resources are useful for classifying resources as educational or not educational.

Chapter 2

Educational Resources, Information Retrieval, and Machine Learning

The growing use of digital resources in education has increased the importance of managing them effectively. The management of educational resources has implications for both institutions providing education and the educators who use the resources. The focus of this thesis is one aspect of that management: effective retrieval to support learning. Therefore, we build on research investigating the management and reuse of educational resources, and we apply techniques and methodologies developed in the fields of information retrieval and machine learning.

We begin with an overview of the impacts and outcomes of the use of digital resources in education, including how this increased use is viewed by academics and teachers, in Section 2.1. Finding appropriate resources is an important aspect of the management of digital resources in education. In Section 2.2, we provide background to the use of information retrieval systems and their evaluation. For a retrieval system to only return resources that are likely to be educational, it needs to have some method of distinguishing those resources likely to support learning from those unlikely to support learning. A common approach to this type of classification problem is to use machine learning. In Section 2.3 we discuss machine learning techniques that can be applied to the classification of resources. Finally, in Section 2.4 we provide a summary of the themes developed in the reviewed literature.

2.1 Digital resources in education

The traditional approach to course development and delivery involves individual teachers producing material and delivering courses for relatively small numbers of students [Schlusmans et al., 2004]. Using this approach, delivery could be altered and resources could evolve as a course progressed, and associated costs are spread over the development and delivery of the course. Distance education institutions in the 1970s and 1980s used a different approach, where quality resources were developed and delivered to large and diverse cohorts of students. Resource development in these institutions became the most expensive aspect of the educational life cycle, and improved management and reuse of resources became essential.

As of the mid-1990s, academics began to make use of the Web in course delivery. Distance education, predominantly on the Web, is growing at a rate three times faster than classroom-based education, and corporate training is worth billions of dollars a year [Christensen et al., 2003]. At many institutions, delivery of courses on the Web was done in much the same way as traditional course delivery, and many resources were simply transferred online. Universities began to recognise that a translation of traditional teaching to online was not sufficient, and that the production of higher quality resources was necessary. However, production of high quality resources is expensive, and therefore, mirroring the earlier experience of open and distance learning institutions, management and reuse of educational resources became more important [Schlusmans et al., 2004].

Much of the research on educational resources focusses on technical aspects necessary for reuse, however to make systems for the sharing and reuse of resources viable, human issues are likely to be the most important [Phillips et al., 2003]. Several authors have explored the complexities involved in the reuse and sharing of educational resources. McNaught [2003] identifies seven factors that contribute to this complexity. These factors include: how to maintain scholarly standards while pursuing the reduction of the costs of production of educational resources; how educators should be supported in the production, reuse, and sharing of resources; the need to manage a balance between adopting change and increasing stress for educators; incentives for contributing quality educational resources; how to foster communities of practice within institutions; methods for encouraging educators to begin contributing resources, despite the technical difficulties that they might encounter; and the need to maintain the best of traditional education practices while embracing a more adaptable

and accessible model.

Campbell [2003] suggests that issues relating to the reuse of educational resources can be grouped into four areas: cultural, educational, technological, and issues of interoperability. While constructing a comparative case study Littlejohn et al. [2003] point out that the specific issues are likely to be different in higher education institutions when compared to vocational education and training (VET) institutions. They highlight several key issues: standardised curricula, as exist across VET institutions, do not necessarily mean staff will use externally produced resources; the ability to contextualise resources is important; appropriate incentives must exist for staff who reuse and contribute resources; support programs for staff must exist; and efficient and effective search is critical.

For successful communities of reuse to become established, Littlejohn [2003] describes three levels of skill that need to be developed in educators. First, educators need to be able to effectively search for and discover educational resources. Second, resources themselves should be designed with reuse in mind. Third, educators need to be able to take the disparate resources they discover and provide appropriate contextualisation to create pedagogically rich courses.

While teachers have experience reusing their own material and incorporating paper-based and electronic material, Littlejohn et al. [2003] contend that a culture of reuse of digital resources produced by others does not exist. In the context of Scottish Higher Education, Campbell et al. [2001] found that willingness to reuse resources produced outside an educator's institution is contingent upon the resources being quality controlled, peer reviewed and having clearly identified authorship and provenance. The study also provided evidence that educators showed reluctance to contribute their own material for reuse, predominantly because of a fear of loss of copyright or intellectual property.

Koppi et al. [2005] describe the creation of a repository of educational resources that have a high level of quality assurance. This led some content creators to refuse to submit resources for fear they were not high enough quality.

Part of the drive to create reusable resources is the possibility of creating a commercial resource. Reporting on the outcomes of a symposium devoted to exploring the theme of the ownership of educational resources, Twigg [2000] outlines several significant hurdles to universities creating successful commercial enterprises based on the creation and distribution

of educational resources, and points out that, in terms of institutions offering tertiary education, it is the structured educational experience that is of value rather than the resources themselves.

Another concern of educators in relation to the sharing of educational resources is what has been called the "Player Piano" scenario [Noble, 1998]. This scenario is based on the premise that an educator's worth can be embodied in the resources that they produce. Once their knowledge is captured in this way, educators are no longer valuable to their employing institution, and are thus in danger of losing their jobs. Twigg [2000] reports that participants of the symposium acknowledged that the role of faculty is changing, but dismissed the possibility that their jobs are on the line as a result of the reuse of educational resources, arguing that the premise is flawed, and that there is value in interaction between educators and students.

It has been argued that collaborative evaluation of educational resources is a necessary part of their life cycle [Nesbit and Belfer, 2004]. Such evaluations might include ratings and reviews of resources, as has been successful on many commercial websites, with Amazon[1] being a well-known example. Forms of collaborative evaluations, such as user generated reviews and tracking of usage patterns, are becoming more widespread as mechanisms facilitate retrieval and trust [Campbell, 2003]. In addition to helping to improve the ability of searchers to make a decision about whether a resource is a good candidate for reuse, this user participation could assist in other important aspects necessary for resource sharing, such as community building [Nesbit and Belfer, 2004].

Though academics are used to sharing resources through the publication of research papers, there is a lack of incentive to share teaching material [Campbell, 2003]. Academics do not tend to think of their learning activities or teaching resources as objects to be shared, in marked contrast to how they view research outputs. An appropriate reward system for the contribution of resources is lacking [Koppi et al., 2005]. There is a perception among academics that research is more highly valued than teaching [McNaught, 2003] and this perception needs to be addressed before significant numbers of resources will be contributed [Koppi and Lavitt, 2003]. Taylor and Richardson [2001] attempt to address the issue of reward by presenting a scheme for peer review of ICT-based resources in teaching and learning. They

[1] http://amazon.com

suggest that the adoption of such a system of peer review of resources by educational institutions may increase the possibility that educators would be willing to contribute resources for reuse by others.

In a review of projects run by the Joint Information Systems Committee (JISC) in the United Kingdom focussing on the sharing of eLearning content, Charlesworth et al. [2007] examine 30 JISC-funded or identified projects and 70 papers and outputs from those projects. Broadly, they report that many stakeholders feel that the cultural, legal, and organisational issues that are of most concern are often overshadowed by a focus on technical solutions. Specifically, they identified large amounts of informal sharing, but concluded that contributing resources for more formal sharing within an institution would require cultural change. Sharing was more common in institutions with a focus on teaching as opposed to research. They found that authorship attribution was important to many considering contributing resources for sharing, and that altruism was as strong an incentive to contribute as financial rewards. They did not address other possible incentives for contribution. A greater willingness to share resources was demonstrated in institutions where there is trust within the organisation, and clear explanation of how resources might be used was required in order to promote trust. The report stated that attitudes to intellectual property and copyright inhibit innovation in content management, and that there appeared to be considerable frustration at the apparent complexity of copyright processes. This perceived complexity, as well as onerous sign-off processes, act as a substantial disincentive to contribution.

2.1.1 Learning objects

The concept of a learning object grew out of the need to manage and reuse digital resources for educational purposes. Downes [2004] argues that the need for learning objects is economic, specifically that it makes no sense for many institutions to spend great amounts of money producing resources that cover the same educational need, when one resource properly shared could meet the educational needs at a much lower cost.

Wiley [2007a] provides a valuable review of the learning object literature, concluding that the literature lacks cohesion and canonical works. The definition of a learning object is contentious [Friesen, 2004; Polsani, 2003; Wiley, 2000a], however, the definition proposed by Wiley [2000b] is commonly cited, “any digital resource that can be reused to support learn-

ing." Some authors simply enter into discussion of learning objects in a manner that suggests the definition should be well known [Boskic, 2003; Brusilovsky and Vassileva, 2003; Futrelle et al., 2001; Heinrich and Chen, 2001], provide a trivial definition [Oldenettel and Malachinski, 2003; South and Monson, 2001], or refer to general characteristics of learning objects without defining them [Downes, 2001; Simon et al., 2003]. Several other terms have been used to describe similar things, such as knowledge objects [Merrill, 2000] and instructional objects [Gibbons et al., 2001].

The term *granularity* is often used in relation to learning objects, however there are different conceptions of what granularity means in this context. The granularity of a learning object is often referred to as its size; however, this is misleading. If a learning object is a digital resource, its size can be objectively measured, in kilobytes for example, which says little about its reusability. The SCORM Content Aggregation Model [SCORM, 2009] discusses learning object granularity in relation to the aggregation of resources, where media assets are combined to form new resources, and these are in turn combined to form resources of higher grain, et cetera. This is an example of a media-centric view of granularity, where granularity is conceived in terms of characteristics of the media used to construct learning objects [Wiley et al., 2000]. An alternative view is that granularity can be measured by the conceptual density of a learning object. This means that an object of small grain has a narrow concept focus, whereas an object that covers several different concepts has a larger grain [South and Monson, 2001]. This is a message-centric view of learning object granularity [Wiley et al., 2000].

To address the problem of designing learning objects of appropriate granularity, Boyle [2003] suggests taking inspiration from the software engineering concepts of cohesion and de-coupling, where coherence implies that each learning object should be based around only one concept and de-coupling implies that learning objects should minimise references to other learning objects. Other metaphors that have been employed to guide learning object design and research include LEGO building blocks [Hodgins, 2001] and atoms [Wiley, 2000b]. However, in what has been called the reusability paradox, the wider the range of educational contexts in which a resource can be used, the less pedagogically useful it is likely to be [Wiley et al., 2004].

In part growing from research into learning objects and inspired by the open source

movement [Wiley and Gurrell, 2009], the open educational resource movement encourages the release of resources, usually on the Web, licensed for free reuse and repurposing, and based on the notion that knowledge is a public good [Smith and Casserly, 2006]. By 2008 more than 6000 courses had been released on behalf of the Open CourseWare Consortium,[2] a worldwide organisation with representatives from hundreds of universities and other institutions that aims to encourage the release of open educational resources and enhance their impact globally [Carson, 2009]. Central to open educational resources is the granting of specific rights in the form of licenses, such as Creative Commons licenses [Bissell, 2009].

2.1.2 User search goals

While examining the use of Wikipedia[3] by students, Head and Eisenberg [2010] found that 95% of students make use of Google[4] to find resources to support their learning. Griffiths and Brophy [2005] found that students overwhelmingly use a Web search engine as the starting point for finding learning resources.

The search goals of users underlie many of the problems in information retrieval, yet it is an area that has not been extensively researched. Broder [2002] presents a taxonomy of web search where queries are classified as navigational, informational and transactional. The taxonomy is supported anecdotally and not empirically, although real queries are presented and classified according to the taxonomy. Broder described a survey delivered to web users that aimed to test the breakdown, but not the validity, of each element of the taxonomy. The survey question design process is not described in the paper, and it is therefore difficult to make conclusions about the rigour of the design process or the validity of the instrument. Aiming to test the breakdown of queries according to this taxonomy, Broder describes an analysis of a log of 400 queries submitted to a live search engine, with the disclaimer that inferring user intent from queries log is at best inexact and at worst guesswork. Queries were classified as either transactional or navigational, with the remainder assumed to be informational, giving a breakdown of 50% informational, 20% navigational and 30% transactional.

Rose and Levinson [2004] take a similar starting point and create a taxonomy of user goals in web search. They also define navigational and informational goals but substitute resource

[2] http://www.ocwconsortium.org/
[3] http://www.wikipedia.org/
[4] http://www.google.com

Table 2.1: Portion of the Search Goal Hierarchy Table by Rose and Levinson [2004] into which searches for educational resources would fit.

SEARCH GOAL	DESCRIPTION	EXAMPLES
Informational	My goal is to learn something by reading or viewing web pages	
1. Directed	I want to learn something in particular about my topic	
1.1 Closed	I want to get an answer to a question that has a single, unambiguous answer.	`what is a supercharger` `2004 election dates`
1.2 Open	I want to get an answer to an open-ended question, or one with unconstrained depth.	`baseball death and injury` `why are metals shiny`
2. Undirected	I want to learn anything/everything about my topic. A query for topic X might be interpreted as "tell me about X."	`color blindness` `jfk jr`

goals for transactional goals and provide a hierarchy for informational and resource goals. Table 2.1 shows a reproduction of the portion of the hierarchy they present. Informational, open directed and undirected search goals are situations in which retrieving educational resources would be appropriate. Approximately 1500 queries were manually classified against this taxonomy. To aid classification, related data were also presented; results clicked by the user and further searches issued by the user. More than 30% of the queries were classified as directed (4.4%) or undirected (26.8%) informational queries. As with the Broder [2002] study, no attempt was made to validate the taxonomy. Validation could have been performed by having queries classified by multiple classifiers and testing inter-classifier consistency measures.

2.1.3 Educational resource retrieval

The dominant technique used for learning object retrieval is through the querying of human-assigned descriptive metadata, and there are few authors who doubt that metadata is critical to learning object retrieval and reuse [Campbell, 2003]. Some authors insist that metadata

must be included for a resource to be considered a learning object [Dalziel, 2002; Polsani, 2003]. Others have bestowed almost magical properties on metadata, such as Hill and Hannafin [2001], who say metadata can "universally signify the nature of their contents and make explicit their attributes for other potential uses."

The use of human-assigned descriptive metadata for retrieval, however, has significant drawbacks. It is expensive to generate and manage [Charlesworth et al., 2007; Hedstrom, 2001], difficult to ensure that metadata will be meaningful [Charlesworth et al., 2007; Najjar et al., 2003], and impossible to guarantee a consistent interpretation of resources over time by a single person, or at any time by a population of people [Smeaton, 2001]. There exists a gap between the interpretation of the person entering metadata, who knows in advance what is in a particular resource, and the person searching for a resource, who only has an information need and who might use an almost infinite variety of search terms [Bates, 1998]. Topic metadata has been shown to be of little value in ranking the results of search [Hawking and Zobel, 2007]. Charlesworth et al. [2007] report that educators are quite reluctant to provide more than the barest descriptive metadata, and that it is futile to try to compel them to create metadata.

Several authors address aspects of improving retrieval from repositories of educational resources. Ochoa et al. [2005] describe a framework for the automatic indexation of educational resources in learning management systems. They propose that the metadata produced by this system could then be searched similarly to human-assigned metadata. Goldrei et al. [2005] describe the harvesting of metadata based on an automatically produced user model of a learning object's author.

Based on the PageRank algorithm for ranking web pages [Page et al., 1998], Duval [2006] proposes LearnRank, the goal of which is to rank retrieved learning objects based on their usage and context. PageRank uses the relationships between Web resources, as described by hyperlinks, to determine which resources are quality resources. The hyperlinks are included in resources as part of the authoring process because the resource author believes there is value in the target page and not as a separate step focussed on retrieval. Using the same underlying idea, the LearnRank of a resource aims to capture how useful people have found it for their learning, without directly asking the learner. For example, LearnRank has been implemented using data derived from how learners interact with resources—or contextualised

attention metadata—in a learning management system [Ochoa and Duval, 2006].

Institutional repositories often expose application programming interfaces (APIs) that allow access to their resources externally. A widely used example is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [de Sompel et al., 2004]. Ternier et al. [2008] propose and describe the ProLearn Query Language, which can be used to retrieve learning objects based on their metadata, or conduct approximate search against resource metadata or indexes constructed from the contents of resources. However, while these protocols use the Internet as their transmission medium, and the resources may themselves be additionally exposed on the Web using the Hypertext Transfer Protocol (HTTP), the search technologies employed are still centred around individual repositories and metadata that provide access to the resources via protocols that are not standard on the Web.

Previous work focusses on searching repositories or federations of repositories of resources that are known to be—or at least are expected to be—educational resources, rather than filtering results returned when searching heterogeneous resources, as can be found on the Web. While many factors contribute to the quality of online learning experiences [Oliver and Herrington, 2003], no absolute measure of the quality of educational resources is possible, regardless of whether the resource is retrieved from a homogeneous or heterogeneous collection. Rather, the learner must make a judgment as to whether resources are appropriate, given the context in which the resource is to be used.

In the next section, to provide background for the evaluation of systems that filter resources, we describe the evaluation of information retrieval systems.

2.2 Evaluating information retrieval

Information retrieval systems attempt to deliver resources to users to satisfy some information need. They generally use some form of natural language query terms, and return results based on relevance. This is in contrast to data retrieval systems such as relational database systems, which are deterministic, have structured query languages, and return data or resources based on matching criteria [van Rijsbergen, 1979].

Baeza-Yates and Ribeiro-Neto [1999] define two information retrieval models; *ad hoc retrieval* and *filtering*, which Belkin and Croft [1992] describe as "two sides of the same coin." The first involves a set of relatively stable resources, where users input a query and

the system returns results based on that query. In the second, new resources are added to the collection and the search criteria, which remain relatively constant, determine whether the new resource should be returned to the user. Croft et al. [2009] additionally describe classification, which is used to automatically assign resources to pre-defined categories, and question answering, which is similar to ad hoc search but addresses more specific questions.

To ensure effectiveness, and to measure potential improvement, information retrieval systems are systematically evaluated [Croft et al., 2009], and this evaluation is based on their ability to return relevant resources [Zobel, 1998]. In this section we describe the ad hoc and filtering retrieval models and the evaluation of information retrieval systems.

Ad hoc retrieval

Ad hoc retrieval is typified by the classic web search engine. A user issues a query to the system, and the system compares the processed query to a previously processed representation of the document set, selecting documents to return to the user based on some matching algorithm [Zobel, 1998]. The query is the user's translation of what may be a vague or ill-formed information need into the format expected by the system [Ingwersen, 1996], a process at which many users are not skilled [Markey, 2007].

In non-trivial search applications, the returned resources may exceed the number that can be effectively processed. It is often useful to manipulate the output, most often by ranking more relevant documents more highly [Sanderson, 2010]. Another manipulation method is clustering, whereby similar resources in the result set are grouped together to try to address the problem of query ambiguity [Kural et al., 1999].

Filtering

Information filtering systems are designed to process incoming resources and either select or remove items based on some either implicit or explicit criteria. The detection of spam email messages in incoming email is a common example of a filtering system [Cormack, 2008].

Malone et al. [1987] identify three different types of information filtering systems: *cognitive*, *social*, and *economic*. Cognitive filtering systems target resources based on characteristics of the resources themselves, such as selecting news items relevant to a user's profile [Morita and Shinoda, 1994]. The focus of social filtering systems is the relationship

between the originator of the resource and the recipient in the context of a broader community. Collaborative recommendation systems, in which ratings are collected from users in production systems and used to suggest resources to other users [Adomavicius and Tuzhilin, 2005], are an example of social filtering systems. Finally, economic filtering systems make a cost-benefit assessment of resources, where the cost may be an explicit monetary cost or an implicit cost such as time or prestige. While Shapira et al. [1999] show that cognitive and social filtering can be used effectively, and that combining them can further improve their effectiveness, research into (or practical systems using) economic filtering is scarce.

2.2.1 Test collections and ground truths

Evaluation of the effectiveness of systems for ad hoc retrieval and filtering is generally carried out by assessing performance on a dataset where each resource has been labelled according to whether it should be returned by the system in response to specific queries [Croft et al., 2009]. This labelled dataset is often called a *ground truth* or *gold standard*.

The ground truth for IR systems evaluation is typically constructed by having relevance judgments assigned to resources by human judges. Thus, systems are measured based on their ability to approximate the human-assigned relevance judgments. This is known as the *Cranfield method* after experiments carried out at Cranfield University in the 1960s [Cleverdon, 1967].

The Cranfield method requires a collection of documents, a set of queries, and a set of relevance judgments linking the documents and the queries, which are often referred to as *qrels* [Sanderson, 2010]. It is the standard approach used in information retrieval evaluation, and is used for evaluation in the Text REtrieval Conference (TREC) [Buckley and Voorhees, 2005], the Cross Language Evaluation Forum (CLEF) [Braschler and Peters, 2002], and has also been adapted for use in other domains, such as for XML retrieval evaluation [Kazai et al., 2003].

Experiments based on the Cranfield method make the assumption that relevance is a property of resources in relation to a query, independent of the user. Under this assumption, the user and the context of retrieval is completely represented by the query [Saracevic, 2007]. While there has been debate about the validity of this assumption, it has been a useful starting point for IR experiments in general [Buckley and Voorhees, 2005].

When a resource is relevant to a topic, the topic has a relevance relationship, through translation into a query, to the resource [Saracevic, 2007]. On the face of it, relevance seems straightforward, however, it is one of the most controversial aspects of information retrieval evaluation [Voorhees, 2000]. For an extensive analysis of relevance with discussion of the pertinent literature, see Saracevic [2007].

In situations where multiple systems are being tested and it is not possible for all resources to be judged, for example in collections of hundreds of thousands of resources, Sparck Jones and van Rijsbergen [1975] propose the use of *pooled judgments*. When pooling judgments, resources returned by each of the systems being evaluated are judged, and all other resources are considered to be irrelevant [Voorhees, 2001b]. While pooling has previously been shown to be a reliable technique for information retrieval experiments [Zobel, 1998], more recent work has demonstrated that it can lead to biased judgment sets as collections become larger [Buckley et al., 2007].

Robertson [1978] describes two assumptions that systems use for ranking result sets. The first is based on the assumption that relevance is a continuous scale, and that resources have a degree of relevance to a given query. The second sees resources as either relevant or not, and systems include resources in the result set based on a probability of relevance. Most experimental evaluation of information retrieval systems use a dichotomous rating system, which emphasises the probability of relevance; a resource is either relevant or not to a particular query [Voorhees, 2001a]. Given the complexity of users' information needs, Järvelin and Kekäläinen [2000] argue that binary relevance judgments cannot reflect the fact that the degree of relevance should be taken into account in system evaluation. As such, systems should be rewarded for ranking highly relevant documents higher.

Measuring agreement

It is common in IR experiments to use a test collection where each resource has been assigned a relevance assessment by a single assessor, and this has been a criticism of experiments based on the Cranfield method [Harter, 1996]. However, this methodology has been shown to be adequate on small collections [Burgin, 1992].

The adequacy of using single assessments for the evaluation of retrieval systems was experimentally supported in relation to the TREC collections by Voorhees [1998], who showed

that, despite variability of individual relevance assessments, the relative ranking of systems is stable. In this work, TREC collections were reassessed by additional judges, and the level of agreement between all judges was measured using *overlap*, the mean of the size of the intersection of positive ratings divided by the size of the union of positive ratings for each resource.

A further simple measure commonly used is *raw agreement*, which is the proportion of observed agreement to possible agreement. Raw agreement measures provide a common-sense value, and reporting only more complex measures of agreement can lead to a failure to adequately communicate practical findings [Uebersax, 2008]. We can think of raw agreement as the probability that if a random resource is selected from a test collection, and we select a random rater who has judged the resource positively, what is the probability that another random judge will agree? If the proportion of negative and positive judgments differ greatly, overall agreement will be biased towards the dominant judgments [Kundel and Polansky, 2003]. This often happens in relevance judgments, where there are likely to be far fewer documents that are relevant to a query (positive) than irrelevant documents (negative). It is therefore also important that the levels of positive and negative agreement are reported separately.

The disadvantages of the overlap and raw agreement measures are that they are not corrected for chance, and it is not possible to estimate a confidence interval [Kundel and Polansky, 2003]. The index κ has been developed to address these issues; Cohen's κ for two raters [Cohen, 1960] and Fleiss' κ for multiple raters [Fleiss, 1971]. κ can be defined as the observed agreement minus the agreement that would be expected by chance, divided by the best possible agreement.

Landis and Koch [1977] developed a table of threshold values that is sometimes used as a way to interpret values of κ. However, the levels chosen are arbitrary, are not applicable across experiments [Sim and Wright, 2003] and can lead to unreliable conclusions [Gwet, 2001]. For these reasons, when reporting κ in this thesis, we do not report our results in relation to the Landis and Koch [1977] table.

While κ is used as a measure of agreement, it is not a test of the effect of classifying resources using two methods. If we want to be able to compare the level of agreement between the two methods for rating or classifying resources, we use Fisher's exact test [Agresti, 1992],

which tests the null hypothesis that there is no difference in the proportions that raters assign resources to different categories under each condition.

2.2.2 Evaluation measures

To assess the effectiveness of retrieval systems, we need to measure their performance. Two basic measures of performance are *recall*, which is defined as the proportion of all relevant documents returned in a result set, and *precision*, which is defined as the proportion of the returned result set that is relevant [Robertson, 1969].

Formally, recall and precision can be defined as follows.

$$recall = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$precision = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

The importance of maximising recall and precision depend on the task. For example, when searching the Web for a fact such as the date the professor's Delorean visited in *Back to the Future*, precision is more important than recall; we only need retrieve one document with the correct answer. If we want to learn in depth about the physics of time travel, it is likely that we would want to consult multiple sources, and so recall becomes more important. Finally, if we want to file a patent for a time machine, we need to find all relevant prior art, so recall must be maximised [Wallis and Thom, 1996].

At times a single summary measure is useful. The harmonic mean of recall and precision, or F_1, is based on the more general F_β originally developed by van Rijsbergen [1979], and is defined as follows.

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

The most commonly used summary measure of ranked retrieval systems is mean average precision (MAP), a measure first introduced by Voorhees [1993] which assumes that resources are returned by a system based on their computed probability of relevance. The average precision for a query can be calculated by taking the precision at each rank where a relevant

document is returned by a system, and dividing by the total number of relevant documents. MAP is then the mean of scores across all topics or queries.

An alternative measure for ranked retrieval systems is discounted cumulative gain (DCG), which aims to take into account the degree of relevance of a resource in relation to a query. This measure is based on the premise that highly relevant resources are more useful higher in ranked results and should therefore be penalised if they appear lower [Järvelin and Kekäläinen, 2002].

2.3 Machine learning in information retrieval

So far we have covered how to evaluate information retrieval systems that return resources, but we have not addressed how those systems decide what resources to return. Machine learning is one method that can be used to differentiate relevant from irrelevant resources. Machine learning involves taking features derived from a set of example resources and using a machine learning algorithm to create a model that is capable of making predictions of some kind about new resources [Witten and Frank, 2005]. Machine learning algorithms are often called induction algorithms or inducers.

There has been a long history of employing machine learning techniques in information retrieval, especially for information filtering applications. For example, machine learning can be used to classify incoming email as spam or legitimate email [Sahami et al., 1998].

Schaffer [1994] proved that it is impossible to produce a perfect machine learning algorithm, an algorithm capable of creating a classification model that outperforms all other models on all data inputs. As there is no universally superior machine learning model, we must evaluate models within specific domains, and therefore a method is required to compare models. When testing classifiers, it is necessary to quantify performance with a summary statistic of some kind. We would also like to be able to say whether a particular classifier significantly outperforms another.

2.3.1 Evaluation measures

In evaluating machine learning classifiers, we are interested in how well they perform on new data. As we cannot directly measure future performance, a method is needed to estimate future performance based on test data.

A classifier is produced by running a machine learning algorithm over a set of labelled training data. Testing the resulting classifier on a set of labelled test data yields a measure of the performance of the classifier. This set of labelled test data is directly analogous to the ground truth in information retrieval evaluation. There are many measures that attempt to quantify the performance of a classification algorithm, providing estimates of the performance of the classifier on unseen data.

The simplest measure is *sample accuracy*, which is the proportion or percentage of the test instances that were correctly classified. This can be used to estimate the underlying and unmeasurable *true accuracy* of the classifier. The term *generalised accuracy* is also used in the literature to refer to true accuracy, and its complement, *generalised error* is used to refer to the estimate of the unmeasurable true error of a classifier. Where we use the term *accuracy* in this thesis, we are referring to sample accuracy.

Using the classification of spam email as an example of a two-class problem, a legitimate email that is correctly classified is called a *true positive* (TP), and a correctly classified spam email is called a *true negative* (TN). A spam email that is incorrectly classified as legitimate is called a *false positive* (FP), and a legitimate email incorrectly classified as spam is a *false negative* (FN). Thus, if P is the total number of positive classifications and N is the total number of negative classifications, accuracy can be defined as follows.

$$accuracy = \frac{TP + TN}{P + N}$$

The information retrieval measures of recall and precision are also commonly used in machine learning. If we let the set of TP and FN be the relevant resources, and the set of TP and FP be the retrieved resources, we can redefine recall and precision as follows.

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

Taking recall and precision from above, we can now also redefine F_1 in terms of TP, FP, and FN as follows.

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Measures should also be sensitive to incorrectly classified resources. If we create a classifier that maximises the number of true positives, such a classifier is unlikely to be fit for its purpose due to ignoring errors. A classifier can perform poorly on one class and still achieve a high percentage of true positives, especially if there is an imbalance of class instances. For example, if most of the resources in the test data are educational, classifying all resources as educational maximises the number of true positives, but every non-educational resource would be a false positive. Therefore, in addition to maximising the true positive rate, it is useful to minimise the error rates, where instances are incorrectly classified.

Accuracy will give an inflated indication of performance where classes are imbalanced, and also treats all errors as equally important, whereas in a particular setting a false negative may be much more costly than a false positive or vice versa [Provost and Fawcett, 1997]. Another weakness of this measure is that it does not account for the level of success that would be achieved by chance; that is, by classifying test instances randomly.

As with the agreement measures discussed in Section 2.2, when evaluating classifiers, the κ statistic can be used to account for the agreement expected by chance, and it has been shown to be more realistic than accuracy [Ben-David, 2008a].

In their critique of accuracy maximisation for comparing classification algorithms, Provost et al. [1998] suggest researchers use receiver operator characteristic (ROC) analysis in its place. Originally developed in the field of signal detection theory [Green and Swets, 1966], ROC curves have been shown to be effective for classification measurement in machine learning [Flach, 2003]. ROC curves are constructed by stepping through the test instances sorted in order of the probability that the instance belongs to a particular class, as produced by a classification algorithm, and graphing the number of true positives (on the Y axis) and false positives (on the X axis) at each instance. This gives a stepped graphical representation of the trade off between benefits and costs and allows for the visual comparison of classification models [Fawcett, 2004].

While it is useful to be able to visualise this trade off, it is often valuable to have a single measure. The area under the ROC curve (AUC) is a suitable single measure [Bradley, 1997], and Sokolova and Lapalme [2009] showed that AUC was an appropriate measure for document classification. It has the added attraction that it can be interpreted as a probability that a classifier will rank a random positive instance above a random negative instance [Witten and

Frank, 2005]. The AUC is equivalent to the p statistic of the Mann-Whitney U test [Mason and Graham, 2002], and is often calculated using this test.

ROC curves and the κ statistic, though their origins are quite different, have been shown to cover closely related concepts [Ben-David, 2008b], though the κ statistic is not cost sensitive [Witten and Frank, 2005]. A thorough discussion of machine learning statistics can be found in Powers [2007].

The behaviour of many machine learning algorithms can be altered through the manipulation of input parameters. To avoid bias or overly optimistic results, any tuning of the parameters of an induction algorithm should be performed on a separate tuning data set, prior to gathering the final test data [Salzberg, 1997].

2.3.2 Resampling

The ideal way to evaluate a classifier involves having access to large amounts of data with class labels, and using most of the data for classifier training, leaving aside a portion as test data. Dividing the available data in this way is known as the *holdout method*. However, the production of collections of labelled data is often expensive and time consuming, so it is rare to have a situation in which unlimited data are available. To make the most of available data, it is common practice to perform *resampling*, dividing and reusing the available data in different ways. The three most commonly used resampling methods are *leave-one-out*, *bootstrap*, and *cross-validation* [Witten and Frank, 2005].

In *leave-one-out*, one instance at a time is removed from the training data used to build a classification model, and a performance measure is obtained using the removed instance as test data. This process is repeated for each instance in the dataset, and the performance measure is averaged. As a classification model is built for every instance in the dataset, leave-one-out is computationally expensive. Additionally, Efron [1983] has shown leave-one-out to have unacceptably high levels of variance.

Using *bootstrap*, in a dataset with n instances, we construct training data by randomly choosing instances, with replacement, n times. As we have sampled with replacement, the resulting training data will contain multiple copies of some instances and some instances will not have been chosen. The data that do not occur in the training set are then used to test the resulting classifier. However, for some problems, bootstrapping has been shown to have

large biases, making the performance measures unreliable [Kohavi, 1995].

Cross-validation involves segmenting all data into k random folds of roughly equal size, and hence it is often called k-fold cross-validation. The classifier can then be trained on all but one of the folds, with the remaining fold being used as test data to obtain a performance measure. This is repeated k times, so that each of the folds has been used once as test data and $k-1$ times as training data, with the overall performance measure being the mean of the performance measures for each run.

To obtain stable estimates, it is common to apply k-fold cross-validation repeatedly and take the average of the results, with 10 times 10-fold cross-validation considered standard [Witten and Frank, 2005]. Stratified cross-validation, where folds are stratified so that the class proportions in each fold are representative of the entire dataset, has been shown to be an effective method of reducing the variance of measure estimation [Kohavi, 1995].

Figure 2.1 shows a collection of instances being divided into 10 stratified folds. In 10-fold cross-validation, each fold is then used as a test set with a classifier trained on the remaining folds, as shown in Figure 2.2, which produces a summary measure. This process is then repeated 10 times, as shown in Figure 2.3, and the summary measures are averaged.

2.3.3 Comparing classifiers

To be confident that the differences observed between classifiers are not due to chance, a method for testing significance is required. To test statistical significance, an accurate measure of variance is required [Nadeau and Bengio, 2003], and that variance should be taken into account when choosing a statistical test. The random selection of data used to train a machine learning algorithm and the random selection of data to test the trained classifier are sources of variation, and there is often variance associated with randomness internal to the induction algorithm itself [Dietterich, 1998]. Variance internal to the induction algorithm can be reduced through repetition, as described above [Witten and Frank, 2005].

Just as performance estimates are more accurate when there is sufficient data to use separate data for testing and training, the ideal method of comparing two classifiers is to split the available data and use the holdout method to obtain unbiased estimates of mean and variance for each classifier according to our chosen measure [Bouckaert and Frank, 2004]. However, as previously discussed, in most cases the scarcity of data means that some form

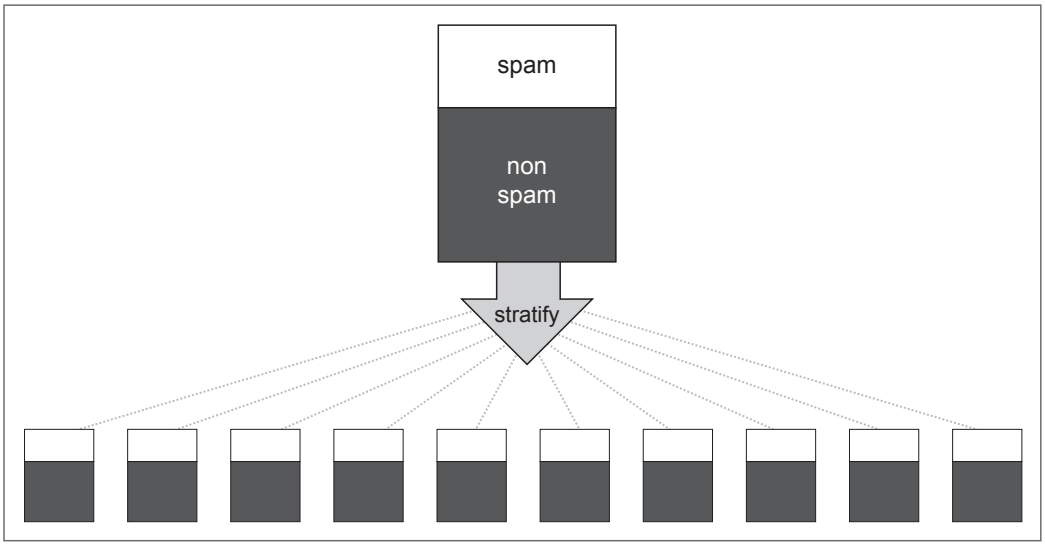


Figure 2.1: During stratification, a set of instances is divided into folds, with each fold having a class distribution approximately the same as the overall dataset. This figure shows 10 folds, as would be used in 10-fold cross-validation.

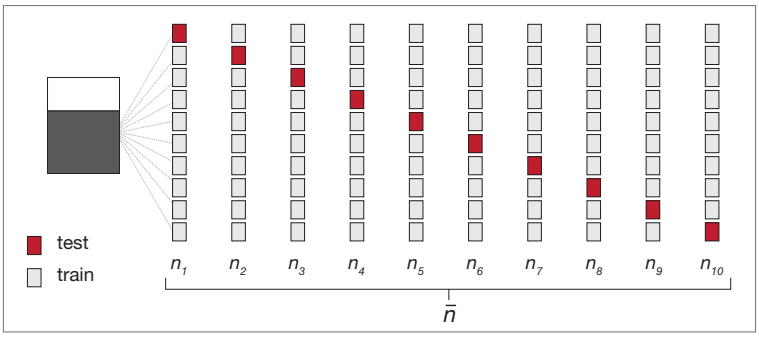


Figure 2.2: Each of the 10 stratified folds is used as the test dataset with a classifier trained on the remaining nine folds.

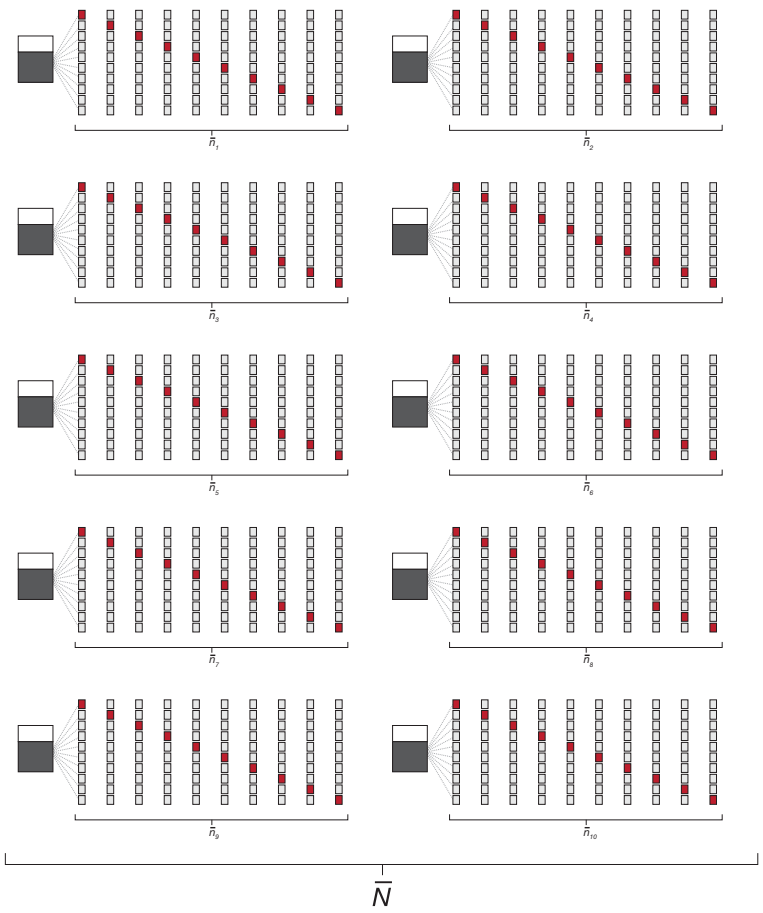


Figure 2.3: Cross-validation is repeated to produce a stable performance estimate.

of resampling will be required to obtain stable performance estimates.

When comparing classifiers, care must be taken that the experiment design and statistical methods lead to valid statistical conclusions [Salzberg, 1997]. For example, if a test is to be performed across multiple runs of classifiers where the datasets are not independent, the significance test and level should be appropriately adjusted to avoid Type I error, erroneously finding a significant difference where none exists. This is a problem where resampling is used, as resources will be common across multiple datasets, so any assumption of independence will be broken.

Dietterich [1998] investigated the appropriateness of five statistical tests for comparing classifiers. Two of the tests, McNemar's test and a test for the difference of two proportions, use the holdout method and do not allow for resampling and thus reduce the amount of data that can be used in training, making them inappropriate for domains where limited data is available. The remaining tests investigated all use resampling of some kind. The resampled paired t-test, in which multiple trials of classifiers are performed where the available data is randomly divided into training and test sets, was shown to have an elevated chance of Type I error due to the dependence of both training and test data. Though the test sets are independent, the k-fold cross-validated paired t-test also showed a risk of Type I error due to dependent training sets. Dietterich [1998] then introduces the 5x2cv test, which performs five 2-fold cross-validation trials and manipulates the t statistic to achieve a more stable variance estimate, but this has the disadvantage of reducing the data on which classifiers are trained, and has been criticised as heuristic and lacking statistical rigour [Nadeau and Bengio, 2003; Witten and Frank, 2005].

Nadeau and Bengio [2003] investigated the variance introduced by the training data under a variety of testing schemes that use resampling, including a corrected version of the resampled t-test. They showed that their corrected test introduces a more reliable estimate of variance than the 5x2cv, and this test is recommended by Witten and Frank [2005]. Bouckaert and Frank [2004] investigated the sensitivity of several statistical tests to the variance introduced by the random partitioning of training and test data when resampling is used, which they define as replicability. They showed that the corrected t-test designed by Nadeau and Bengio [2003] using 10 times 10-fold cross-validation showed high levels of replicability, and therefore recommend the use of this test.

2.4 Summary

In this chapter, we surveyed the background literature on educational resources, information retrieval, and machine learning that we build on in the remainder of this thesis.

We began by describing the use of digital resources in education, and research into their management and reuse, a theme that we investigate in Chapter 3. We also outlined current research into the retrieval of educational resources, the topic of the remainder of this thesis.

We introduced models of information retrieval and the methodology used for the evaluation of information retrieval systems, including measurements of rater agreement. In Chapter 4, we build on this background and describe how systems that filter educational resources should be evaluated, specifically using the agreement measures described in this section to assess whether a single assessor is reasonable for this evaluation. We report agreement in terms of overlap, raw agreement, and κ.

We introduced machine learning, describing the appropriate measures and techniques to use when assessing the performance of classifiers, and significance tests for comparing that performance. In Chapter 5, we describe various machine learning induction algorithms, and develop machine learning models to classify Web resources as educational or non-educational. When comparing classifier performance, we use stratified 10 times 10-fold cross-validation and consider AUC as the most important measure. Additionally, we report accuracy, recall, precision, F-measure, and the κ statistic. In Chapter 6 we test whether those models perform significantly better than a baseline classifier using the corrected 10 times 10-fold cross-validated t-test.

Chapter 3

Management and Reuse of Educational Resources

In an information retrieval system, success is achieved if resources returned are useful or relevant to a user [Baeza-Yates and Ribeiro-Neto, 1999]. To identify what might be useful or relevant in a given domain, we must have an understanding of that domain. Successful retrieval of educational material requires an understanding of how the resources being searched are managed, how searchers interact with those resources and the systems that manage them, and the needs of the people searching for resources.

This chapter describes our investigation of the management of digital material for teaching and learning, which was completed in two stages. First, we investigate how organisations currently manage educational resources. Second, we explore in greater depth how academics and teachers use and reuse educational resources.

The following section presents our research methodology for answering these research questions. In Sections 3.2, 3.3, and 3.4 we present the research procedure and data analysis for the first stage of this work, focussed on resource management. In Section 3.5, we present the second stage, concerning resource reuse. We draw together the themes that emerge from this research in Section 3.6, which informs our work in following chapters.

3.1 Research methodology

Qualitative research methods are appropriate for exploring complex domains, such as how educational resources are managed, used, and reused, including the interaction between the individuals and systems involved. Among the most commonly used methods for qualitative data collection are interviews, focus groups, and surveys [Kayrooz and Trevitt, 2004]. We use each of these methods, as collecting data from a variety of methods is useful for triangulation, which reduces the risk of systematic biases or limitations of methods and helps to verify the validity of results [Maxwell, 1994].

3.1.1 Resource management

In the first stage of our research, the aim of which was to broadly understand how educational resources are currently used and managed, data was collected using focus groups, interviews and a survey.

Focus groups are useful for exploratory data collection [Fontana and Frey, 2000], and were therefore a logical starting point for our work. They encourage themes to emerge through the interaction of participants [Kayrooz and Trevitt, 2004]. Additionally, we use the focus groups as part of the survey design process, of which they are often an integral part [Fowler, 2002]. In Section 3.2, we describe how focus groups were used to investigate educational resource management issues.

Interviews are particularly useful in complex domains [Kayrooz and Trevitt, 2004], making them appropriate for use in the first stage of our research. We used interviews for description and interpretation, as outlined by Peshkin [1993], investigating the processes, settings and systems involved in the management of educational resources. Section 3.3 presents our discussion and analysis of interviews with domain experts.

To avoid biasing results in our preliminary investigations, we did not use pre-determined codes when analysing transcripts of interviews and focus groups conducted as part of stage one. Rather these were analysed for themes and categories using the editing method, which Kayrooz and Trevitt [2004] describe as, "the text is examined from the point of view of an editor, searching for categories and/or themes until the text is reduced into a summary distillation."

As well as being analysed directly, issues raised in the focus groups and expert interviews

were used to inform the construction of a wider survey of a sample of staff from RMIT University. Surveys can be used to collect data from a wide sample, to learn about the overall population [Leedy and Ormrod, 2005]. The results of our survey of academics and teachers are presented in Section 3.4.

3.1.2 Successful reuse

The second stage of our research, which focussed more closely on educators who had successfully reused educational resources, is presented in Section 3.5. The interviews conducted in this stage were used to encourage reflection and establish informative generalisations about how resources are used and reused. Our general approach followed that recommended by Leedy and Ormrod [2005], identifying key questions in advance, and using them to form an interview schedule to guide the interviews.

Following Miles and Huberman [1994], a provisional set of codes to be used for analysis was produced prior to the interviews. These codes were developed concurrently with the interview schedule and questions, and were based upon a subset of specific themes that emerged during stage one of our research. The codes are described in Section 3.5.

3.2 Focus groups

To identify broad issues regarding the management, use, and reuse of learning objects within RMIT University, we ran four focus groups with three to ten people in each group. Participation in the focus groups was voluntary.

3.2.1 Procedure

To ensure coverage of a broad range of issues, it was decided to run two sets of focus groups. For the first two focus groups we aimed to recruit RMIT staff with an awareness of or involvement with learning objects; these groups were dubbed *aware*. The participants were staff who had been involved in projects supporting the reuse of digital material for teaching and learning.

The remaining two focus groups, dubbed *naïve*, were made up of RMIT staff with teaching experience who had not been involved with learning objects projects. Participants' experience

or knowledge was not specifically tested before being invited to participate, so the level of awareness or naïvety of group members varied.

From the aware participants we expected to find out how educational resources were being managed at that time, and get higher level views of the successes and failures of various management projects. From the naïve participants we sought to gain insight into how people use and reuse resources, and how they might like to in the future, without the bias that might arise from being involved in a particular project.

The first aware group, with ten participants excluding the researchers, was the Project Reference Group of the Reusable Learning Objects (RLO), a project that aimed investigated how RMIT University should become engaged with learning object research and implementation. Members of the Project Reference Group also recommended the five participants in the second aware group. The naïve focus groups had three and five participants, excluding the researchers, and were also recommended by members of the Project Reference Group.

Potential participants were emailed an invitation to participate in a focus group, presented in Appendix B.1, which contained questions for the focus groups, with the questions and email text differing between the aware and naïve participants. The email also included a plain language statement, attached in Appendix B.2, giving the background, context and goals of the research, and a description of the focus group process. Participants were required to sign an informed consent form indicating that they had read the plain language statement, they understood their rights, and were willingly involved in the research.

Discussion in the focus groups was led by either the author of this thesis, the other project research officer, Henric Beiers, or both. They were based on the initial questions, but were largely unstructured. Transcripts of the focus groups were produced from audio recordings, and the transcripts analysed using the previously described editing method.

3.2.2 Aware focus groups

The fifteen participants across the two aware focus groups included academic and teaching staff from each of the three academic portfolios,[1] from both higher education and voca-

[1]When this research was conducted, RMIT University was divided into organisational units called *portfolios*, three of which were academic. The three academic parts of the university have since been renamed *colleges*, while the remaining non-academic parts retain the designation portfolio.

tional education and training, educational designers and developers, and senior information technology staff.

Questions to guide the aware focus groups, sent to participants before the focus group was conducted, covered the participants' knowledge of projects related to the use of learning objects, systems to support such projects, and how RMIT University could most productively support the field of learning objects.

The following sections describe the themes that arose from the aware focus groups, with some illustrative quotes from participants.

Motivations and current situation

Discussions in the aware focus groups were wide ranging and generally positive about the possibility of improving the management of educational resources, though many difficulties and barriers were raised. None of the participants expressed any reluctance to contribute educational resources they had prepared, though there was a recognition that others may be reluctant to do so.

The picture that emerged of how educational resources are managed was one of individual silos of resources. The university-wide system for managing digital assets was predominantly used to store resources for display on the university web site, and was regarded as deficient for centralised management of educational resources. Several participants described smaller scale resource management activities, in units that serve the whole university, an academic portfolio, and a school.

Two main motivations for better management of educational resources were raised. The first was to provide a centralised and easily searchable location or portal through which staff could access intellectual property owned by RMIT University, licensed third-party learning objects, and other external resources. The second was to reduce duplication of effort and allow educators to use their time more effectively by not having to recreate resources that already exist.

Participants noted that many academics already used digital resources produced by others. Publishers of texts used as required reading in classes often provide resources online, and participants had both used these resources directly and had contributed resources to publishers. Participants also spoke of formal reuse, where "modules" were reused between

different courses, and informal reuse, where resources were stored on a networked drive and used by multiple teachers.

Several participants at the portfolio and school levels said that a common problem was when a resource was used in multiple offerings of the same course. They wanted to be able to update a single master copy of a resource and have that update propagate through other versions. The ability to reuse resources in different modes and contexts with the same content, without having to have multiple copies, was attractive to them.

It was recognised that other institutions were grappling with the same issues in regards to the sharing and reuse of resources, and the experiences they had shared should not be ignored.

What resources should be managed

Three dimensions emerged to the issue of what resources should be managed: granularity, quality, and interoperability.

There was no clear agreement about the granularity of resources that would be useful to manage centrally. Opinions ranged from a belief that it was not worth managing small resources such as images, to wanting small resources but only if one was able to easily distinguish individual resources from larger topics, to a desire to store larger topics or even whole courses as long as they can be disaggregated into smaller elements.

There was also disagreement as to the quality assurance requirements of managed resources. On one hand was the belief that resources should be rigorously quality assured and that not to do so ran the risk of having the system populated with resources of little use, making it more difficult to find worthwhile resources. On the other hand, there was concern that having a high level of quality assurance may hinder the iterative improvement of resources that might arise as a result of a resource being used in multiple contexts by different people. Given that successful informal sharing and reuse has already been observed, support for less formally quality assured resources was seen as important.

To enable the widest possible reuse of resources, one participant emphasised that resources should be standards based. However, it was recognised that reuse cannot be achieved simply by adhering to a standard format, and several participants said reuse would be difficult to achieve unless there was facility to customise resources. One participant cautioned against

introducing a purely technology-driven resource management solution, echoing the findings of Charlesworth et al. [2007] described in Chapter 2.

Finding appropriate resources

Having a system with effective search capability was seen as critical, and allowing educators to effectively and efficiently find suitable resources was a recurrent theme, supporting the findings of Littlejohn et al. [2003]. It was evident that previous systems implemented within the university, both for managing educational resources and for other tasks, led to poor user experience and created a lack of engagement from staff. The need for a simple and effective user interface for any software system that supports the reuse of educational resources was highlighted by several participants. One participant said they felt it was important that it was easy to interact with search results and view resources returned in response to a search, so that staff did not have to spend a lot of time searching.

Other mechanisms for finding appropriate educational resources were also raised, with Amazon's product recommendations and user reviews being cited as an example by more than one participant. Participants felt that comments from educators about how they had used a resource and whether they found it useful could be valuable, and such comments should be recordable in a management system and accessible to searchers. This suggests the collaborative evaluation tools proposed by Nesbit and Belfer [2004], and discussed in Section 2.1, would be useful.

It was noted in one focus group that many staff already use Google to try to find resources on the Web.

Culture and workflow

Several participants expressed the opinion that if reuse was to become widespread, significant cultural change was required, and stressed that institutions need to place great importance on supporting academic staff to encourage commitment to that change. Despite the technical challenges that were identified, several participants maintained that if the most important factors related to the sharing of resources were seen as technical rather than organisational or cultural issues then any effort to change the status quo would be wasted, similar views to those found by Charlesworth et al. [2007].

> "***We need to look at the patterns of what people already do. That will make it easier to add electronic learning objects into their normal patterns of behaviour.***"

Advertising the existence of any system for managing learning resources was identified as an important initial step to encouraging adoption. Further, training in the use of software to support sharing processes was seen as crucial. One participant was particularly concerned about training, having felt that this had been neglected in similar ventures and was hopeful that training would not be provided only as an afterthought. Providing professional development and assistance in the use of a new system, one of the key areas of complexity identified by McNaught [2003], had support in both focus groups.

While participants discussed widespread informal sharing of resources within departments, to be successful, they stated that contribution to a system designed for sharing must be integrated into the existing workflows of staff. This was seen as an important aspect to managing the balance between adopting innovation and increasing staff stress, another area of complexity discussed by McNaught [2003].

> "***If academics aren't of a mindset where they're going to share, then we're going to be here in 10 years time doing the same thing.***"

Further to incorporation into existing workflow, participants suggested that an important incentive to contribute resources would be to have such contributions recognised for purposes such as promotion, much in the same way that research output is recognised. Part of the issue was seen to be that research outputs such as papers and grants were seen as a measure of success, whereas production and contribution of high-quality educational material was not directly measured. While the production of such resources could affect career progression through improved teaching scores, there was no formal recognition of contribution of resources. The need to create incentives for the contribution and reuse of educational resources similar to those that exist in relation to research has been raised by several authors [Campbell, 2003; Koppi and Lavitt, 2003; McNaught, 2003].

Drivers identified for the desire to contribute resources included opportunities for the

iterative improvement of resources and the kudos associated with producing high-quality material that is reused by others. This kudos was compared to the prestige achieved by publishing research work. Participants noted that the primary use of learning objects is in the classroom, and any release of these materials for reuse by others is additional secondary work, in contrast to research publication output.

Even after successful contribution of resources, several participants believed that it would be difficult to encourage some teachers and academics to use material created by other people due to a belief that they can produce better resources themselves or that the nuances of the content would not match what they thought was ideal. One participant recounted situations where course content was completely rewritten repeatedly in a short period of time when the academic responsible for teaching the course changed.

Rights

> "***It's kind of pointless going down this path if everyone's still a bit confused about who owns what.***"

Supporting the findings of Charlesworth et al. [2007], further areas of concern raised by participants were intellectual property, moral rights, and copyright. A system for the management of intellectual property issues was seen as essential for the reuse of learning objects, as was an appropriate framework for moral rights attribution of learning objects submitted for reuse. Digital rights management issues were raised in each aware focus group.

> "***People need to understand what they're entitled and what they're not entitled to do with these objects.***"

Participants expressed concern that a system to share learning resources may lead to the abuse of their moral rights. Moral rights belong to the creator of a resource, ensuring they have the rights to be properly attributed for work that they have created and that their work is not treated in a derogatory manner. Several participants stated this concern as a fear that they would lose control of resources submitted for reuse, specifically in that the resources may be altered in

ways that the original creator did not approve. While most participants did want to be recognised as the authors of resources they contribute, it was felt that as resources changed over time, possibly away from the original intent of the work, this attribution may be negative.

Distrust was also expressed, in that there was a concern that an institution would attempt to sell the resources for profit, with little recognition going to the original creator. Participants said they would want to know how resources were being reused and in what circumstances, as there was a fear that students may see the same resources in different but related courses if this was not managed.

Funding

One of the reasons identified in the focus groups for better management of digital resources for teaching and learning is to more easily reuse resources in different contexts. Participants raised the fact that universities are complex institutions and it is likely that different teaching organisational units will have separate funding. They expressed concern that where there was no system in place to facilitate reuse of resources across organisational units, negotiations would have to occur to ensure that the originating unit was adequately compensated, and that these negotiations would likely be complicated and conducted on a one-off basis. The need to enter into such negotiations to get approval for sharing material between schools, as well as the lack of an appropriate funding model for this sharing, were seen as significant organisational barriers to reuse across the university.

3.2.3 Naïve focus group

The two naïve focus groups had three and five participants, excluding the researchers, and included academic and teaching staff from each of the three academic portfolios, from both higher education and vocational education and training, and educational designers and developers.

The guiding questions included in the naïve focus group invitation email addressed the participants' reuse and contribution of resources, their knowledge of similar teaching areas within the university, as well as issues of staff support and system requirements. Participants raised many of the same themes as had been raised in the aware focus groups. While many of the participants in the aware groups were less involved with teaching and more with the

management of others who teach, the participants in the naïve groups were generally focussed on their own teaching. The following sections describe the themes that arose from the naïve focus groups.

Contributing and reusing resources

Widespread willingness was expressed to contribute resources for other people to reuse and to engage in contributive activities, both within and between portfolios. Several of the participants said they felt the same willingness also existed among their colleagues.

> "***What credit does a lecturer get out of putting up learning objects? There needs to be a reward as well or you just wouldn't do it.***"

However, as with the aware focus groups, there was recognition that goodwill was insufficient and that more direct incentives would be required to encourage contributions. A system of recognition modelled upon research output was again suggested, as were work load recognition and direct financial incentives. One proposed model of financial reward was similar to royalty schemes for artists, where contributors are rewarded with a nominal amount each time a resource they contributed was reused by others.

> "***There is resistance from a lot of people to willingly give up work that they have developed without some sort of recognition.***"

The other aspect important for a system to manage educational resources is the willingness of academics and teachers to reuse resources contributed by others. Again, participants said that reuse will not happen without motivation, and that educators would have to know that reusing resources would save them time, that they would have access to a reasonable range of resources, and that they could alter at least some of those resources to meet their own needs.

What resources should be managed

As with the aware focus groups, there was no consensus among participants in the naïve focus groups about the quality of material that should be managed. Some participants felt that all material in a managed system should be quality controlled, others believed works in progress could encourage collaboration, and that some "junky stuff" would be fine. One focus group discussed the possibility that a quality review could be carried out as an informal peer review performed by colleagues teaching in the same area as the contributor.

Finding appropriate resources

In support of our aware focus group findings, participants said they viewed search as key, and that for reuse to be successful, finding relevant and appropriate resources to reuse should be as painless as possible. Aspects of search that were raised included ease of use, effectiveness and efficiency.

> **"*I think [efficient search] is a really key thing.*"**

One focus group supported the idea of being able to search for related content. For example, if a resource had been used successfully, other related resources might also be useful. This led to the idea of being able to subscribe to the system, so that people could be notified when new material is added that is related to resources that they have previously used.

> **"*You'd better have a bloody good search engine with it.*"**

Rights

Perhaps the greatest amount of concern expressed in the naïve focus groups was in relation to intellectual property and moral rights. This reflected much of the discussion in the aware focus groups, but there was a higher level of distrust expressed.

Some participants felt it was important that they were able to retain control over resources they had submitted for reuse. Tracking the usage of resources was seen as important by several participants. The main reason for wanting usage tracking was so that educators could

avoid using resources that were likely to have already been seen by students in prerequisite courses.

The moral rights of educators were also discussed, with one participant likening changing his material to debasing a piece of art. Another was concerned that authorship attribution be preserved.

> ***"I wouldn't mind who used whatever I developed. Don't have a problem. But if they were pretending it was theirs then I've got big problems."***

Participants feared that the university would see a collection of resources as an asset to be sold or traded. There was a distrust of the motivation of the university management in promoting sharing, with several participants expressing fear that an organisational push to encourage reuse was a rationalisation exercise and likely to be a precursor to job losses. Some participants expressed doubt university management understood that the value added by educators in teaching cannot be contributed along with resources, exactly the concern expressed by Noble [1998] and further explored by Twigg [2000].

Opportunities for collaboration

One of the complexities raised by Littlejohn et al. [2003] related to the reuse of educational resources was the need to develop communities of practice. There was sustained discussion in one of the naïve focus groups about the opportunities for collaboration afforded by a system facilitating sharing of learning resources, and the possibility of developing such communities. One participant was enthused by the possibilities for informal collaboration that a formal system of sharing might allow to develop.

Discussion about searching for appropriate resources led to the suggestion that a system for managing learning resources should allow for requests for resources to be created, which in turn could lead to collaboration on resource creation.

3.3 Expert views

Following the focus groups, we interviewed both internal and external people who had experience running projects focussing on the management of educational resources. From internal

university staff we aimed to gather perspectives from as many different areas as possible, to clarify the situation as it stood at the time and past experiences within RMIT University. From external experts, our goal was to gain insight into the experiences at other institutions with regards to learning objects, including specific enterprise-wide projects. We also sought to allow other important issues to be raised.

3.3.1 Procedure

Interviewees were selected because of a recommendation from the RLO Project Reference Group or another interviewee, or because of their involvement with a particular project or organisation. They were approached either through email or by phone and given an outline of the project, and no set contact text was used. Participants read a plain language statement and signed the interview consent form prior to being interviewed. As with the focus groups, the plain language statement outlined the background, context and goals of the research, and described the process of the interviews. It is provided in Appendix C.1.

Interviews were conducted by either the thesis author, the other project research officer, Henric Beiers, or both. Interviews were unstructured, as this is the most appropriate form of interview for this type of exploratory task [Kayrooz and Trevitt, 2004].

In total 16 interviews were conducted, 11 individual interviews, three interviews with two participants, and two larger group interviews with people from a single organisational unit, with seven and nine attendees.

Complexity

A theme that was raised repeatedly in the interviews was that enterprise-wide management of educational resources is extremely complex, and should not be embarked upon lightly, echoing the findings of Littlejohn et al. [2003]. Thus, institutions should acknowledge the risks and approach projects to manage resources with care. One participant pointed out that this complexity, and hence the financial risk, is often underestimated. Another interviewee also cautioned that institutions be aware that a return on investment would not necessarily be simple, and that considerable reuse would be necessary to amortise the costs of the system and its ongoing maintenance.

> **"*That didn't really prepare me—despite the reasonably thorough background in the area—for the complexities that we were going to need to unravel.*"**

The complexity is exacerbated by the fact that there are generally a large number of stakeholders in such a project. As pointed out by one interviewee, stakeholders generally approach projects from their own point of view, which can lead to stakeholders seeing things from an educational development point of view, from a technical point of view, from the point of view a user of a system, or from a student point of view, leading to a lack of communication and difficulty establishing a common vision for a resource management project. Overcoming these difficulties in communication was seen as extremely important, however, interviewees warned that resistance was inevitable.

People may also come from different theoretical backgrounds, making communication even more difficult. After questioning whether the term "learning object" was useful, one interviewee warned against tying a system to a particular theoretical context, and that it was important to try to keep the system pedagogically neutral.

Several participants echoed the views expressed in the focus groups that although it may be tempting to see issues as being of a technical nature, most of the concerns raised were organisational and cultural. One interviewee warned that it was a real risk that if a reuse project is driven by technology, the system may meet no-one's needs.

Given the complexities and risks involved, it was suggested by several interviewees that it was worthwhile running a pilot or trial project to try to work through issues.

Organisational motivation

Given the complexity of managing educational resources, several interviewees said organisations should be clear about their motivations for embarking on management projects and realistic about the prospects of meeting their goals. One participant warned that organisations should not simply follow what other institutions are doing or what people perceive the institution should be doing. The same participant went on to say that organisations also need to be clear about how such projects should be evaluated.

> ***"We know we've got for instance close on 100,000 files ... of which I wouldn't be surprised if the duplication is 50 per cent."***

Reuse of resources between staff was cited as a driver by several interviewees, however, some interviewees expressed doubt that such sharing is likely. An alternative motivation, the better management of resources used by individuals in different locations and modes of delivery, was mentioned more than once as an important goal. Managing resources on a small scale was seen by one interviewee as the start of wider reuse, suggesting that the initial focus could be to help individuals manage resources across the courses they teach, then be expanded to include a few people using the same resources, then to sharing across schools or departments.

Managing resources offered in courses in a diversity of cultures offers challenges that are not insignificant, according to one participant, who believed that improved management of resources across cultures in a manner consistent with the best principles of transcultural education was a problem worth solving.

A significant attraction of resource reuse mentioned by several interviewees is the reduction in duplication of various kinds. For example, better management of educational resources could decrease the number of times a file has to be stored. An interviewee who was responsible for negotiating the purchase of resources from external providers saw an educational resource management system as a way to stop different parts of the organisation buying the same resource, a situation they had seen many times.

Contributing resources

Encouraging content producers to contribute resources is fundamental to the success of a system to share and reuse educational resources, as McNaught [2003] points out. There are two parts to unravelling this issue; understanding the motivations that encourage people to contribute and understanding the barriers that stop them. Interviewees did talk about situations in which people had expressed reluctance, or even outright hostility, to the idea of contributing resources they had created, but they also found that some were enthusiastic.

> ***"I'm an educator and education is about getting education out to as many people as you can. I actually find it abhorrent that people won't share."***

Interviewees said they felt some educators saw their value as being keepers of knowledge, and the idea of sharing was a threat to their academic identity. One interviewee recommended that the people who were unwilling to share should simply be ignored and focus should be placed on supporting the positive contributors.

According to one interviewee, if the main perceived benefit of sharing is to help others, large-scale sharing is unlikely to occur, given the existing pressures that teachers and academics are already under. Other interviewees said that, from a practical point of view, systems should not make the task of contributing resources onerous, from either a technical or a bureaucratic point of view.

Several of the interviewees had been involved in relevant pilot programs involving academic and teaching staff who were generally self-selecting. It was felt that these early adopters were important to the success of any project. They had various motivations in being involved. Many had an academic interest and felt that they could produce research output based upon their involvement. Being able to speak theoretically about learning objects was generally seen positively, however, in one project that brought together people from around an institution, this meant that planning in the pilot was often side-tracked by theoretical discussion, to the detriment of actual contribution.

> ***"Is that the only way to value teaching, to write a paper about what you're doing?"***

The issue of the recognition of the importance of teaching and learning was again raised. There was a repeated perception that teaching is not recognised as an equally important activity as research in most institutions, especially higher education institutions. It was the common feeling that whereas writing a paper for a refereed journal is rewarded, good teaching is not, particularly the development of good teaching materials. Some saw the introduction of teaching awards as one small step towards recognition of teaching. However, they also commented that unless budget allocations and promotions equally

rewarded good teaching and the development of quality teaching resources there would be little incentive for staff to spend the time necessary to create and contribute quality learning objects. It was suggested that the most effective change that could be made to encourage resource contribution would be to actively recognise it as a criterion for promotion.

Part of the value in contributing resources was seen to be the exposure that contributors could gain, both internally to their institution and externally if resources were exposed to Web search engines.

What resources should be managed

In the expert interviews, most of the discussions about what resources should be managed revolved around quality. Some interviewees had been involved in projects that mandated a high level of quality assurance, while others were involved in projects that had no quality review at all. In the former, there was a tension between the high overhead of contributing resources and the desire to be guaranteed that anything in the system would be usable. In the latter, there were a large number of resources that were not educational, or even useful, at all.

Interviewees commented on the fact that quality review processes are already embedded in the processes around course design and approval, and that to require another quality approval process at the level of individual resources would be redundant. The resources would effectively be quality reviewed whenever they were actually included as part of a course.

Several interviewees suggested the possibility of having a repository that allowed for two tiers of resources to be stored. One tier would be quality reviewed while the other would be for resources that were in development or versions of reviewed resources that had been altered and not yet reviewed again.

In a project described by an interviewee, educators expressed reluctance to accept the need for quality review. There were two aspects to this issue. First, people were unsure as to whether they could trust that reviews were done fairly by qualified reviewers and, second, that people were personally and professionally threatened by the prospect that reviewers would be unhappy with the quality of resources. The interviewee suggested that professional development activities might be able to allay some of those fears.

Versioning of managed resources was discussed by some interviewees, specifically in terms of managing iterative changes over time, which one interviewee pointed out was an active area of research and existing commercial interest, and in terms of managing slightly different versions of a resource being used in different offerings. In one pilot project, solving the latter problem was a big attraction to the academics involved.

Finding appropriate resources

> ***"If it doesn't function as an effective retrieval tool then we're going to have problems."***

The ability to retrieve useful resources was seen as critical to the reuse of digital resources. A successful search will result in staff finding resources that will be of use in their own teaching. Interviewees discussed various aspects of this issue.

Some interviewees had used systems with no or very poor indexing or cataloguing of resources, and they expressed frustration at not being able to find resources easily. Several interviewees felt that the requirement to input descriptive metadata for retrieval was redundant, as modern search engines were capable of indexing the content of resources.

A simple search interface was seen as essential by many of the interviewees. The user interface provided for Google Search was mentioned more than once as an exemplar. Google was also mentioned as the model by which to measure resource retrieval effectiveness.

Finally, take up of a repository will depend in part on there being enough material in it to enable people to have a good chance of conducting a successful search.

Metadata

Several interviewees commented that metadata requirements must be kept to a minimum and that, where possible, metadata should be harvested automatically. Interviewees said that as metadata entry is costly and time consuming, there should be clear reasons for mandatory metadata, whether it be for search and retrieval, interoperability, or research, and this should be balanced with an understanding of the effort required to enter it. Several discussed how much people dislike entering metadata, and that if the value of the metadata was not clear, it was likely that any metadata entered would be of poor quality.

In one interview, participants said that in their institution each faculty has one or more information specialists providing e-learning support to academic staff who assist in the loading of metadata records. They either enter the metadata for the academic staff or train academics how to enter the metadata.

Culture and workflow

As with the focus groups, cultural change was identified by interviewees as an important issue, likely the most important. This relates to an element of complexity identified by Littlejohn et al. [2003], that of balancing innovation with the possibility of increasing stress levels. One interviewee said that in their experience some people were keen to share educational resources, but others were quite "fearful".

> **"*Anything that happens needs to be introduced in a way where there is choice and where it wins support because it's able to demonstrate that it's valuable.*"**

Several interviewees pointed out that reuse of educational resources was in practice already common, and that educators reuse text books and other content provided by publishers, as well as articles and case studies. However, people were not as used to sharing and reusing digital resources.

Comments were made by interviewees on the tendency within their institution to look for "the quick fix" and, as a consequence, to mandate procedural change rather than facilitate incremental cultural change. Several interviewees stated that the implementation of a shared teaching resources system should promote usage by staff based on education, persuasion and facilitation rather than the introduction of enforced targets or procedures, and that the enforcement path fails to acknowledge the complexities inherent in attitudinal and cultural change.

> **"*You will just have to cope with the fact that some people share, some people don't.*"**

Interviewees identified situations and organisational units where the existing culture does allow and even encourage sharing of resources. The tertiary and further education sector

(TAFE), also known as vocational education and training (VET), was cited several times as being more likely to encourage sharing. This supports the findings of Charlesworth et al. [2007], who found that institutions that focus on teaching over research were more likely to share. It is perhaps significant that ventures aimed at the VET sector, such as resources managed through AEShareNet[2] and the Flexible Learning Toolboxes,[3] are already in existence, whereas in the higher education sector there are no widely available shared teaching resource banks in Australia. Several explanations were put forward as to why this was the case. One interviewee believed there was less criticism of teaching resources produced by VET teachers. Several interviewees made the supposition that teaching loads and casualisation influenced the willingness of VET staff to reuse and share educational resources, as these added to staff time pressure. However, one interviewee cautioned against generalisations, suggesting that while situations allowing for sharing may be more common in VET, personality was the most important issue.

Another organisational area that was reported as having a significant level of sharing and reuse was a medical school. In one interview, it was reported that teaching staff in the medical area are happy to use materials developed by others, and that reuse is expected. One reason cited was the support provided to create educational resources, and therefore more acceptance of using resources created by others. This was contrasted with other areas in the institution that had a strong tradition that people could only be sure that resources were of high quality if they had created them.

> "***Having some monolithic system plonked on top of everyone and being told, 'There you go, use it,' is not really a sensible approach.***"

The multimedia production units that are responsible for the creation of resources for others were also identified as having a culture predisposed to the sharing of their output.

Mandating the use of a system for managerial purposes was seen as detrimental to its chance of success, and management support should be perceived to be working towards solving issues of teaching and learning. Most interviewees were aware of aspects of people's workflows that should integrate with resource

[2] http://www.aesharenet.com.au

[3] http://toolboxes.flexiblelearning.net.au/

management systems, and that any system should solve problems people have now, being available at their point of need. To do this, the system should be enmeshed with other services so that people will be able to find it through their current normal behaviour. However, interviewees pointed out that there was a danger in coupling systems too strongly; as a new system should be independent enough that a system for managing resources could allow groups to control how they use it rather than pre-specifying how it should be used.

Speaking about a previous project to share and reuse resources that was perceived to be a failure, one interviewee felt the system did not encourage meaningful reuse because no culture had developed around the system. Another problem raised is that because large-scale management of digital educational resources is new to many enterprises, many people do not have the vocabulary to talk about the issues involved.

> "***Just having the best system doesn't mean it is going to be taken up.***"

Rights

As with the focus groups, issues of intellectual property, moral rights, and copyright were raised repeatedly. In regards to moral rights, a balance between a creator's right to be acknowledged and the subsequent modification or contextualisation of resources was seen as important and difficult to achieve. One interviewee with experience in copyright clearance discussed the issues of copyright and moral rights attribution. As is common in higher education institutions, the interviewee's university holds the copyright of any material produced by its staff, however the moral rights stay with the creator unless specifically signed away. This has an impact on the right of the original creator to be acknowledged as the creator in both the original and subsequent versions, and on the right of the creator to have some level of control over subsequent modifications to resources. Both of these interact with the need to maintain version control so that such modifications can be tracked.

> "***Staff are aware that their materials are used without attribution or without being acknowledged or even contacted for reuse of the material.***"

The same interviewee also discussed difficulties associated with hosting external material in regards to the licensing agreements controlling their use, stating that there is a need to ensure that any external material included in teaching resources is managed within the terms of the original agreement for the use of that material. The perceived danger was that staff mistakenly assume that because they have a license to use material for one purpose they therefore have a license to use it for another purpose. Some changes are obvious enhancements, and are unlikely to be an issue, but it is up to the originator to make the judgment as to whether a change is an enhancement or a mutilation. A further issue raised is that a resource may lose its applicability not through being changed but because time has passed and it goes out of date. Automatic or semi-automatic resource expiry was suggested to avoid some moral rights issues.

Collaboration and community

In describing the complexities involved in the reuse of resources, McNaught [2003] discussed the fostering of communities of practice. The possibility that a system to reuse educational resources, through its propensity to encourage sharing, could also encourage collaboration between academics, between developers, and between academics and developers, was mentioned by many of the interviewees. They perceived the potential for collaboration as being one of the side benefits of being able to identify the creators and other users of resources that they were interested in. Interviewees also mentioned that it would be desirable to have a group of people who were in effect part of the repository system. These people could be like liaison librarians in the library, keeping track of people's teaching interests and be able to act as signposts directing others towards potential collaborators.

3.4 Surveying academic and teaching staff

Following preliminary analysis of our focus group and interview data, we undertook a survey of RMIT staff, to test the extent to which issues raised were important to a wider population of staff involved in teaching. Further input on the construction of the survey was taken from the RLO Project Reference Group.

In addition to seeking to collect the opinions of subjects regarding the use, reuse and storage of learning objects, other classificatory information was collected, including:

- School/Portfolio;
- Length of teaching experience;
- Amount of experience with technology-supported education;
- Modes of teaching experience currently and in the past.

3.4.1 Procedure

The survey was presented in four sections, addressing the following areas.

Section A General background information of the respondent.

Section B Questions about reusing resources contributed by other university staff.

Section C Questions about contributing resources for use by other university staff.

Section D Questions about the technical features that a hypothetical system for the management of educational resources should provide.

In Sections B, C, and D, respondents were asked to respond to questions using a five-point scale; *Unimportant*, *Somewhat Important*, *Important*, *Very Important*, *Vital*.[4] Each section provided space for further comment regarding the issues raised in the questions. The full survey is presented in Appendix D.3.

The survey was presented on RMIT University's web site, with responses collected anonymously. Respondents were asked to read the plain language statement, shown in Appendix D.2, followed by a consent form on which they had to click "I agree" prior to accessing the survey. The survey was only accessible after consent had been given.

To encourage the greatest level of response, several different methods were used to inform staff of the survey, mirroring a common practice with mail surveys [Fowler, 2002]. The weekly internal email newsletter, RMIT Update, was used to advertise the survey. Members of the RLO Project Reference Group approached staff directly, either face to face, by phone or by email, and encouraged them to complete the survey. Additionally, reference group members

[4] In hindsight, the scale may have been more useful if it had a neutral mid-point.

requested key Teaching and Learning personnel within their respective areas to encourage other staff to complete the survey.

The university is a dual-sector institution, and the 52 survey respondents were involved in teaching and learning in either a higher education (HE) setting, or a vocational education and training (VET) setting. Due to a technical fault, the results for sections B, C, and D were not collected for 9 of the respondents, though comments were still recorded and included in our analysis. Additionally, there were 11 instances where a respondent did not respond to a question.

Respondents to the survey were drawn from both the Higher Education (35 respondents) and VET (12 respondents) sectors of the university as well as staff involved in training from the Library (4 respondents) and ITS (3 respondents). Respondents had between 1 and 33 years of teaching experience, with a mean of 10.8 years and a median of 9.5 years.

Points on the five-point scale were weighted from zero for unimportant to four for vital. We then took the sum of the weighted responses and divided by the total number of valid responses to obtain a level of importance associated with a question. Importance was calculated overall, as well as being broken down by the respondents' organisational group.

We present the results for each section of the survey and discuss our findings. Graphs of the breakdown of results on the response scale are provided in Appendix D.4.

3.4.2 Using resources created by others

The responses to questions about using resources created by others are shown in Table 3.1. Average responses for each question fell in the range of somewhat important to very important, with none reaching the extremes of unimportant and vital. Knowing if the resource had changed, and being able to use and change the resource without restriction were the most important aspects of reusing resources. Being involved in a quality review was of the lowest importance.

Several commenters on Section B of the survey noted that if a resource were to be changed, the original author should still be acknowledged, suggesting at least an implicit understanding of moral rights issues. There were also respondents to the survey who commented that it did not really matter whether a particular object had been quality reviewed, as any such quality review was likely to be fairly general, and they would perform their own quality

Table 3.1: Section B: Importance associated with survey questions regarding the use of resources created by others, sorted by importance.

Question	Importance
4. Know if the resource had been changed?	2.786
5. Be free to use the resource as is without restriction?	2.744
6. Be free to change the resource without restriction?	2.619
1. Know who created the resource?	2.381
3. Know how others had used the resource?	2.093
7. Know that the resource had undergone a quality review?	2.047
10. Have the opportunity to annotate the resource for the benefit of others?	1.860
8. Have access to a quality review regarding the resource?	1.744
2. Know who else had used the resource?	1.674
9. Have the opportunity to participate in a quality review of the resource?	1.209

review to determine if it was the appropriate for inclusion in their own courses. In reference to annotation, one respondent expressed concern that annotations should be positive and constructive, while another suggested user ratings, such as those used by Amazon. One respondent expressed the view that reuse should be voluntary, and that no-one should be forced to use resources created by others. Two of the comments suggested it was very unlikely that they would ever use resources created by someone else. Finally, one respondent said it was difficult to answer questions about conditions imposed upon someone wanting to reuse a resource in a general fashion, and that they would be more willing to abide by restrictions for high-quality resources.

3.4.3 Contributing resources for use by others

Average responses to questions about contributing resources for others to reuse, shown in Table 3.2, were in a similar range as questions regarding reuse. Being acknowledged as the creator of a resource, both in the management system and in its subsequent use, was viewed as among the most important aspects, and this was supported in the comments. Several of the comments talked about having experienced lack of appropriate acknowledgement. In addition to being acknowledged for the contribution of resources, several commenters suggested the

Table 3.2: Section C: Importance associated with survey questions regarding the contribution of resources for use by other staff, sorted by importance.

Question	Importance
10. Know of changes made to the resource?	2.651
1. Be acknowledged as the creator of the resource in the storage system?	2.628
2. Be acknowledged as the creator of the resource in its subsequent use?	2.419
14. Be personally rewarded through your workplan, promotion, awards or other mechanism for the use of the resource?	2.233
4. Know who uses the resource?	2.186
7. Know how the resource is used?	2.186
3. Be acknowledged as the creator of the resource if it was subsequently changed?	2.093
17. Know if there was a quality review of the resource?	2.050
18. Have input into a quality review of the resource?	2.025
11. Set general conditions on how the resource can be changed?	1.884
5. Set general conditions on who can use the resource?	1.721
15. Have your group/school/portfolio financially recompensed for the use of the resource?	1.581
8. Set general conditions on how the resource can be used?	1.558
12. Control on a case-by-case basis how the resource can be changed?	1.488
16. Be free to share a resource without it undergoing a quality review?	1.476
9. Control on a case-by-case basis how the resource can be used?	1.262
13. Be personally financially recompensed for the use of the resource?	1.093
6. Control on a case-by-case basis who can use the resource?	1.000

idea that contribution should count towards promotion.

Controlling the use of the contributed resource on a case-by-case basis was among the least important issues, with one commenter suggesting that such control was "petty". Personal financial compensation was also among the least important issues.

Knowing if the resource had been changed was seen as important. The point of this question was to ascertain whether people would want to know if another person had modified a resource that they had contributed. However, comments suggest that some people interpreted the question to mean that the originally contributed resource would be lost. As this question is ambiguous, the results are unreliable. One commenter discussed the moral rights issues in relation to changes, specifically the difficulty in maintaining acknowledgement for a resource that had been changed significantly.

Several commenters suggested that knowing who had used a resource might help to identify possible collaborators for future projects, while others talked about the possibility that information about changes made to a resource might suggest ways to iteratively improve it.

In relation to quality review, some commenters dismissed the idea of sharing resources without a review. One commenter said they would need to be assured that those performing a quality review were competent to do so. Finally, one commenter said that since the resources they develop would only be of interest to a small group, they were much more likely to share informally.

3.4.4 Using a computerised system for the reuse of resources

> "***Search is the key feature of this sort of repository.***"

Average responses to questions about the features that a system to support the reuse of resources should have are shown in Table 3.3. Average responses for these questions were higher than the previous sections, clustering around very important. Having a simple user interface was the most important, followed by having a fast and efficient search capability.

Most of the comments focussed on search, and many specifically mentioned that the search functionality for a system for the management of educational resources should be

Table 3.3: Section D: Importance associated with survey questions regarding the use of a computerised system for the reuse of resources, sorted by importance.

Question	Importance
6. Has a simple user interface?	3.279
4. Has a fast and efficient search capability?	3.163
7. Has a consistent look and feel?	3.023
1. Is able to search on key words?	2.953
5. Has an Advanced Search feature?	2.907
2. Is able to refine a previous search?	2.814
3. Has a find more like this function?	2.651

like Google, and one specifically suggested purchasing a Google search product. Ease of use was the other significant theme in comments, with one saying they would be unlikely to use the system at all if it had not been designed by human computer interaction experts. Having an Advanced Search feature, which would unavoidably include a more complicated user interface, was seen as less important than being fast and efficient.

It was suggested that the driver should be making the system effective for the user, rather than as a tool for compliance or management. Commenters said that the system should integrate with existing workflow. One person commented that metadata entry requirements should be kept to a minimum.

3.4.5 Types of resources

Respondents were asked what type of resources they were thinking of while answering the questions in the survey. A range of types of resources were identified. Some were static objects such as images, slides and lecture materials, while others were non-interactive objects such as animations, audio and video. Activities that a student could engage in, such as quizzes and exercises, were mentioned, as were interactive tools and activities. At the most interactive level, some respondents were thinking of forums or chat rooms, though it is unclear whether this means reuse of computer applications providing this functionality or reuse of the content added to them. Several respondents were considering a combination of many of these resources.

3.5 Successful reuse interviews

In the second stage of our work, we interviewed academics and teachers who had successfully reused educational resources, with the aim of examining problems they faced in their reuse, but also to look at the things that helped them, and how they would like to retrieve and reuse educational resources in the future. We also investigated the types of resources people were able to successfully reuse.

3.5.1 Procedure

The subjects interviewed were academic staff of RMIT University teaching in either VET or in higher education. Potential interviewees were identified through discussions with members of groups involved in the production, use, reuse and management of reusable learning objects. To attempt to identify individuals from as broad and representative a base as possible, three groups were asked to nominate staff for interview: one was a university-wide unit supporting development of educational resources, another was a unit supporting online learning across an academic portfolio, the last was a unit supporting online learning in a school within a different academic portfolio. These groups were identified as being involved in the production, use, reuse and management of educational resources in the first stage of our research.

As much as was possible, interviews were conducted in quiet locations where interviewees were comfortable, in most cases in the participants' own offices. The process used to select interviewees is described in detail below.

Participation in the interviews was voluntary and confidential. Confidentiality was important for ethical considerations, as participants had been nominated by people who may have been their superior, and likely had an interest in the successful outcome of the research project, and known refusal to participate had the potential, however remote, to damage that relationship.

Potential interviewees were emailed an invitation to participate in the project, which is shown in Appendix E.1. This email also included the interview schedule, with questions, and a plain language statement, presented in Appendix B.2, describing the research, including a brief description of prior work, the context and goals of the current work, and a description of the interview process. Participants were required to sign an informed consent form indicating that they had read the plain-language statement, they understood their rights, and were

willingly involved in the research. Ten interviews were conducted from the pool of suggested interviewees. The interview schedule is presented in Appendix E.2.

The interview questions used to guide the interviews covered the sorts of resources that the interviewee had successfully reused, how they reused them, and how they would like to reuse resources in the future. The interviews were semi-structured, as the questions were treated as a set of topics to be discussed, rather than to be delivered verbatim and in order [Robson, 2002]. This method allows for thorough exploration of opinions and experiences [Gall et al., 1996]. Every effort was made to establish a rapport with the interviewee, and to make them comfortable, as is vital in semi-structured interviews [Fontana and Frey, 2000].

There were several reasons why suggested interviewees declined to be interviewed, including lack of availability and a belief that they were not appropriate candidates for interview.

This latter reason exposed a difficulty in our experimental design, that of identifying academics who reuse in the absence of a system supporting reuse. In some cases individuals initially believed they did not reuse material. However, after informal discussion in which incidences of reuse were mentioned, they realised they did in fact reuse. In other cases, the individual was indeed not an appropriate interview candidate. In all cases these potential interviewees had been identified by the university level group. While the other two groups work closely with academics on a regular basis, the university level group responds to specific requests from throughout the university. Hence there are less likely to be close ongoing relationships between the university level group and the academics for whom they produce material. While members of this group were able to say they produced material for a particular academic, they were not able to say whether that academic had made the resource available for others to reuse, or even whether the academic had used it themselves.

3.5.2 Analysis codes

A set of analysis codes, developed concurrently with the interview schedule and questions, was produced prior to the interviews. These were based upon a subset of themes that emerged during the focus groups, survey, and earlier interviews. Four high-level code groups were used: *what*, to code passages about what resources were reused; *how*, to code passages about how those resources were reused; *why*, to code passages about the needs the teacher or academic was attempting to address when reusing the resources, and how they might

translate them into search terms; and *want*, to code passages about how people want to find and reuse resources. Each high-level code was divided into sub-codes. The codes and sub-codes are presented in Table 3.4. The analysis codes were used to organise segments of the transcripts into themes.

Themes that emerged from the interviews are presented below, using the high-level analysis codes for organisation. The few passages that were coded using the high-level code *why* were not helpful in furthering our understanding of the reuse of educational resources, so there is no corresponding section for that code.

3.5.3 What resources were reused?

An understanding of the types of resources that academics reuse, and the difficulties that they faced reusing them, is essential for judging the effectiveness of future learning object retrieval research. Therefore, interviewees were first asked questions about the sorts of resources that they have reused, including the characteristics, file types and granularity of the resources.

Interviewees reported the reuse of a wide range of resources. Most common were sets of slides, either using Microsoft PowerPoint or LaTeX, and varying in length from approximately 10 to 60. Several interviews also reported successfully reusing web resources, from a single page to complex web sites, such as one representing an imaginary company. Other resources reused were Macromedia Flash objects, video, journal articles, quizzes, and assignments.

Resources that were reused were developed by the interviewees themselves or on their behalf by a university production group, a colleague previously responsible for a course, or had been bundled with the recommended textbook for the course, or had been retrieved from other third-party sources. Most of the reuse reported was of learning objects that focus on a single topic, though some interviewees reused significantly larger resources. The granularity of desired results ranged from a single resource on a single topic to an entire semester's material, as well as much in between.

Some participants had tried to reuse resources available from third parties but were unsuccessful because of the difficulty in extracting useful parts.

The reuse interviews also support our findings that there are differences in attitude between those involved in HE and those involved in VET. It seemed that material of a larger granularity was reused in VET. Some HE staff expressed exasperation at the unwillingness

Table 3.4: Codes used in the analysis of transcripts of interviews with internal university staff who had successfully reused educational resources.

Code	Sub-code	Used to code sections that cover ...
what	gran	the granularity of reused resources
	ft	the file types of reused resources
	drm	issues relating to digital rights management, such as licensing, intellectual property, and moral rights
	type	more general issues about the characteristics of reused resources
how	offer	issues relating to the offering in which the resource was reused
	found	how the reused resource was found
	context-item	how the reused resource was contextualised
	context-offer	how the offering was contextualised to accommodate the reused resource
	support-got	support the interviewee received in finding and reusing the resource
	support-want	support the interviewee had wanted when finding and reusing the resource
	diff-tech	technical difficulties encountered when trying to reuse the resource
	diff-cult	cultural difficulties encountered when trying to reuse the resource
	diff-org	organisational difficulties encountered when trying to reuse the resource
	diff-ed	educational difficulties encountered when trying to reuse the resource
	diff-drm	rights-related difficulties encountered when trying to reuse the resource
why	info	information need the interviewee was trying to address
	terms	how the interviewee might translate the information need into search terms
want	ui	views about the user interface requirements of a system to manage resources
	gran	the granularity of resources the interviewee would like to use in the future
	ft	resource file types the interviewee would like to use in the future
	drm	interviewees' views about digital rights issues in regards to their future reuse of resources
	type	more general discussion about the type of resources the interviewee would like to use in the future
	support	the nature and amount of support the interviewee would like to receive in reusing resources
	context	the contexts in which the interviewee would like to reuse resources

of their colleagues to share material, or to reuse material that others have made available. In support of findings from the first stage of our research, interviewees surmised that the greater acceptance of sharing may be a result of the high number of contact hours and extensive use of sessional employees as VET teaching staff. Additionally, VET courses are competency-based, where the competencies are mandated by government agencies, so there is a constrained range of topics to be covered.

3.5.4 How were resources reused?

Information about how resources are reused successfully may assist in making judgments about the sorts of resources that are appropriate for retrieval. Additionally, information about the difficulties academics faced, and the support they received in their attempts to reuse, are important for the development of systems that support reuse. Therefore, interviewees were then asked about how the resources were reused, including the course offerings they were reused in, what contextualisation was necessary, and what difficulties they encountered and support they received when attempting to reuse.

Material was reused in a number of different ways. Academics often manipulated or contextualised third-party resources or had the material developed specifically for them. Respondents reused material in offerings of the same subject targeted at different student cohorts or run at different times. Some reuse also involved the conversion of the material to different formats.

The level of support received varied between respondents. Groups exist within some academic units whose sole purpose is to assist academics to develop and reuse digital material. Suggestions for the reuse of material often came through these groups, and difficulties in reuse were absorbed by them rather than passed on to the individual academic. In another case, while appropriate content was suggested by a central university source, the interviewee was left to manage the conversion of the material to an appropriate format. The interviewees suggested by the business school group expressed a high level of satisfaction with the level of support they received in their attempts to reuse learning objects.

Most interviewees felt that it was important to be able to make changes to learning objects before they reused them. These changes include contextualising resources for integration with existing courseware and altering to suit the desired topic depth or educational level. Several

interviewees said they would not use learning objects that they were not able to change. On the other hand, two interviewees wanted to reuse material with no or minor adaptation and were not concerned with the contextualisation of the resource.

One of the difficulties mentioned by interviewees is related to their ability to modify resources prior to reuse. In several cases the material that academics wanted to reuse was in a format that was difficult to manipulate or edit. Academics did not want to be required to use specialist software or have specialised skills to be able to contextualise learning objects.

While only two of the interviewees were VET staff, resource reuse was taken as a given, with much ad hoc sharing happening in their academic units already. For example, most academic units provide a shared network drive where teachers can store their material, with the understanding that resources should be used in different subjects where possible.

3.5.5 How do people want to reuse resources?

As findings from the focus groups and interviews indicated, poor user interface design can hinder the acceptance of new systems. It is difficult to establish the champions needed to make a project successful if users are put off by the system interface. Therefore, participants were asked to reflect on their experiences and discuss how they would like to reuse resources in the future.

Interviewees overwhelming wanted the user interface of a repository to allow for key word searches. When asked what sort of user interface they would like to use to find learning objects, many interviewees said they would prefer a Google-like interface, with a simple text box. Three interviewees said they would like to additionally be able to browse the resources using a topic hierarchy. Two interviewees stated that while they supported the idea of reuse, they were unlikely to access material in a general repository of learning objects.

3.6 Summary

In this chapter we presented our exploration of issues surrounding how institutions manage and how individuals and groups successfully reuse educational resources. We used different modes of qualitative exploration to triangulate and to verify our findings.

We ran two sets of focus groups to identify broad issues regarding the management, use, and reuse of learning objects within RMIT University. The participants of the first two focus

groups were staff who had been involved in projects supporting the reuse of digital material for teaching and learning, while the remaining two focus groups were made up of staff with teaching experience who had not been involved with learning objects projects.

After the focus groups, we interviewed people who had experience running projects focussing on the management of educational resources, both internally to RMIT University and externally. In total 16 interviews were conducted with 33 participants, who were approached on recommendation or because of their involvement with a particular project or organisation.

We then undertook a survey of RMIT staff to test the extent to which issues raised in the focus groups and interviews were important to a wider population of staff involved in teaching. The survey presented questions about reusing resources contributed by other university staff, about contributing resources for use by other university staff, and about the technical features that a hypothetical system for the management of educational resources should provide.

Finally, we interviewed educators who had successfully reused educational resources to examine problems they faced in their reuse, to look at the things that helped them, and investigate how they would like to retrieve and reuse resources in the future.

Our results provide strong evidence that there is substantial disagreement about the quality of resources that should be managed in a system to encourage reuse, with some people insisting that all resources should receive quality review before being shared, and others wanting more informal sharing. The latter appears to contradict the findings of Campbell et al. [2001] that educators are only willing to reuse resources that are quality reviewed, though this is impossible to confirm without analysis of actual working systems. There was little discussion about who should be responsible for quality review if it was mandated.

The need for incentives both to reuse resources contributed by others and to contribute resources was raised in multiple data collection modes. As McNaught [2003] points out, research is seen to be more highly regarded than teaching, which was a concern to many participants. Some methods to encourage contribution of resources that have not been discussed in the literature were suggested, such as a royalties scheme. A significant barrier to contribution was perceived to be the complexity associated with moral rights attribution.

The importance of an effective search mechanism for educational resources was discussed

by participants in every mode of data collection used. Google was suggested many times, both as an exemplar of ease of use, simplicity, and effectiveness, as well as directly as a tool to search for educational resources. These findings were critical in our decision to investigate the effective retrieval of educational resources from the Web, the focus of the following chapters, rather than retrieval from closed repositories of learning objects.

There are several aspects to the research described in this chapter that limit the generalisability of our findings. First, data collection was constrained to educators and experts in Australia, so it may be that results would be different internationally. Second, the pool of eligible respondents to the survey was restricted to one university, and within that pool the response rates were low. Though we attempted to reach as wide an audience as possible by asking reference group members to encourage people to undertake the survey, we are unable to say with certainty that the respondents, who were self-selected, were representative.

In the following chapter we present our work developing a methodology for evaluating the effectiveness of systems that filter educational resources from results returned when searching the Web.

Chapter 4

Evaluating Effectiveness of Educational Resource Filters

As discussed in Section 2.1.3, previous research on the effective retrieval of educational resources has assumed that all resources in the collection being searched are learning objects. However, many educational resources are released on the World Wide Web, and clearly there is much more on the Web than just learning material.

We showed in Chapter 3 that when teachers and academics want to find digital material to support learning they prefer to use a public search engine, such as Google. Similar findings have been reported in relation to the information seeking behaviour of students [Griffiths and Brophy, 2005]. This suggests that users searching for educational resources—whether teachers, students, or people learning informally—may be more satisfied with search engine results if only resources likely to support learning were presented. To meet the needs of these users and provide more satisfactory result sets, resources that are unlikely to support learning should not be present. A filter to detect material that is likely to support learning is therefore needed.

We propose that, using the Cranfield method for the evaluation of information retrieval systems as a base, systems that filter learning material can be evaluated based on their ability to select those resources that have been categorised by human judges as educational. To carry out this investigation, the development of a ground truth is required.

In IR systems, the ground truth is constructed by assessing relevance, a concept which

is complex and multi-dimensional [Saracevic, 2007]. For filtering educational resources, the ground truth requires assigned judgments of whether resources are educational, which is also a complex concept. To make it easier for people to make judgments, when collecting ratings we use the phrase "likely to support learning" to mean that a resource is educational.

In Chapter 2 we mentioned the assumption necessary for experiments based upon the Cranfield method that relevance is a property of resources in relation to a query, independent of the user. In this work, we make a similar simplifying assumption, that a resource can be judged educational or not, independent of the specific educational context. Though context obviously plays an important role in education, we believe making the assumption of context independence when making judgments is appropriate for the development of an evaluation methodology for retrieving educational resources.

This chapter explores two issues related to the construction of collections appropriate for the evaluation of systems that filter web resources to identify learning material. First, given that the concept "likely to support learning" is not precise, complete agreement between judges rating resources according to that concept is unlikely. We investigate whether people can broadly agree on what resources are likely to support learning in the face of this ambiguity and complexity. Second, we examine whether displaying the query used to retrieve a resource, which is necessary in relevance evaluation experiments, influences judgments and affects agreement.

The remainder of this chapter is organised as follows. We begin in Section 4.1 by describing the parts of a system used for building a ground truth. In Section 4.2 we describe a user experiment to assess the level of consistency of human judgments of whether web resources are educational. In Section 4.3 the calculation of agreement evaluation is detailed. The results of our user experiment are analysed in Section 4.4. We then discuss in Section 4.5 what these results suggest about the ability of judges to agree on whether a resource is educational, and what impact this has on an evaluation methodology for systems that filter educational resources.

4.1 Elements of a system for building an evaluation collection

There are four elements in a model for building a collection based on the Cranfield method: resources, queries, judges, and the judgment subsystem. We briefly introduce these ele-

ments in this section, and give concrete examples when describing our experiment design in Section 4.2.

Resources are categorised by judges, and are one part of the input to retrieval systems. The resources to be judged are usually be retrieved by issuing keyword queries to one or more search engines. These queries should be representative of likely real queries in the domain under investigation. For example, IR experiments often use queries from query logs, as these queries can be thought of as a real user's attempt to distil their information need into a query.

Judges are the people who classify resources, providing the relationship or concept assigned to the resources that represents the ground truth. In the case of classic IR collections, that relationship is one of relevance between a query and a resource. In the context of the current work, that relationship is between a resource and whether the resource is educational.

The final part of the model involves how the judges assign the relationship in which we are interested. We will call this the judgment subsystem, and it includes methodological considerations such as the instructions given to judges, and the number of categories into which the judges categorise resources.

In the next section, we describe our experiment design, taking the Cranfield method as a starting point for assessing agreement when judging whether resources are educational.

4.2 Experiment design

Our proposed method for the evaluation of systems that filter e-learning material differs from evaluation of relevance using the Cranfield method in that the ground truth it seeks to collect involves classifying resources according to a concept (supporting learning) as opposed to drawing a relationship (relevance) between a query and a document.

In most cases, the number of judges and the judging time available will be constrained. Therefore, consideration needs to be given to the number of resources that each judge needs to assess, and the time commitment required of judges. To investigate how such classifications should be collected under these constraints, we conducted a user experiment investigating whether human judges can agree on what resources are likely to support learning.

The most common way to retrieve resources to be judged is to issue keyword queries to one or more search engines, and use some subset or all of the resources that are returned.

When judging relevance, judges draw a relationship between queries and resources. However, when judging whether a resource is likely to support learning, the resource can be considered independently of the query. In this experiment, we further investigate whether visibility of the query used to retrieve the resource has an effect on judgments.

Eight judges were recruited for the experiment. These judges represent the hypothetical users who might use a system to find educational resources, whose needs the retrieval system aims to meet. Participants were acquaintances of the author, from diverse backgrounds, and all had some experience with using web browsers and web interfaces.

A total of 20 resources were judged by the eight judges under one of two conditions: the query used to retrieve the resource being visible (q) or not visible (q'). Each judge viewed ten resources under condition q and ten under q'. A Latin square design [Kelly, 2009] was used to control for ordering effects.

4.2.1 Resource selection

There are many types of resources available on the Web, such as HTML pages, images, multimedia, PDFs, and Word documents. In this thesis, we restrict our consideration to resources that are HTML pages.

To retrieve appropriate resources, queries should represent likely queries in the domain under investigation. It is common in IR experiments to use queries from query logs, as these queries can be thought of as a real user's attempt to distil their information need into a query.

The Flexible Learning Toolboxes [Oliver et al., 2005] are a collection of educational resources managed by e-Works[1] that comply with the Sharable Content Object Reference Model (SCORM). Further description of the resources in the collection is provided in Appendix A. A log of queries submitted to the live repository of the e-Works collection was obtained, containing 21 139 queries, 7764 of them unique. Queries for the experiment were drawn at random from the unique queries. If it was judged improbable that submitting a particular query to a search engine would return educational resources, that query was discarded. For example, the queries "rte2606a" and "a*" were discarded. A total of 20 queries were selected for our experiments, and these are shown in the query terms column of

[1] http://www.eworks.edu.au

Table 4.1.

While the queries were originally used for seeking resources from an educational repository, the resources used for our experiment were retrieved from a search across the entire web, with each query being submitted to the Yahoo! Search API.[2] Alongside each selected query, Table 4.1 shows the resources used in our experiment, which were selected as follows.

For the first ten queries, resources returned at rank position one were selected for judging. Call this set of resources R_A, resources 1 to 10 from Table 4.1.

It is also important that an evaluation collection contains representative examples of the classes of resources that retrieval systems want to retrieve. To ensure that our collection contained an adequate proportion of resources for which a positive judgment was probable, resources returned using the second ten queries were judged by the author of this thesis, in rank order, according to the same criteria that would ultimately be used by the participants. For each query, the highest ranked resource judged likely to support learning was added to the collection. These judgments were made without reference to the search query used to retrieve the resources. Call these resources R_B, resources 11 to 20 from Table 4.1.

4.2.2 Presentation and user interface

Five resources from R_A and five from R_B were combined and their order randomised to form the first pool, P_1. The remaining resources were combined and their order randomised to form P_2. The judgment pools contained the following resources.

$$P_1 = [4, 11, 15, 5, 3, 13, 2, 12, 1, 14]$$

$$P_2 = [19, 6, 16, 18, 10, 8, 17, 20, 9, 7]$$

Resources were presented to judges in four ways, as described below.

Group 1) P_1 with the query followed by P_2 without the query.

Group 2) P_1 without the query followed by P_2 with the query.

Group 3) P_2 with the query followed by P_1 without the query.

Group 4) P_2 without the query followed by P_1 with the query.

[2] http://developer.yahoo.com/search/web/

Table 4.1: Query terms, result list rank, and retrieved resources used in the user experiment collection.

resource	query terms	rank	URL (http://)
R_A			
1	Communicate with colleagues and clients in an office environment	1	www.visualdatallc.com/clients.aspx
2	food safety practices	1	www.ehow.com/how_2075133_practice-food-safety.html
3	cash flow	1	en.wikipedia.org/wiki/Cash_flow
4	Maintain equipment for activities	1	www.govliquidation.com/list/c7484/lna/1.html
5	safe beauty	1	www.safeinternetshops.com/beauty.htm
6	search internet	1	kaftos.com
7	software development	1	en.wikipedia.org/wiki/Software_development
8	lifting safely	1	www.smarter.com/---se--qq-Lifting+Safely.html
9	cash budget	1	www.investopedia.com/terms/c/cashbudget.asp
10	Process customer complaints	1	www.bcuc.com/Complaint.aspx
R_B			
11	costing ingredients	4	www.ces.ncsu.edu/depts/poulsci/tech_manuals/ingredient_sampling.html
12	prepare cook and serve food	3	www.cfsan.fda.gov/~dms/hret2-2.html
13	computing skills	3	www.cnn.com/TECH/computing/9806/03/autism.idg
14	treat weeds	1	www.blm.gov/weeds/FAQs/FAQs.htm
15	library skills	31	www.rock-hill.k12.sc.us/schools/high/sphs/Media/libraryskills.htm#Mod1
16	(plan and conduct and meetings)	3	www.azskillsusa.org/Teachers/meetings.htm
17	administer projects	8	sais-jhu.edu/cmtoolkit/issues/evaluation/index.html
18	Coordinate implementation of customer service strategies	50	en.wikipedia.org/wiki/Customer_relationship_management
19	write simple documents	1	nadoka.vipnet.org:8080/doc/user/08_is1.htm
20	Arts administration	3	en.wikipedia.org/wiki/Arts_administration

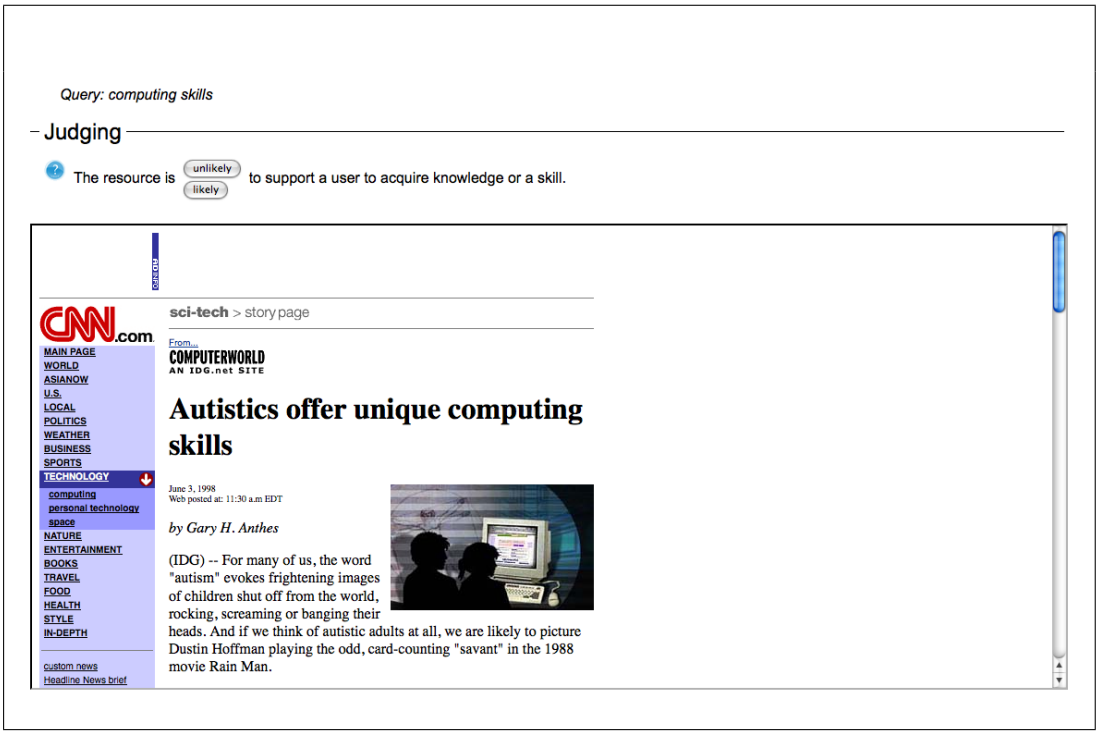


Figure 4.1: The judgment interface used in our user experiment, with the query used to retrieve the resource visible.

For example, Group 1 were presented the resources from P_1 with the original search query displayed, and then the resources from P_2 were presented without the original search query displayed. The ten resources in P_1 and P_2 were always presented in the same order.

Two judges were randomly assigned to each group.

Resources were presented sequentially via a web interface. Judges were asked to classify resources as likely or unlikely to support learning. Specifically, they were presented with the statement, "The resource is likely/unlikely to support a user to acquire knowledge or a skill," where the words "likely" and "unlikely" were buttons that recorded the judgment. The judgment interface with the query visible is shown in Figure 4.1. The instructions given to judges are presented in Appendix F.

An HTML `iframe` element was used to embed single page resources in the judgment interface. All links within the resources were disabled, and raters evaluated the resources without reference to other web pages.

The resource displayed in Figure 4.1 is one of the resources used in our experiment,

resource 13 from Table 4.1. The judgment interface without the query visible was identical, save for the removal of the query from the top of the page. When discussing our results in Section 4.4, we use judgments for this resource to illustrate the analysis.

Each participant answered several questions after each judgment, including being asked for comments about the resource they had just judged, and also completed a post-experiment questionnaire after they completed all judgments.

4.3 Measuring agreement

The level of agreement between raters will assist in deciding how many judgments are needed for each resource to establish an accurate ground truth.

In this section we revisit the measures of agreement described in Chapter 2. Alongside our discussion of agreement measures, we present a worked example of each method, in the domain of judging educational material, using the judgment data collected in the user study described in this chapter.

4.3.1 Overlap

Overlap is the mean of the size of the intersection of positive ratings divided by the size of the union of positive ratings for each resource. For two judges, that is the size of the set of resources judged relevant (for our purposes, resources judged educational) by both judges divided by the size of the set of resources judged relevant by either judge.

In work that has been used to justify the use of a single assessor for many retrieval experiments, Voorhees [1998] used overlap between each pair of three judges, and overlap across three judges, to show that relative rankings of retrieval systems are stable despite variability in relevance assessments. However, the value of this overlap when calculated across all judges will decrease as the number of judges increases, as a single dissenting judge counts as disagreement. For this reason, mean pairwise overlap is a more useful measure.

The judgment data we use to illustrate all agreement measures is shown in Table 4.2. Let there be J judges and R resources, and thus we have eight judges ($J = 8$) each rating 20 resources ($R = 20$). A value of 0 represents a judgment that a resource was not educational, and a value of 1 represents a judgment that a resource was educational.

Table 4.2: Ratings of eight judges on twenty resources, with 0 and 1 representing a judgment that a resource is unlikely and likely to support learning respectively.

	Resource																			
Judge	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	1	0	0	1	1	1
2	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	1	0	1	1	1
3	0	1	1	0	0	0	1	0	1	0	1	1	0	1	1	1	0	1	1	0
4	0	0	1	0	0	0	1	0	1	1	1	1	1	1	1	0	0	1	0	1
5	0	1	1	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	1	1
6	0	1	1	0	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1
7	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	0	1	1
8	0	1	1	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1

Pairwise overlap can be calculated as follows. Consider judges 1 and 2; they agree that resources 3, 18, 19, and 20 are educational, so the intersection is 4. However, there were a further 8 resources that were rated educational by one judge and not by the other, so the union is 12. Overlap for these two judges is therefore $\frac{4}{12} = 0.333$. Mean pairwise overlap is the average overlap across all pairs of judges, as shown in Table 4.3, which in this case is 0.595.

4.3.2 Raw agreement

Raw agreement is the proportion of observed agreement to possible agreement. In the context of assessments of whether resources are educational, if a random resource is selected from a test collection, and we select a random rater who has judged the resource to be educational, what is the probability that another random judge will agree?

We derive the measure for raw agreement, limiting our discussion to the binary case, which we use in our user experiment. See Uebersax [2008] for calculation of raw agreement with an arbitrary number of categories.

We calculate raw agreement as follows. Let p_r be the number of times resource r was positively rated and n_r be the number of times resource r was negatively rated. To illustrate, we can see from Table 4.2 that resource 2 has six positive judgments, giving $p_1 = 6$, and resource 8 has six negative judgments, giving $n_8 = 6$. The number of times a rating was

Table 4.3: Pairwise overlap, calculated by dividing the number of resources that both judges rate as likely to support learning (intersection) by the number of resources that either judge rates as likely to support learning (union).

judge							
2	$\frac{4}{12}$						
3	$\frac{6}{13}$	$\frac{6}{13}$					
4	$\frac{7}{12}$	$\frac{4}{15}$	$\frac{8}{14}$				
5	$\frac{7}{14}$	$\frac{6}{15}$	$\frac{10}{14}$	$\frac{10}{14}$			
6	$\frac{7}{15}$	$\frac{7}{15}$	$\frac{11}{14}$	$\frac{10}{15}$	$\frac{12}{15}$		
7	$\frac{7}{16}$	$\frac{6}{17}$	$\frac{10}{16}$	$\frac{10}{16}$	$\frac{13}{15}$	$\frac{13}{16}$	
8	$\frac{8}{15}$	$\frac{8}{15}$	$\frac{11}{15}$	$\frac{10}{16}$	$\frac{12}{16}$	$\frac{13}{16}$	$\frac{13}{17}$
	1	2	3	4	5	6	7

judge

given to each resource in our data is shown in Table 4.4.

Table 4.4: The number of judges that rated each resource as likely to support learning, or positive ratings p, and the number of judges that rated each resource as unlikely to support learning, negative ratings n.

	Resource																			
Rating	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n	8	2	0	8	7	8	1	6	1	3	2	2	4	1	1	2	5	2	1	1
p	0	6	8	0	1	0	7	2	7	5	6	6	4	7	7	6	3	6	7	7

There are p_r positive ratings for resource r. If we take one judge who rated resource r positively, there are $p_r - 1$ judges that agree. Raw agreements are bi-directional, so total positive agreement is calculated by $p_r(p_r - 1)$. Negative agreement is calculated in the same way, taking negative ratings instead of positive ratings.

From the example judgments, take resource 8 from Table 4.4. We see that $p_8 = 2$, meaning that 2 judges said resource 8 is educational. Thus, the total number of agreements that resource 8 is educational is $p_8(p_8 - 1) = 2(1) = 2$

Therefore, the observed agreement across all R resources can be expressed as follows, with

A^-_{obs} representing observed agreement on negative judgments and A^+_{obs} observed agreement on positive judgments.

$$A^-_{obs} = \sum_{r=1}^{R} n_r(n_r - 1)$$

$$A^+_{obs} = \sum_{r=1}^{R} p_r(p_r - 1)$$

The number of possible agreements, A^-_{poss} for negative agreement and A^+_{poss} for positive agreement, can be calculated similarly, but instead of taking the number of judges that agreed with the original judge, we take the number of judges that *could* have agreed. Since it makes no sense to count judges agreement with themselves, this means possible agreement is one fewer than the total number of judges, or one fewer than the total number of ratings, $(p_r + n_r - 1)$, and thus positive agreement for a resource is $p_r(p_r + n_r - 1)$.

$$A^-_{poss} = \sum_{r=1}^{R} n_r(n_r + p_r - 1)$$

$$A^+_{poss} = \sum_{r=1}^{R} p_r(n_r + p_r - 1)$$

The observed and possible agreement for our example are shown in Table 4.5.

Table 4.5: Observed and possible agreement for each resource. Given the rating of a judge, observed agreement is the number of judges who agreed with that rating, and possible agreement is the number of judges who could have agreed.

	Resource																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
A^-_{obs}	56	2	0	56	42	56	0	30	0	6	2	2	12	0	0	2	20	2	0	0	288
A^-_{poss}	56	14	0	56	49	56	7	42	7	21	14	14	28	7	7	14	35	14	7	7	455
A^+_{obs}	0	30	56	0	0	0	42	2	42	20	30	30	12	42	42	30	6	30	42	42	498
A^+_{poss}	0	42	56	0	7	0	49	14	49	35	42	42	28	49	49	42	21	42	49	49	665

Therefore, we can calculate specific agreement for both the positive (A^+) and negative (A^-) cases to be the proportion of observed agreement to possible agreement.

$$A^- = \frac{A^-_{obs}}{A^-_{poss}}$$

$$A^+ = \frac{A^+_{obs}}{A^+_{poss}}$$

For our data, this means we have $A^- = \frac{288}{455} = 0.633$ and $A^+ = \frac{498}{665} = 0.749$.

Overall agreement can similarly be calculated by dividing the observed agreement from both positive and negative judgments by the number of possible agreements.

$$A = \frac{A^-_{obs} + A^+_{obs}}{\sum_{r=1}^{R}(n_r + p_r)(n_r + p_r - 1)}$$

Of course, $n_r + p_r$ is constant, the number of judgments made on a resource, and therefore the number of judges. We defined the number of judges earlier as J, so the overall agreement can simplified to the following.

$$A = \frac{A^-_{obs} + A^+_{obs}}{R \cdot J \cdot (J-1)}$$

Using our example data, for overall agreement we have $A = \frac{288+498}{(20)(8)(7)} = \frac{786}{1120} = 0.702$.

4.3.3 Kappa

The measure κ developed by Fleiss [1971] can be used to measure agreement between multiple raters. It was developed to overcome the fact that overlap and raw agreement are not corrected for chance, and it is not possible to estimate a confidence interval.

In calculating Fleiss' κ we ask the question, given that we have some set of observed judgments, what agreement would we expect by chance? The proportion of agreement expected by chance can be represented as $\bar{P}_e$. If we take this value away from perfect agreement, we have the best agreement possible, $1 - \bar{P}_e$. If we take the chance agreement away from what was observed, which we can call $\bar{P}_o$, and divide it by the best possible agreement, we have the proportion of agreement that is not due to chance.

Therefore, κ can be defined as follows, with a value of 1 indicating complete agreement, and a value less than 0 representing agreement less than would be expected by chance.

$$\kappa = \frac{\bar{P}_o - \bar{P}_e}{1 - \bar{P}_e}$$

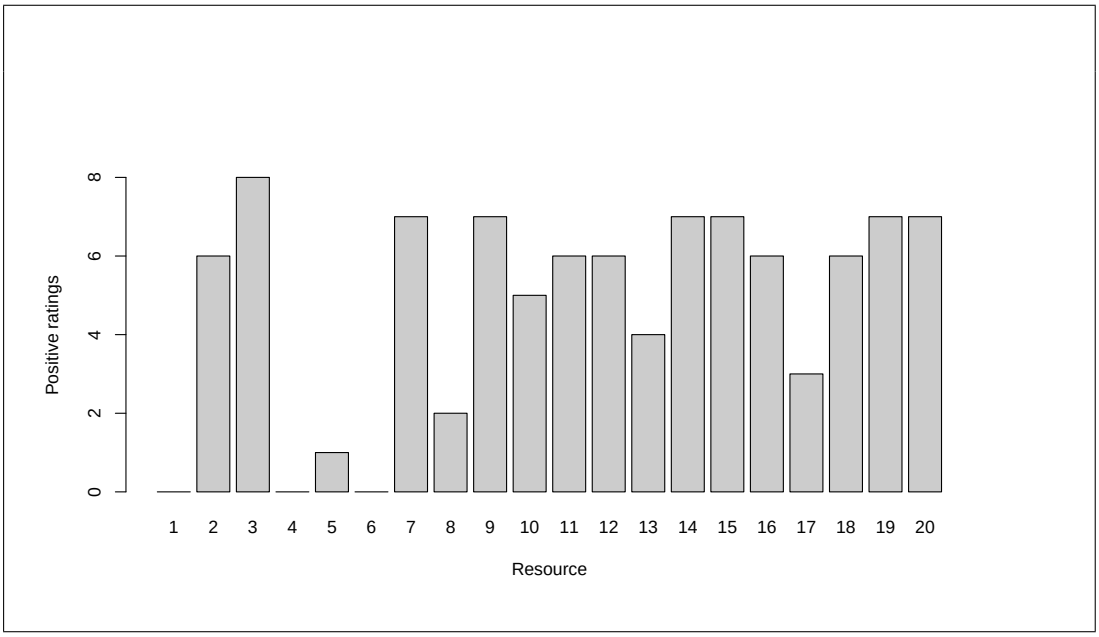


Figure 4.2: The number of times each resource was judged positively, or likely to support learning.

It is then possible to calculate the standard error, and a confidence interval. Calculating κ for our data, we have $\kappa = 0.382$ and $p < 0.001$. Therefore, we observe agreement above chance, and our data does show statistically significant agreement.

4.4 Analysis

This section reports on the results of our experimental evaluation of rater agreement, and provides analysis of these results. To give a general picture of the judgments, the number of times a resource was judged educational is shown in Figure 4.2. For our analysis, we begin by investigating rater agreement across all judgments regardless of other factors. Second, we discuss the impact of query visibility on rater agreement. We then report on the influence of resource type, that is, whether the resource was included in the judgment pool as a first ranked resource, or as a resource that was pre-selected to be a likely educational resource. Finally, we conclude by providing a discussion of comments that raters made after judging each resource.

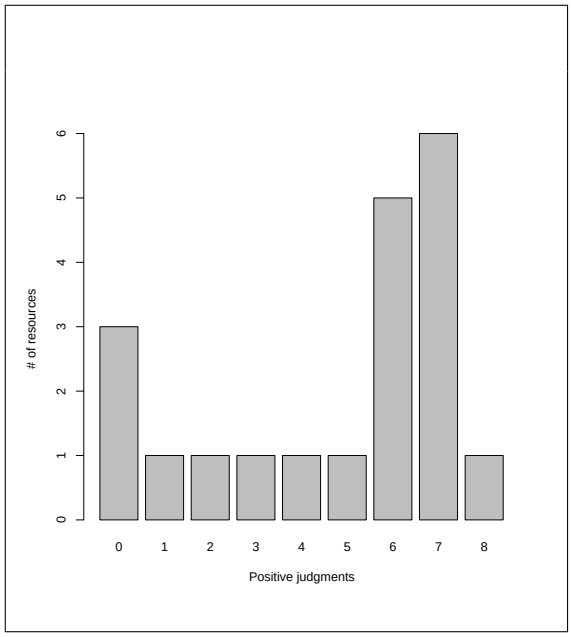


Figure 4.3: The frequency with which a number of raters judged resources as likely to support learning, with the number of resources that no rater judged likely to support learning on the left and the number of resources that all raters judged a resource likely to support learning on the right.

4.4.1 General rater agreement

The frequency with which a number of raters judged resources as likely to support learning is shown in Figure 4.3. The leftmost bar represents the number of resources that all raters judged as unlikely to support learning, and the rightmost bar represents the number of times that all raters judged a resource as likely to support learning. The example resource in Figure 4.1 was the most contentious resource, with four judges believing it was educational and four believing it was not. Overall, we see a bimodal distribution, with higher frequencies at the extremes. This is as expected if there is a high level of agreement.

The agreement measures between the eight judges observed across all resources, as calculated when describing the measures in Section 4.3, are presented in Table 4.6. All measures indicate a high level of agreement, and the value of κ is highly significant. The calculated mean pairwise overlap measure between the eight judges is 0.595, compared with the mean pairwise overlap measure between three assessors of 0.447 shown in Voorhees' work that

Table 4.6: Summary of measures of agreement between all eight judges.

Overlap	0.595
Negative	0.633
Positive	0.749
Overall	0.702
κ	0.382 ($p < 0.001$)

justified the use of a single judge in relevance assessments.

Our results show a high level of general agreement. Indeed, our results show a level of agreement higher than that used in the IR literature to justify the use of a single assessor. Given this high level of agreement, we conclude that it is appropriate that resources are categorised by a single judge rather than have multiple judges categorise each resource. In particular, for fixed time and number of judges, it is more useful to judge a larger number of resources than have multiple judgments on fewer resources.

4.4.2 Query visibility

When building a collection for assessing systems using the Cranfield method, an assessor makes a judgment about the relevance of a document to a query. The query is therefore central to the process, and different queries will cause the resource to be judged differently. However, when judging whether a resource is educational, the judgment criteria are stable, and it is unclear what effect query visibility would have on the judging process. Here we report the results of varying query visibility.

Each resource has an *a priori* probability of being judged likely to support learning. In this case, we are interested in the conditional probability, that is, the probability a resource will be judged likely to support learning given query visibility.

Figure 4.4 shows the frequency with which a number of raters judged a resources as educational, separated by query visibility. Each resource was judged by four judges under each condition. The leftmost bar shows that, with the query visible five resources received no positive ratings, while without the query visible three resources received no positive ratings. The rightmost bar indicates that all raters judged a resource educational on three occasions

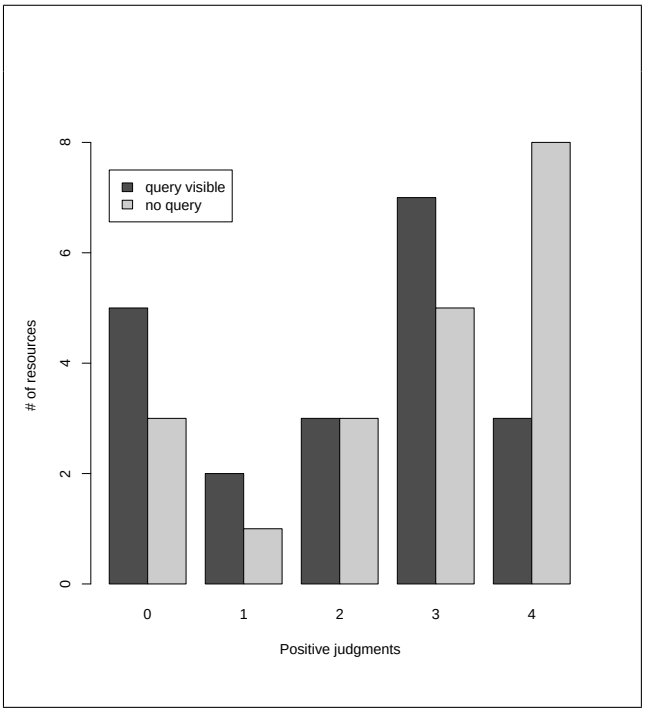


Figure 4.4: The frequency with which a number of raters judged resources as likely to support learning, separated by query visibility, with the number of resources that no rater judged likely to support learning on the left and the number of resources that all raters judged a resource likely to support learning on the right.

when the query was visible and on eight occasions when the query was not visible. While judgments of the example resource from Figure 4.1 were evenly split overall, when the query used to retrieve the resource was not displayed, three of the four judges who assessed this resource said it was educational material. However, when judges could see the query, only one of the four assessors judged the resource as educational material.

As with Figure 4.3, bimodal distributions indicate a high level of agreement. The distributions of frequency with and without the query being visible do appear to be generally bimodal, however, inspection suggests that displaying the query makes it less likely that a resource will be judged educational.

The agreement measures when split by query visibility are presented in Table 4.7. We can see that on all measures except negative agreement, agreement is noticeably higher when the query is not visible, though κ is significant in both cases. There is a very high level of

Table 4.7: Summary of measures of agreement between all eight judges, separated by query visibility.

	query	no query
Overlap	0.516	0.685
Negative	0.667	0.615
Positive	0.683	0.815
Overall	0.675	0.750
κ	0.350 ($p < 0.001$)	0.430 ($p < 0.001$)

positive agreement when the query is not visible, meaning that when the query is not visible raters very often agree that a resource is educational. It appears that judges use different criteria to rate a resource when the query is visible. Fisher's exact test indicates that query visibility has a weakly significant effect on judgments ($p = 0.053$).[3]

This effect is significant when considering only the case when the resource is not the first ranked result. This result is intuitively reasonable, as the search engine used for retrieving the resource has rated the resource as less relevant than other resources, as reflected in its ranking. We suggest that the query distracts raters from the task of judging whether a resource is likely to support learning, and causes them to judge relevance instead.

When people use search engines generally, they issue a query and judge how well the documents returned meet their information need. That is, they judge the relevance of returned resources to their query. Therefore, when presented with a resource to judge, and the query that was used to retrieve it, it is unsurprising that their judgments reflect relevance. As we are interested in filtering educational resources, relevance is handled by the search engine, and therefore should be factored out for our purposes.

4.4.3 Resource rank

Figure 4.5 shows the positive ratings made on each resource, that is, ratings where raters judged the resource educational, separated by query visibility. As described in Subsection 4.2.1, resources 1 through 10 were included in the judgment pool because they were returned

[3]Given the choice of a significance level is relatively arbitrary [Fisher, 1950], we claim a weakly significant effect, report the value of p, thus allowing the reader to draw their own conclusions.

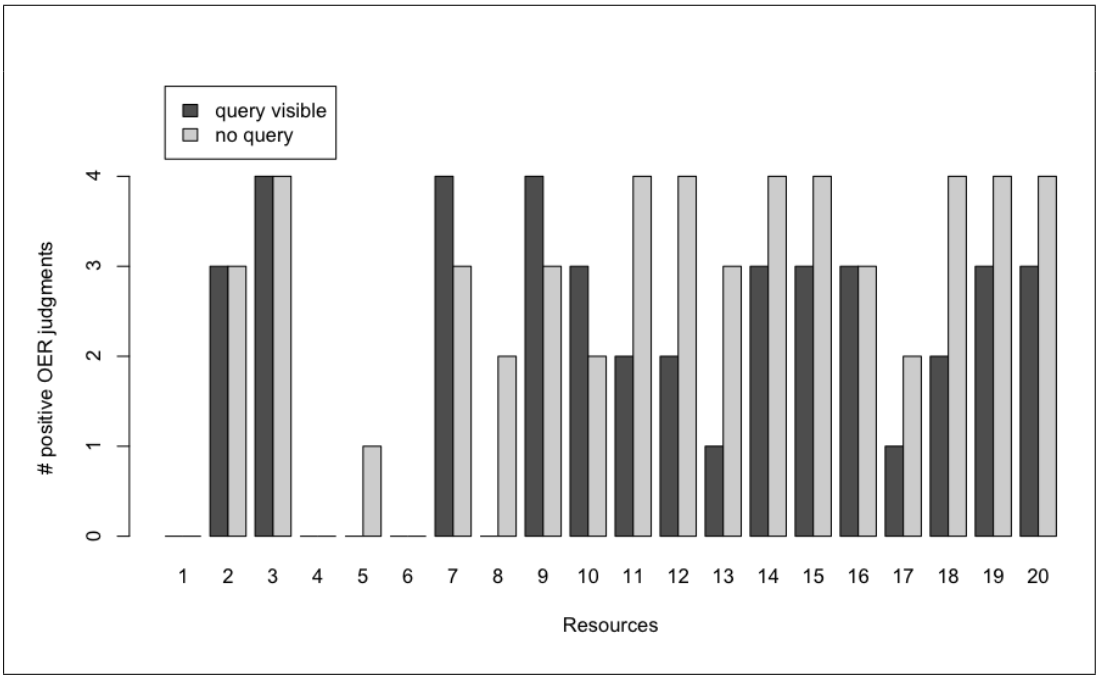


Figure 4.5: Number of times each resource was judged likely to support learning, separated by query visibility.

at rank position one in response to a search for the first 10 queries (R_A), and the resources 11 through 20 were included in the judgment pool because they were the highest ranked resource judged educational from the results returned in response to the second 10 queries (R_B).

The agreement measures when split by resource group are presented in Table 4.8. On all measures, agreement is lower for resources in R_B, and though we see similar values for overlap, positive agreement and overall agreement, κ does not show significant agreement for R_B. Negative agreement is particularly low for R_B when compared with R_A.

Fisher's exact test indicates that how a resource was added to the judgment pool has a significant effect on judgments ($p < 0.001$). This means that the proportions of negative and positive judgments are different depending on whether the resource was a first ranked resource or was the highest ranked resource judged to be educational.

Table 4.1 shows that most of the resources R_B (those included in the pool because they were judged likely to support learning) were ranked in the top 10 results. The mean rank was 7.7 and the median was 3. Also, half the resources in R_A (those included as the first

Table 4.8: Summary of measures of agreement between all eight judges, separated by whether the resource was included in the judgment pool as the first ranked result returned for a query or as the highest ranked resource that was judged likely to support learning by the author.

	R_A	R_B
Overlap	0.622	0.583
Negative	0.805	0.272
Positive	0.762	0.741
Overall	0.786	0.618
κ	0.567 ($p < 0.001$)	0.013 ($p = 0.827$)

ranked result) were judged educational by a majority of the judges. This suggests that a reasonable percentage of highly ranked resources will be judged likely to support learning, and that we need not have taken the precaution of pre-judging some resources. Therefore, it is appropriate to include the first N results from the returned ranked results in the collection. The results of this preliminary study suggest that $N = 10$ is sufficient, which is the method we follow when revisiting collection construction in Chapter 6.

4.4.4 Rater comments

In the judgment interface raters were invited to make comments about their judgments. In total, raters made 91 comments from the 160 judgments. Seven of the eight raters made at least one comment after judging a resource. Approximately a third of the comments make reference to the query used to retrieve the resource. For example, after judging a resource with the query visible one rater said, "Query asking general question; resource for much more specific request which is likely irrelevant. Therefore, easy to judge," and another said, "query not specific enough," and "if it was autism and computing skills this would be a useful resource."

In some cases, the rater stated that they found the resource difficult to judge because the query was not known, "Specific resource and without search terms, difficult to determine whether relevant to query; therefore, difficult to judge."

These comments appear to suggest that raters find it more difficult to judge whether a resource is educational in the absence of the implied context given by a query. It might be

expected that a more difficult judgment decision would take longer to make; however timing measurements reveal that there is no significant interaction between judging times and query visibility.

4.5 Summary

The methodology presented in this chapter produces a ground truth that can be used in the evaluation of systems that filter web search results for educational resources. This methodology is based upon the Cranfield method, which is the most widely used evaluation approach in IR experiments. We explore several elements of the methodology in relation to resources, queries, judges and the judgment subsystem. Specifically, we establish that a single judgment of each resource is sufficient, and that the queries used in the retrieval of resources should not be presented to assessors as part of the judgment interface.

We present a user experiment in which participants judged whether web resources were educational, in a manner similar to the way relevance assessments are collected when building test collections for use in experiments using the Cranfield method.

In relation to query visibility, our results show a high level of agreement both with and without the query visible. The level of agreement is higher when the query is not visible, though this is only weakly significant overall. We conclude that allowing the judge to see the query is detrimental to the judging process.

As there was a high level of agreement, the simplifying assumption made in the work described in this chapter—that a resource can be judged independently of educational context—appears to be reasonable. However, judges did report that they found the task difficult. The task may be made simpler for judges by asking them to rate resources in an artificial context. Judges can then additionally make a judgment as to whether the resource is educational in other contexts, with the subsequent ground truth being the union of these contextual and extra-contextual judgments. This is the method we use when developing the validation collection in Chapter 6.

In relation to the selection of resources appropriate for inclusion in an evaluation collection, the results indicate that it is appropriate to submit queries to a search engine and select returned resources for the collection. In this work, initial queries came from a log of queries submitted to a repository for e-learning material. This is reasonable in that the

users were searching for the type of material in which we are interested. However, a user's search behaviour may be different when searching a specific repository of educational resources as opposed to the wider Web, and thus the queries may not be representative of the sorts of queries that would be submitted to a filtering system. Equally, queries selected from a general query log, without knowledge of user intent, are likely to be inappropriate. When building the validation collection in Chapter 6, we use an alternative method for developing queries, extracting text from a curriculum document relevant to the context we present in that chapter. We believe this method will lead to realistic query formulation.

While we conclude that the use of a single judge per resource meets the requirements of constructing an evaluation collection, this does not answer the question of who should be performing the judgments, or what makes someone an appropriate judge. Aspects of this include the judges' expertise or familiarity with the topic or topics covered in resource and how this affects judgments, the judges' confidence when they classify a resource, and the ease with which the resource is classified. We investigate some of these issues in Chapter 6.

The judgment subsystem used in our experiment included a fixed description when asking judges to rate resources. We recognise that the instructions given to judges may influence the outcome of judgment experiments. Further, there may be differences in the judgments made by individuals belonging to different groups, such as students and teachers. Exploration of these issues is beyond the scope of the current work.

The next chapter presents our investigation into some of the features of resources and techniques that may be useful to a system for filtering educational resources.

Chapter 5

Filtering Educational Resources

Our results in Chapter 3 supported previous work that showed that, when searching for educational resources, people prefer to use a public search engine. Thus, adding the capability of categorising resources as educational or not educational to a retrieval system is likely to improve the satisfaction of users looking for resources to support learning. In the previous chapter, we described how such systems might be evaluated. In this chapter we investigate the development of such systems.

As discussed in Chapter 2, a filter for performing this type of classification can be constructed using machine learning techniques. A machine learning model for resource classification comprises a classification algorithm and a set of attributes. Attributes are name and value pairs, where the value can be, for example, Boolean, numeric, or nominal. These attributes are derived from characteristics or features of resources, with a one-to-one or one-to-many mapping between features and attributes.

The ease with which useful attributes can be derived from a feature varies. For example, the value of an attribute for a feature of the density of outgoing link text of a resource might be simply calculated by dividing the number of words in outgoing link text by the number of words in the document. At the other end of the spectrum, it might be extremely difficult to assess whether or to what extent a resource begins with learning objectives that are later addressed in the body of the resource.

While some features have a one-to-one mapping to attributes, such as the number of links in a resource, in other cases multiple attributes may be derived from a single feature. For

example, a feature might be the actual words used in a collection. This feature could be represented with a set of boolean attributes, one for each word in the collection, indicating the presence or absence of that word in a resource. Preliminary investigation suggests that the words in a collection are likely to be one of the most effective features for distinguishing educational from non-educational resources. We therefore begin by investigating this feature and proceed as follows.

In this chapter, we explore a variety of machine learning algorithms and the parameters that can be passed to them. In this exploration, we use resource text to choose candidate algorithms to investigate further with other features. As many machine learning algorithms cannot use string data directly as an attribute, we also evaluate alternative methods of pre-processing the text to convert it into usable attributes.

After identifying a shortlist of candidate algorithms that work effectively with the text feature, we explore other resource features, with the aim of choosing an overall effective classification algorithm by developing additional attributes with discriminatory power.

When examining classifier performance in this chapter, we report six measures; AUC, accuracy, precision, recall, F-measure, and kappa. Of these measures, we rely on AUC to make choices between classifiers and tuning parameters. We do not perform significance testing when comparing techniques and developing classifier models in this chapter, but simply choose the best performing techniques based on AUC. We test for significance when comparing models in Chapter 6. For details of the measures used, see the description in Section 2.3 of Chapter 2.

This chapter is structured as follows. We begin in Section 5.1 by describing the construction of a development collection to be used for the selection and tuning of a machine learning model to differentiate between educational and non-educational resources. Section 5.2 then introduces various machine learning algorithms and we explore their effectiveness when applied to text extracted from resources. In Section 5.3 we develop further attributes aimed at assisting the automated detection of educational material, giving details about how those attributes are extracted from resources, and select the best machine learning algorithm for our purposes. Finally, we summarise our findings in Section 5.4.

Table 5.1: Queries extracted from the eWorks query log and used to retrieve resources to build the development collection. For each query, 10 HTML resources were retrieved and judged by the author of this thesis.

Query	Number of resources judged educational	Rank of skipped resources
home community care	0	
dealing with customer complaints	10	7
occupational health and safety computers	4	
visual design	3	
Conduct food safety audit	2	1, 3, 5, 7, 8, 11, 13, 14
breathing apparatus	4	
complex workplace communication	5	3
conduct financial transactions	5	1, 3
furniture costing	3	
risk	3	

5.1 Developing a collection for exploring resource features

To develop a system for filtering educational material, it is necessary to have a collection of resources that can be used for testing and tuning the machine learning model. To avoid biasing the evaluation of the system, this collection should be distinct from the collection used to validate it. In this section we describe the construction of the development collection that will be used in the remainder of this chapter to develop our classifier for educational resources.

A new set of 10 queries was drawn from the eWorks query log using the same criteria as were used in Chapter 4 Subsection 4.2.1. These queries were then submitted to the Yahoo! Search API. For each query, the first 10 HTML resources returned in the results list were included in our collection, giving a total of 100 resources.

Resources were judged by the author of this thesis as either educational or not educational. The queries used to retrieve resources are shown in Table 5.1, along with the number of

resources from that query that were judged educational. Where a resource at a certain rank was skipped, either as a non-HTML resource or a dead link, the rank of the skipped resource is noted. Overall, of the 112 resources returned, one dead link and 11 PDFs were skipped, 39 were judged educational, and 61 were judged as not educational.

In the remainder of this chapter, by performing 10 times 10-fold cross-validation, these 100 resources are used to develop and tune a system for differentiating educational resources from resources that are not educational. We then validate the effectiveness of our new system, using a larger independently developed collection, in the following chapter.

5.2 Evaluating classification algorithms using resource text

The simplest feature to examine is words that appear in a resource. It seems intuitive that the words that are used in educational resources will be different from the words used in resources that are not educational. Inspection of the resources in the development collection appeared to support that intuition.

In this section, using the development collection described in Section 5.1, we investigate the effectiveness of various methods of creating a classification model using attributes developed from the text in resources, including text vectorisation, attribute selection and classification algorithm choice. We select a range of classification algorithms representing the most common classification methods.

For classification tasks, we use version 3.6.1 of the Weka Data Mining software,[1] which provides implementations of data mining tools and algorithms [Hall et al., 2009].

5.2.1 Converting resources to word vectors

To make a resource usable for a classification algorithm, text must first be extracted from the HTML source and converted from a string of text to a term vector.

To describe how a resource is converted to a term vector, we refer to the example resource shown in Figure 5.1, which will also be used when describing features later in this chapter. The example is a modified version of resource 20 from the rating agreement work described in Chapter 4. It has been modified for readability and edited for length.

[1] http://www.cs.waikato.ac.nz/ml/weka/

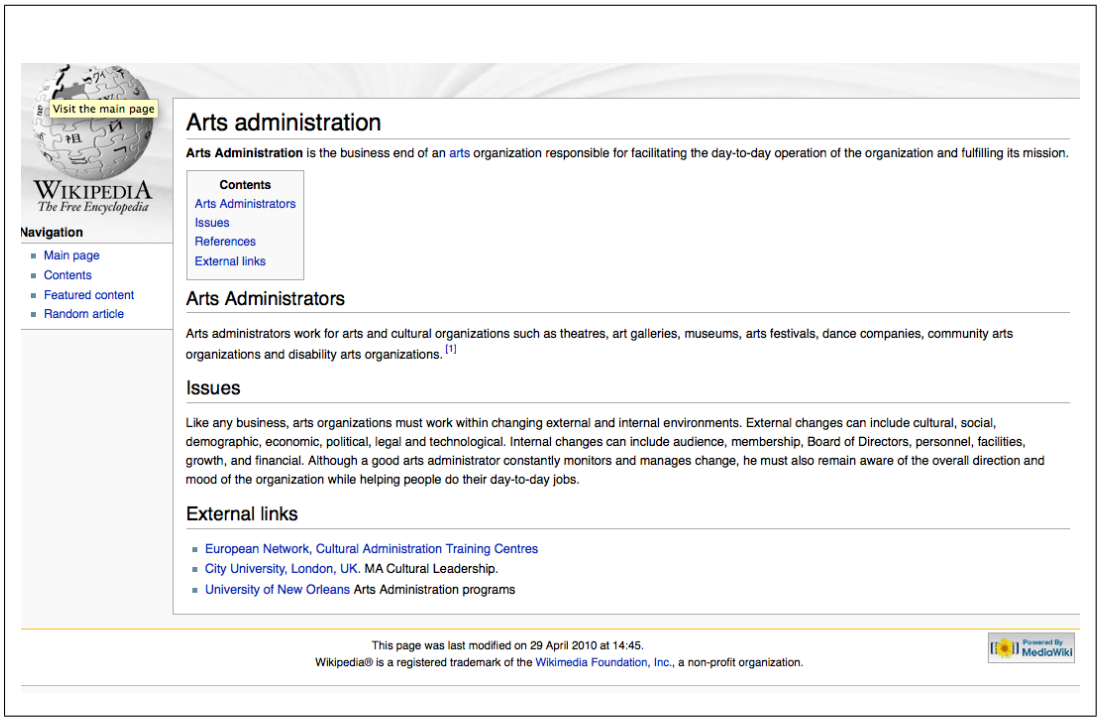


Figure 5.1: Rendered example resource, modified for readability and edited for length.

While unrendered text such as `meta` tags might hold clues as to whether a resource is educational, we are most interested in text that the user can read. We can see from the source code of the example resource, shown in Figure 5.2, that there is substantially less text visible to someone reading a web page than there is in the resource source, most of which is HTML markup. The simplest way to extract text is to remove all HTML tags and use the remaining text.

However, simply removing HTML tags and leaving the elements' contents will leave some text that is unlikely to be of use. Text content of `script` and `style` elements generally refer to the structure of the document rather than its content, and will therefore be unrelated to the topic of the resource itself. In our example, the `style` element on lines 3 to 6, and the `script` element on lines 69 and 70, contain words such as *import*, *skins*, *common*, *shared*, and *window*. At best these elements will have no impact on resource classification, while at worst may degrade performance. As such, `script` and `style` elements and their content are removed entirely.

After removal of the remaining HTML tags, we convert the extracted text into a term

```
<html><head>
  <title>Arts administration</title>
  <style type="text/css" media="screen, projection">
    @import "http://en.wikipedia.org/skins-1.5/common/shared.css?141";
    @import "http://en.wikipedia.org/skins-1.5/monobook/main.css?141";
  </style>
</head><body>
<div id="globalWrapper">
  <div id="column-content"><div id="content">
    <h1 id="firstHeading" class="firstHeading">Arts administration</h1>
    <div id="bodyContent"><p>
<b>Arts Administration</b> is the business end of an <a href="/wiki/Arts"
title="Arts">arts</a> organization responsible for facilitating the day-to-day
operation of the organization and fulfilling its mission.</p>

      <table id="toc"><tr><td><div id="toctitle"><h2>Contents</h2></div><ul>
          <li><a href="#Arts_Administrators">Arts Administrators</a></li>
          <li><a href="#Issues">Issues</a></li>
          <li><a href="#References">References</a></li>
          <li><a href="#External_links">External links</a></li>
      </ul> </td></tr></table>

      <h2>Arts Administrators</h2><p>
Arts administrators work for arts and cultural organizations such as theatres,
art galleries, museums, arts festivals, dance companies, community arts
organizations and disability arts organizations.<sup id="cite_ref-0"
class="reference"> <a href="#cite_ref-0">[1]</a></sup></p>

      <h2>Issues</h2><p>
Like any business, arts organizations must work within changing external and
internal environments. External changes can include cultural, social,
demographic, economic, political, legal and technological. Internal changes
can include audience, membership, Board of Directors, personnel, facilities,
growth, and financial. Although a good arts administrator constantly monitors
and manages change, he must also remain aware of the overall direction and
mood of the organization while helping people do their day-to-day jobs.</p>
```

Figure 5.2: HTML source of the example resource shown rendered in Figure 5.1.

```
37        <h2>External links</h2><ul>
38          <li><a href="http://www.encatc.org">
39            European Network, Cultural Administration Training Centres</a></li>
40          <li><a href="http://www.city.ac.uk/cpm/cultural_leadership_programme">
41            City University, London, UK.</a> MA Cultural Leadership.</li>
42          <li><a href="http://arta.uno.edu/">
43            University of New Orleans</a> Arts Administration programs</li>
44        </ul>
45      </div>
46    </div></div>
47
48    <div id="column-one"><div class="portlet" id="p-logo">
49        <a style="background-image:
50          url(http://upload.wikimedia.org/wikipedia/en/b/bc/Wiki.png);"
51          href="/wiki/Main_Page" title="Visit the main page"></a></div>
52      <div><h5>Navigation</h5><div class='pBody'><ul>
53        <li><a href="/wiki/Main_Page">Main page</a></li>
54        <li><a href="/wiki/Portal:Contents">Contents</a></li>
55        <li><a href="/wiki/Portal:Featured_content">Featured content</a></li>
56        <li><a href="/wiki/Special:Random">Random article</a></li>
57      </ul></div></div>
58    </div><div class="visualClear"></div>
59
60    <div id="footer"><div id="f-poweredbyico">
61      <img src="/skins-1.5/common/images/poweredby_mediawiki_88x31.png"/></a>
62      </div><ul id="f-list">
63        <li> This page was last modified on 29 April 2010 at 14:45.<br /></li>
64        <li>Wikipedia&reg; is a registered trademark of the
65        <a href="http://www.wikimediafoundation.org/">
66          Wikimedia Foundation, Inc.</a>, a non-profit organization.</li>
67    </ul></div>
68  </div>
69  <script type="text/javascript">
70    if (window.runOnloadHook) runOnloadHook();</script>
71  </body></html>
```

Figure 5.2: Example HTML resource (continued from previous page)

```
terms = {
14, 2010, 29, 45, [1], a, administration, administrator, administrators, also,
although, an, and, any, april, art, article, arts, as, at, audience, aware,
board, business, can, centres, change, changes, changing, city, community,
companies, constantly, content, contents, cultural, dance, day-to-day,
demographic, direction, directors, disability, do, economic, end, environments,
european, external, facilitating, facilities, featured, festivals, financial,
for, foundation, fulfilling, galleries, good, growth, he, helping, inc,
include, internal, is, issues, its, jobs, last, leadership, legal, like, links,
london, ma, main, manages, membership, mission, modified, monitors, mood,
museums, must, navigation, network, new, non-profit, of, on, operation,
organization, organizations, orleans, overall, page, people, personnel,
political, programs, random, references, registered, remain, responsible,
social, such, technological, the, theatres, their, this, trademark, training,
uk, university, was, while, wikimedia, wikipedia, within, work
}
```

Figure 5.3: The set of terms extracted from the source of the example HTML resource shown in Figure 5.2. These terms would represent the contribution of this resource to the term vector of the collection.

vector. To do this, we use Weka's `StringToWordVector` filter to take the text from each resource and produce a set of terms that appear in the collection and a vector with an entry for each resource indicating whether the term was present for that resource. Individual terms are extracted from the text using any sequence of whitespace (space, tab, and new line) and punctuation characters (`.,;:"()?!`) as delimiters. Terms are converted to lower case, so that, for example, "arts", "Arts", and "ARTS" are all treated as identical.

The output of this process on our example document is shown in Figure 5.3 as comma-separated values. Terms are extracted from each document in the collection, with every term in the collection becoming an attribute in the term vector. Each resource, or *instance*, can then be represented with attribute names being the terms of the vector and attribute values indicating whether the instance contains the term.

Term attributes can be represented as either the Boolean presence or absence of the term in a resource or the number of times the term appears in the resource. For simplicity in our initial exploration, we choose to use a Boolean value, as term counts introduce the confounding factor of document length, with longer documents containing higher frequency counts for most terms, requiring a choice of method to normalise term counts.

Finally, to keep processing time reasonable, we trim the vector to keep only the top 1000 terms by total number of occurrences across the whole collection. Ties at the 1000th position are retained, so that there may be more than 1000 terms retained. In our development collection, this results in a vector of 1037 terms.

Parameters to the vectorisation process can be altered to potentially change the output vector. Since exploration of each of these parameters in combination with each parameter of every classification algorithm we examine is prohibitively expensive, we first use the parameter values described above while identifying a shortlist of candidate classification algorithms. In Subsection 5.2.11 we tune these vectorisation settings with the shortlisted candidates to see if they improve classification performance. We also investigate term stopping and stemming.

5.2.2 Baseline

To allow meaningful evaluation, we require a baseline against which classifiers can be compared. Two commonly used baselines are ZeroR and OneR.

In the ZeroR classification scheme, resources in the test collection are simply assigned to the majority class found in the training data, ignoring all attributes of the test resources. For example, the collection developed in the Section 5.1 had 39 resources judged educational from a pool of 100 resources. With a ZeroR classification model trained with this collection, test resources would all be classified as not educational.

The OneR classification algorithm classifies test instances based on the value of a single attribute. The attribute chosen is the one whose value is most able to split the classes represented in the training instances. OneR was developed after it was observed that simple classifiers with very simple rules are often as effective as more complex classifiers [Holte, 1993].

Results of using the ZeroR and OneR classification algorithms on the development collection are shown in Table 5.2. Unsurprisingly, the OneR classifier performs better than

Table 5.2: Performance of baseline classifiers using text extracted from resources.

Measure	ZeroR	OneR
AUC	0.4811	0.7528
Accuracy	0.6100	0.7830
Recall	0.5000	0.7528
Precision	0.3050	0.7831
F-Measure	0.3789	0.7609
kappa	0.0000	0.5257
time (seconds)	0.0131	1.4508

the ZeroR classifier in all measures except time. Given that ZeroR simply checks the class attribute for each instance and does not perform any calculations, one would expect it to perform poorly. Given the more realistic performance of the OneR algorithm, we use it as our classification baseline for the remainder of this thesis.

5.2.3 Naïve Bayes

A Bayesian classifier uses the distribution of attribute values in the training instances to make an estimate of the probability that a test instance belongs to a particular class, given its attribute values.

When attributes are terms, a Bayesian classifier counts the number of times terms appear in the training collection, taking into account class labels. It then uses the number of occurrences of terms in resources of each class as an estimate of the probability that a new resource is a member of that class. A Naïve Bayes classifier makes the further assumption that the occurrence of terms is independent. Though independence of terms is unlikely, this assumption allows the multiplication of probabilities, so that the probability that a resource is educational can be simply calculated as the product of the probabilities of each of the terms it contains. In practice, Naïve Bayes classifiers often perform well, despite this simplifying assumption [John and Langley, 1995].

It is common for Naïve Bayes implementations to assume that the values of attributes

Table 5.3: Performance of Naïve Bayes classifiers using text extracted from resources. The number in brackets shows the percentage change in AUC compared to the OneR baseline.

Measure	Naïve Bayes	with kernel	with discretization
AUC	0.8449 (12.23%)	0.8669 (15.16%)	0.9032 (19.98%)
Accuracy	0.7990	0.8090	0.8540
Recall	0.7844	0.7967	0.8484
Precision	0.7899	0.8000	0.8462
F-Measure	0.7868	0.7982	0.8470
kappa	0.5738	0.5965	0.6941
time (seconds)	2.0916	2.5687	2.5741

are normally distributed. This is often not the case. The use of kernel function estimation has been shown to improve distribution estimation in several domains [John and Langley, 1995], and is especially useful if the nature of the underlying distribution is unknown. Alternatively, performance of Bayesian classifiers on numeric data can often be improved by using discretization, transforming numeric data into nominal data [Dougherty et al., 1995].

Table 5.3 shows the results of classification using the Naïve Bayes algorithm, as implemented by Weka.[2] Each of the Naïve Bayes classifiers used performs well, substantially outperforming our baseline measures. Naïve Bayes with discretization obtained the best result of the three, showing a 19.98% in AUC improvement over the OneR baseline. As the pre-processing performed on the resource text output a binary value based on the presence or absence of a word in a resource, it is unsurprising that the algorithms that assume a continuous distribution do not perform as well as discretization.

5.2.4 Rules

Rule-based classification algorithms involve learning rules from training instances that separate as many instances as possible into the correct class. Each rule is a group of conditions based on attribute values. The final model has a set of rules for each class. When testing

[2]For details of the tools mentioned in this chapter, see the Weka API at http://weka.sourceforge.net/doc/

instances, class rule sets are checked until a set matches.

We chose to investigate Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [Cohen, 1995], a rule-based classification algorithm, implemented in Weka as JRip. This algorithm develops rules by iteratively splitting training instances into two sets, one for growing rules by adding conditions and one for pruning conditions from rules. A rule is grown by repeatedly adding conditions until there are no negative examples of a class in the grow set. The rule is then pruned of conditions to minimise error rates in the pruning set. This is called *reduced error pruning*, and helps to avoid overfitting the model to the training data. Pruning techniques reduce the accuracy of the model on training data, but have been shown to reduce rule complexity without reducing accuracy on new instances [Clark and Niblett, 1987].

The JRip implementation allows for varying the number of subsets into which to divide the training data, using all but one subset as the growing set and one subset as the pruning set.[3] The default setting is to use two-thirds of the data for growing the rule, and one-third for pruning.

After a complete set of rules have been learned, JRip provides a mechanism for optimisation. For each rule, in the order the rules were learned, two alternative rules are developed. The first new rule is grown in the same way as the original rule, adding conditions to the empty rule. The second new rule is grown from the existing rule. These rules are pruned so as to minimise the error for the entire set of rules rather than for the individual rule, as when rules were first learned. If one of the new rules performs better than the original rule, the new rule is substituted. This optimisation run is performed twice by default, but the number of runs can be altered.

The results of our tuning of the JRip classifier can be seen in Figure 5.4. We varied the number of subsets used in rule construction. With two subsets, half the data is used as a growing set, and half as a pruning set. Using five subsets, that is one-fifth of the data for pruning, performed best on this dataset. Using this pruning level, we then varied the number of optimisation runs from one to 20, and found that nine optimisation runs performed best. Thus, pruning using one-fifth of instances with nine optimisation runs performs substantially

[3] Weka uses the term "folds" to describe these divisions, but to avoid confusion with cross validation folds we use the term "subsets", as per the original paper by Cohen [1995].

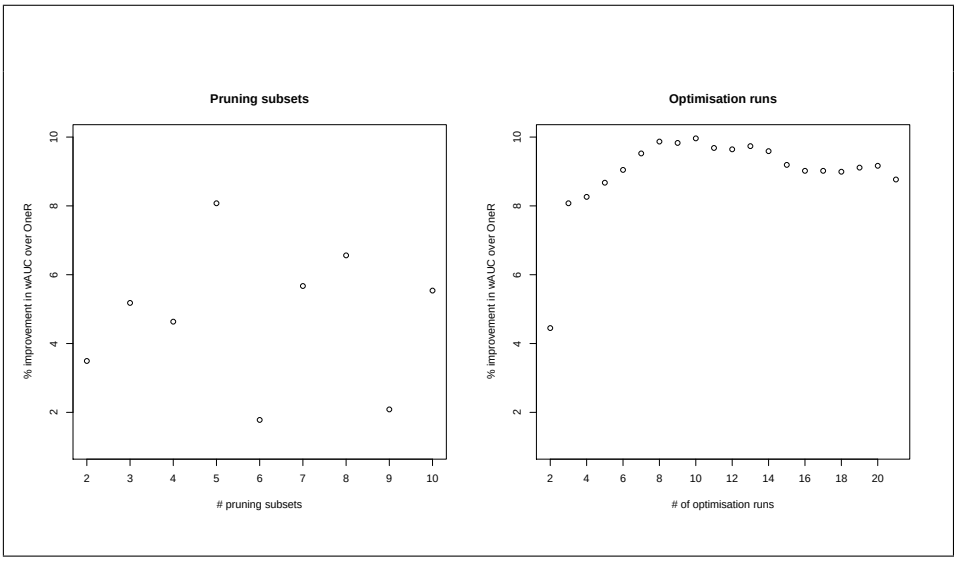


Figure 5.4: Performance of JRip classifier using text extracted from resources varying the number of rule pruning subsets and optimisation runs used in constructing the model.

better than the OneR baseline.

In Table 5.4, we show the performance of the JRip classifier under a variety of settings. The default, using three subsets and two optimisation runs, performs only slightly better than the OneR baseline. Substantial improvement in AUC, 9.96%, is achieved by tuning the number of pruning subsets and the number of optimisation runs.

5.2.5 Trees

Decision trees operate similarly to rule-based algorithms, but focus on attributes rather than classes. A decision tree algorithm recursively creates branches for observed values of a particular attribute until all instances in a branch have the same class. The decision of which attribute to split on is based on maximising the associated *information gain*, which is a measure of how much information is encoded in a particular attribute, a concept described in detail by Witten and Frank [2005]. Leaves of the tree indicate the class of instances. Test instances are assigned to a class by comparing their attributes to each node and branching accordingly until they reach a leaf.

As with the rules algorithm described in the previous section, pruning decision trees can improve their performance on test data. Two important pruning methods are *subtree*

Table 5.4: Performance of JRip classifier using text extracted from resources. The default uses three subsets and two optimisation runs, tuned subsets uses five subsets and two optimisation runs, and tuned runs uses fives subsets and nine optimisation runs. The number in brackets shows the percentage change in AUC compared to the OneR baseline.

Measure	Default	Tuned subsets	Tuned runs
AUC	0.7918 (5.18%)	0.8136 (8.08%)	0.8278 (9.96%)
Accuracy	0.7790	0.7860	0.8090
Recall	0.7717	0.7811	0.8074
Precision	0.7688	0.7761	0.8002
F-Measure	0.7692	0.7773	0.8021
kappa	0.5390	0.5554	0.6050
time (seconds)	12.8842	15.7448	52.8821

replacement, in which subtrees are replaced with a descendent leaf, and *subtree raising*, in which internal nodes and their descendent subtree are raised up the tree to replace an ancestor. Both methods use the training data to calculate a confidence interval of estimated error rates. As this confidence interval is calculated using the training data rather than independent test data, a pessimistic estimate is required so the upper bound is used to decide whether to alter the tree.

Reduced error pruning, in which a portion of the training data is held out and used for pruning, can be used as an alternative pruning method instead of subtree raising or subtree replacement. This has the advantage of using independent data to estimate error but the disadvantage of training the tree on less data. The proportion of data held back for reduced error pruning is controlled by dividing the training data into subsets, with one subset used for pruning and the rest for training.

For our experiments we test a C4.5 decision tree [Quinlan, 1993], as implemented by Weka's J48 classifier. C4.5 uses post pruning with subtree raising and a pruning confidence level of 0.25. We also investigate the use of subtree replacement as an alternative post pruning method. Pruning confidence level can be adjusted for both pruning methods, and we explore

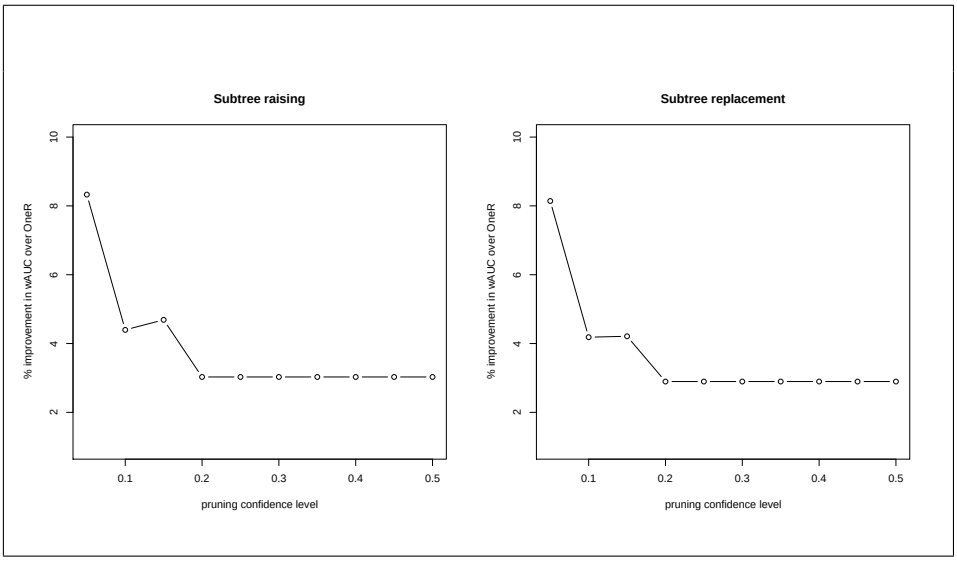


Figure 5.5: Performance of J48 classifier using text extracted from resources varying the confidence level used when deciding whether to prune a rule set using subtree raising when constructing the model.

optimal levels.

Results of varying the pruning confidence levels using subtree raising and subtree replacement are shown in Figure 5.5. The performance of the J48 classifier is similar for each form of subtree manipulation, performing best when aggressively pruning the tree, with the most effective confidence level being 0.05. At that level, subtree raising marginally outperforms subtree replacement, AUC 0.8155 compared to 0.8141. The flat portions of the graph demonstrate that no pruning occurs if the confidence level is set to 0.2. Both subtree manipulation methods perform better than the OneR baseline measure.

Reduced error pruning can also replace subtree manipulation, using three subsets by default. Finally, the tree can be left completely unpruned. The effect of varying the number of subsets used in reduced error pruning can be seen in Figure 5.6. Reduced error pruning does not perform as well as the subtree manipulation pruning methods described above, and only slightly better than the OneR baseline–worse when using five and eight subsets.

The results of the best performing of the J48 classifiers, using the pruning methods described above, as well as the results of an unpruned J48 classifier are shown in Table 5.5. Subtree raising provides the best performance, with 8.33% improvement over the AUC of

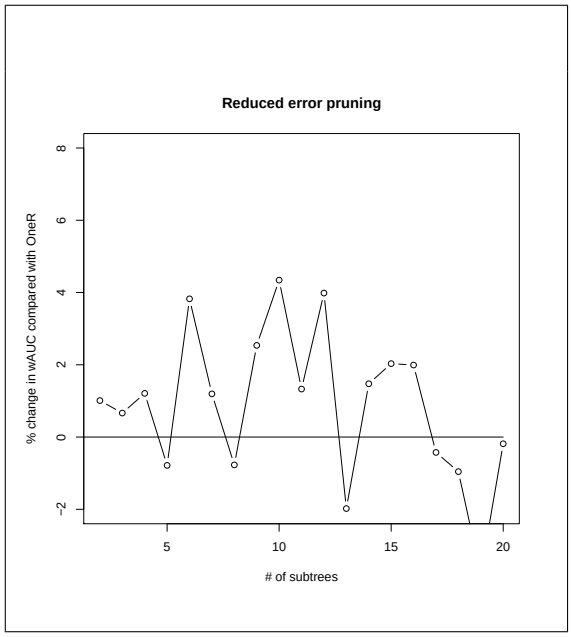


Figure 5.6: Performance of J48 classifier using text extracted from resources varying the number of reduced error pruning subsets used when constructing the model.

the OneR baseline.

5.2.6 Support vector machines

In a linear model, attributes are assigned weights, with the aim of creating a line that can separate instances into the correct class. Such a separating line is called a *hyperplane* [Curtis, 1984]. For non-trivial data, in most cases it is not possible to define a hyperplane that divides, and hence correctly classifies, all instances.

Support vector machines (SVMs) [Vapnik, 1995] extend linear models by transforming the instance space so that classes can be separated by a hyperplane in the new space. The individual instances in the transformed space that are closest to the separating hyperplane are called support vectors. As the hyperplane is effectively defined by the support vectors, changes in other instances have no effect on classification. This makes SVMs particularly resilient to overfitting. SVMs have been shown to be effective for text categorisation tasks [Dumais et al., 1998].

Sequential Minimal Optimization (SMO) [Platt, 1998] is an algorithm that implements a

Table 5.5: Summary of performance of J48 classifiers using text extracted from resources and different pruning methods when constructing the model. The number in brackets shows the change in AUC compared to the OneR baseline.

Measure	Unpruned	subtree raising	subtree replacement	reduced error
AUC	0.7746 (2.90%)	0.8155 (8.33%)	0.8141 (8.14%)	0.7855 (4.34%)
Accuracy	0.7990	0.8280	0.8270	0.8140
Recall	0.7872	0.8220	0.8221	0.7893
Precision	0.7895	0.8193	0.8181	0.8145
F-Measure	0.7878	0.8200	0.8193	0.7973
kappa	0.5759	0.6404	0.6390	0.5967
time (seconds)	6.4944	6.4945	6.4738	7.3369

SVM that performs particularly well with sparse data sets. Since the terms in a document are likely to be a small subset of those in the entire collection, we would expect the term vector to be sparse.

We use Weka's SMO implementation with the default kernel, PolyKernel, with its default options. The results achieved, shown in Table 5.6, are moderately better than our OneR baseline, with an improvement of 7.19% in AUC.

5.2.7 Multilayer Perceptrons

A *perceptron* is a function that takes a vector as input and produces a binary output [Rosenblatt, 1958]. The attribute values of an instance for classification, such as a vector describing the presence or absence of terms in the instance as described above, can be represented as such a vector. This vector can be used as input to a perceptron, where the output would be the classification decision. Multiple perceptrons can be combined to form arbitrarily complex logical expressions [Witten and Frank, 2005]. A combination of multiple perceptrons is called a *multilayer perceptron*, and it is one of the most popular forms of neural network.

We use Weka's MultilayerPerceptron package with default settings. We find that the MultilayerPerceptron performs substantially better than our baseline measures, showing a

Table 5.6: Performance of SMO classifier using text extracted from resources. The number in brackets shows the percentage change in AUC compared to the OneR baseline.

Measure	SMO
AUC	0.8069 (7.19%)
Accuracy	0.8180
Recall	0.8069
Precision	0.8095
F-Measure	0.8079
kappa	0.6160
time (seconds)	2.1026

21.53% improvement in AUC. However, the evaluation run was extremely slow, taking almost a full day to complete a single cross-validation run.

5.2.8 Boosting

Several classes of algorithm have been developed that combine the output multiple classifiers to improve classification performance. These are often called *ensemble* classifiers. One ensemble method is *boosting*, a method developed to combine *weak classifiers*, classifiers that perform marginally better than random guessing [Schapire, 1990]. Boosting algorithms combine the output of classifiers that perform well on different input instances. When using training data to build a classifier to be used in a boosting algorithm, weight is added to instances that were incorrectly classified by the classifiers already built. This makes the new classifier more likely to correctly classify previously incorrect instances. When classifying new instances, the output of each classifier acts as a vote towards the final class value.

In our investigation of boosting, we examine the AdaBoost boosting algorithm [Freund and Schapire, 1996], as implemented in Weka with `AdaBoost.M1`. For the underlying classifier, we use the `DecisionStump`, a single-level decision tree. Table 5.8 shows that this algorithm performs extremely well on our development collection, improving over the OneR baseline's AUC by 22.78%.

Table 5.7: Performance of Multilayer Perceptron classifier using text extracted from resources. The number in brackets shows the percentage change in AUC compared to the OneR baseline.

Measure	Multilayer Perceptron
AUC	0.9149 (21.53%)
Accuracy	0.8440
Recall	0.8291
Precision	0.8403
F-Measure	0.8336
kappa	0.6676
time (seconds)	82615.4117

Table 5.8: Performance of AdaBoost classifier using text extracted from resources. The number in brackets shows the percentage change in AUC compared to the OneR baseline.

Measure	AdaBoost
AUC	0.9243 (22.78%)
Accuracy	0.8860
Recall	0.8756
Precision	0.8840
F-Measure	0.8789
kappa	0.7580
time (seconds)	18.4355

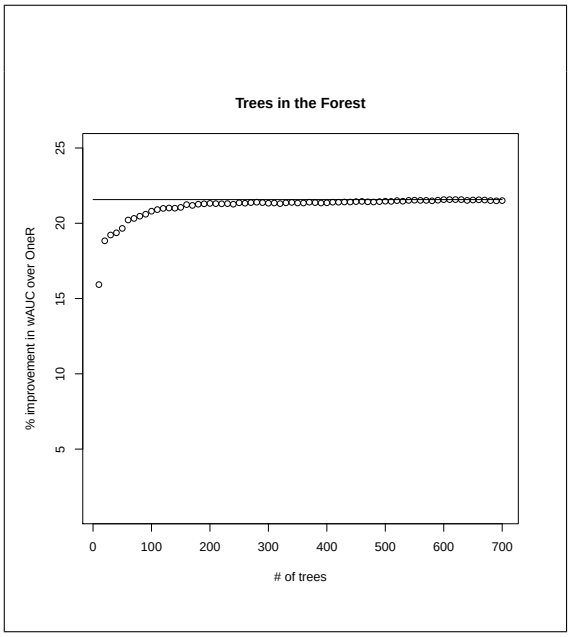


Figure 5.7: Performance of Random Forest classifier using text extracted from resources varying the number of tree classifiers used in the constructed the model.

5.2.9 Bagging

Another popular ensemble method is *bootstrap aggregating*, more commonly known as *bagging* [Breiman, 1996], which resamples from the training set to produce new training sets to build individual classifiers. These individual classifiers are then combined using unweighted votes, with test instances being assigned to the class that receives the most votes. Classifiers used for bagging are built independently, as opposed to classifiers produced for boosting, which are weighted based upon the previously built classifiers in the ensemble.

The Random Forest classification algorithm is a bagging algorithm that combines the results of multiple decision trees [Breiman, 2001]. In Weka's Random Forest implementation, which we use to investigate bagging, the default number of trees in the forest is 10, and we varied this number in increments of 10 up to 700. The results are shown in Figure 5.7. The largest improvement in AUC over the OneR baseline is seen at 600 trees, where AUC is 0.9152, an improvement of 21.57%, and is marked on the graph as a horizontal line, and the results of this model are show in Table 5.9.

Table 5.9: Performance of Random Forest classifier using text extracted from resources and 600 tree classifiers in the constructed the model. The number in brackets shows the percentage change in AUC compared to the OneR baseline.

Measure	Random Forest
AUC	0.9152 (21.57%)
Accuracy	0.8450
Recall	0.8272
Precision	0.8437
F-Measure	0.8334
kappa	0.6675
time (seconds)	112.1671

5.2.10 Candidate models

In terms of classification, the best-performing classifiers are the Naïve Bayes with discretization, AdaBoost, Random Forest, and Multilayer Perceptron. Given the extremely long evaluation time of the Multilayer Perceptron, exceeding 22 hours while most other classifiers complete in less than a minute and the longest running of the remaining classifiers being the 600 tree Random Forest which takes approximately 20 minutes, we do not consider the Multilayer Perceptron further. We therefore use Naïve Bayes with discretization, AdaBoost, and Random Forest to investigate tuning of the vectorisation process and the development of other features.

In the following section, we discuss tuning the vectorisation process and investigate whether this tuning leads to increased performance. For easier reference and comparison, the performance of these models using the default vectorisation process is shown together in Table 5.10.

5.2.11 Tuning vectorisation

When preparing the resources of the development collection to build the above classifiers, we performed simple vectorisation of the text. However, there are many parameters to the

Table 5.10: The performance of the most effective models using text extracted from resources and the default vectorisation processes. The number in brackets shows the percentage change in AUC compared to the OneR baseline.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.9032 (19.98%)	0.9243 (22.78%)	0.9152 (21.57%)
Accuracy	0.8540	0.8860	0.8450
Recall	0.8484	0.8756	0.8272
Precision	0.8462	0.8840	0.8437
F-Measure	0.8470	0.8789	0.8334
kappa	0.6941	0.7580	0.6675
time (seconds)	2.6013	18.4357	112.1452

vectorisation that may change the resulting term vector. We investigate those parameters here.

Stopping

Words that are common and likely to appear often in multiple classes, such as "at" or "yourself", are not useful when trying to classify documents. The removal of these words is called *stopping* [Baeza-Yates and Ribeiro-Neto, 1999]. In addition to the potential removal of terms that may be unhelpful for classification, another benefit of stopping is that it can reduce the size of the word vector, which in turn will decrease the amount of time required to train each classifier. A full list of the words we removed in our investigation of stopping can be found in Appendix H.

We show the results of stopping during the vectorisation process in Table 5.11. The number in curly brackets shows the effect of stopping compared to the default vectorisation process, shown in Table 5.10, as we are no longer comparing the classifier performance to the OneR baseline. Each classifier suffers a degradation in performance when stopping is performed.

Table 5.11: Using stopping to tune the performance of the most effective models using text extracted from resources. The number in curly brackets shows the percentage change in AUC compared to the default vectorisation.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.8580 {-5.00%}	0.8643 {-6.49%}	0.9021 {-1.43%}
Accuracy	0.7830	0.8020	0.8330
Recall	0.7505	0.7716	0.8053
Precision	0.7862	0.8065	0.8415
F-Measure	0.7591	0.7811	0.8161
kappa	0.5232	0.5661	0.6349
time (seconds)	2.4632	17.9302	135.1506

Stemming

Many variants of words will be used across a collection of text resources, such as plurals and past tense suffixes. Stemming involves the reduction of words to their stem or root form. For example, the words "ending" and "ends" would be reduced to the stem "end". The stem is not always a word in its own right, as with the words "retrieval" and "retrieved", which would be reduced to the common stem "retriev". The stems are then treated as the same term for vectorisation purposes. As with stopping, stemming can reduce the number of terms in the output vector. When stemming, we use the Snowball stemming algorithm [Porter, 2001].

The results of stemming during the vectorisation process are shown in Table 5.12. The Random Forest classifier shows no change in performance, while each of the other classifiers displays poorer performance when stemming is performed.

Word frequencies

In our previous vectorisation we used the binary occurrence of a term in a document; either a particular term was in a document or it was absent. However, the number of times a term appears in a document may provide useful information. As such, we investigate using term counts instead of binary presence. However, this introduces the issue of document length, as

Table 5.12: Using stemming to tune the performance of the most effective models using text extracted from resources. The number in curly brackets shows the percentage change in AUC compared to the default vectorisation.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.9039 {-1.13%}	0.8647 {-6.72%}	0.9134 {0%}
Accuracy	0.8570	0.8290	0.8490
Recall	0.8541	0.8127	0.8328
Precision	0.8486	0.8247	0.8473
F-Measure	0.8509	0.8172	0.8379
kappa	0.7020	0.6350	0.6766
time (seconds)	2.9531	20.3987	133.3467

longer documents will have more terms. We can correct for this by normalising for document length. Used with binary values, this effectively weights attributes by the length of the resource in which they occur; that is, all attributes in an instance will have a value of either zero or, given that the document length is constant within an instance, the same normalised number. Used with word counts, normalisation weights attributes by both the length of the document and the number of times they appear in a resource.

Given that we hypothesised that Naïve Bayes with discretization outperformed Naïve Bayes with kernel estimation due to fact that attribute values were binary, we reintroduce kernel estimation when exploring non-binary values.

Table 5.13 shows the results of changing the vectorisation method to use word counts and document length normalisation. The Naïve Bayes classifier using discretization and the Random Forest classifier show a small performance improvement when word counts are used in place of the binary presence or absence of terms and the word count is normalised to take into account document length. Under all other settings, performance is degraded. We observe that Naïve Bayes using kernel estimation substantially under performs compared to using discretization when word counts are used in place of binary values, and therefore we do not investigate it further.

Table 5.13: Using the number of words in resources to tune the performance of the most effective models using text extracted from resources. The number in curly brackets shows the percentage change in AUC compared to the default vectorisation.

	Naïve Bayes			
Measure	Discretized	Kernel Estimated	AdaBoost	Random Forest
Word Counts				
AUC	0.8973 {-0.65%}	0.7210 {-20.17%}	0.9031 {-2.29%}	0.9085 {-0.73%}
Accuracy	0.8380	0.7060	0.8430	0.8500
Recall	0.8256	0.6827	0.8348	0.8354
Precision	0.8318	0.6904	0.8354	0.8468
F-Measure	0.8282	0.6854	0.8349	0.8399
kappa	0.6566	0.3717	0.6698	0.6802
time (seconds)	2.6681	2.7722	19.3025	103.8198
Normalise				
AUC	0.8718 {-3.48%}	0.8488 {-6.02%}	0.9030 {-2.30%}	0.9151 {-0.01%}
Accuracy	0.8370	0.8040	0.8640	0.8430
Recall	0.8359	0.8125	0.8511	0.8278
Precision	0.8280	0.7988	0.8614	0.8393
F-Measure	0.8309	0.8000	0.8552	0.8324
kappa	0.6621	0.6028	0.7107	0.6652
time (seconds)	2.8263	2.7980	20.6016	97.8341
Word count & normalise				
AUC	0.9089 {0.63%}	0.8149 {-9.78%}	0.8758 {-5.25%}	0.9191 {0.43%}
Accuracy	0.8790	0.7710	0.8170	0.8330
Recall	0.8832	0.7707	0.8056	0.8146
Precision	0.8711	0.7619	0.8087	0.8305
F-Measure	0.8752	0.7641	0.8067	0.8204
kappa	0.7508	0.5295	0.6136	0.6417
time (seconds)	2.8252	3.0742	20.7486	97.4984

Table 5.14: Using term frequency and inverse document frequency to tune the performance of the most effective models using text extracted from resources. The number in curly brackets shows the change in AUC compared to the default vectorisation.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.8973 {-0.65%}	0.9032 {-2.28%}	0.9096 {-0.61%}
Accuracy	0.8380	0.8430	0.8440
Recall	0.8256	0.8348	0.8287
Precision	0.8318	0.8354	0.8403
F-Measure	0.8282	0.8349	0.8334
kappa	0.6566	0.6698	0.6672
time (seconds)	2.6610	19.2754	103.1244

Term frequency and inverse document frequency

Term frequency and inverse document frequency are measures used to estimate the importance of a term in a resource. Term frequency is based on the intuition that the more times a term appears in a resource, the more important it is to that resource, while inverse document frequency is based on the intuition that the more a word appears in the collection as a whole, the less discriminating that term is [Sebastiani, 2002]. As can be seen in Table 5.14, $TF \cdot IDF$ does not provide any improvement over the default vectorisation parameters.

Of the vectorisation methods described, only word count and normalisation for Naïve Bayes and Random Forest show improvement over the default methods described in Subsection 5.2.1.

We examine the characteristics of the vectors produced by the methods described to ensure that the methods are producing differing vectors using these methods. These results were produced by sorting the term vector by total number of occurrences in the collection. Table 5.15 shows the total number of terms produced using the default vectorisation method, stopping, and stemming. Producing term vectors based upon counting word occurrences in documents rather than using binary presence or absence, normalising by document length, and weighting values based on frequency, all produce the same number of terms as the default

Table 5.15: Total terms in development collection using different vectorisation methods.

Method	Total Terms
Default	12475
Stopping	12000
Stemming	8624

Table 5.16: Number of terms in common between different vectorisation methods when retaining the 1000 most common terms.

Method (total terms)	Default	Stopping
Default (1031)		
Stopping (1044)	797	
Stemming (1014)	488	359

method. That is, only the values assigned to terms are being manipulated, not the terms themselves. We see that 475 words have been removed from the term vector after stopping. Stemming has reduced the term vector by almost a third. While the vector after stopping is a subset of the default vector, this is not the case with stemming.

Given the default 1000 terms, including ties, Table 5.16 shows the number of common terms between the default method, stopping, and stemming. Again, the other vectorisation methods, counting word occurrences in documents, normalising by document length, and weighting values based on frequency, contain the same terms as the default. Stopping has increased the total number of terms, due to an increase in the number of tied terms at the 1000th position. The table shows that, while there is some term overlap, the vectors produced are not identical.

We also investigated the effect of altering the length of the term vector. Figure 5.8 shows the performance of the candidate classifiers using different vectorisation methods at various vector lengths. Using word counts and document length normalisation consistently performs better for the Naïve Bayes and Random Forest classifiers, while binary presence or absence performs best for AdaBoost. The performance of Naïve Bayes rises slightly with vector

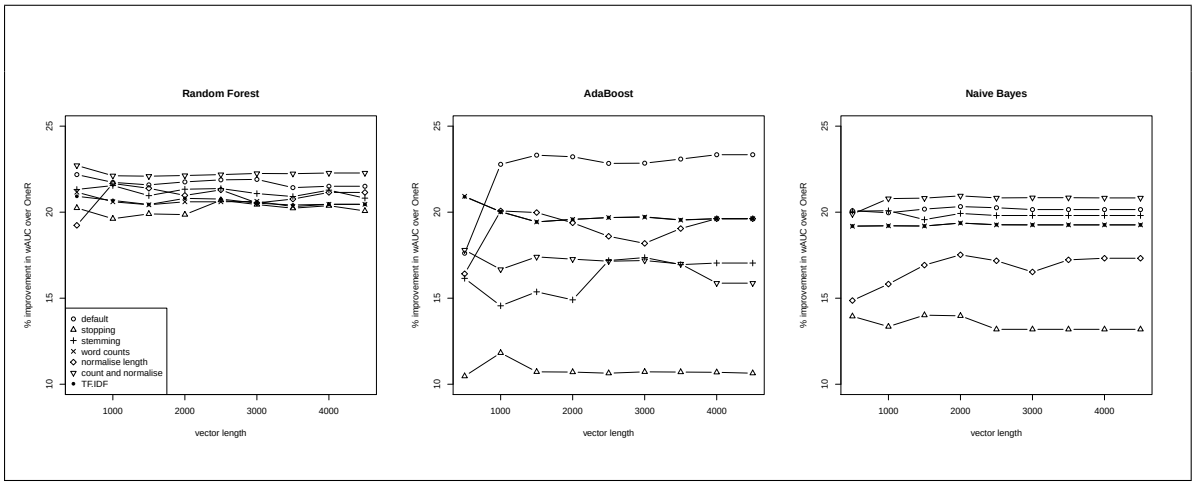


Figure 5.8: Performance of classifiers built using text extracted from resources and varying the vector length for different vectorisation tuning methods

length, peaking at an AUC of 0.9105 with 2000 terms, AdaBoost's performance is relatively steady after 1000 terms, and AUC peaks at 0.9285 with 4000 terms, and Random Forest's performance length grows. We use the best-performing method and vector lengths for each classifier when developing further attributes in the following section.

5.3 Developing further attributes

In the previous section we investigated using the words that occur in a collection of resources to determine whether or not a resource is educational. Inspection of the resources in the development collection suggested other features that might be used to differentiate between educational and non-educational resources. Those features are explored in this section, using the candidate algorithms identified in Subsection 5.2.10: Naïve Bayes, AdaBoost, and Random Forest.

We add each new feature to the text classification described in the previous section, and investigate its effect. As each algorithm performed optimally with a different number of terms, we run all classification algorithms at each length and report AUC.

5.3.1 Hyperlinks

Hyperlinks are fundamental to the Web, and analysis of link text and link structure in hyperlinked collections has been effective for retrieving and ranking documents [Kleinberg, 1999; Page et al., 1998]. We hypothesise that the way that links are used in a resource may help to indicate the type of resource it is, and explore several features related to links that can be extracted from web resources.

Number of internal links

Hyperlinks can be used to point to a specific named position within a resource. These named positions are called fragment identifiers. We hypothesise that educational resources have more internal links than resources that are not educational. For example, educational resources often have a table of contents at the beginning that points to sections within the document.

A fragment identifier is represented as an HTML anchor with a `name` attribute. Links can link directly to the fragment identifier using a hash followed by the fragment identifier's name.

To calculate the value of this single attribute, we simply count the number of links whose end point, or `href` attribute value, begins with a hash. It is possible to link to an internal fragment identifier within the same resource using an absolute URL rather than a relative URL, but as this did not occur in any of the resources in the development collection, we do not take these into account.

The results of including an attribute representing the number of internal links to our tuned text classification models is shown in Table 5.17, which shows no change in the classification performance of the Naïve Bayes classifiers, but the performance of the AdaBoost and Random Forest classifiers worsen.

Outgoing link count

Another simple feature related to hyperlinks is the number of times a resource links to external resources. We hypothesise that educational resources have fewer external links than non-educational resources. For example, resources such as product search results have many links to the search results themselves.

Table 5.17: Performance of classifiers using text extracted from resources and the number of internal links. The number in square brackets shows the percentage change in AUC compared to using text alone.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.9105 [0%]	0.9270 [-0.16%]	0.9186 [-0.56%]
Accuracy	0.8770	0.8840	0.8390
Recall	0.8779	0.8726	0.8199
Precision	0.8689	0.8823	0.8378
F-Measure	0.8724	0.8766	0.8266
kappa	0.7451	0.7535	0.6540
time (seconds)	5.3778	89.1380	167.9821

This attribute is calculated by taking the total number of links in the resource and subtracting the number of internal links.

The results of including an attribute representing the number of outgoing links to our tuned text classification models is shown in Table 5.18. Again, Naïve Bayes shows no change in performance, while AdaBoost and Random Forest show a performance degradation.

Ratio of link text to overall text

The length of resources is likely to be correlated with the number of links they contain, with longer documents having more links. Therefore, it may be useful to normalise the link count to take into account document length. We also observe that many resources that are not educational appear to have a higher proportion of links to text.

We take the text that appears in link text and remove all characters that are not alphabetical or whitespace. The number of terms in the link text (after splitting by whitespace) represents the number of link terms. We similarly process the entire text of the document, and divide the number of link terms by the total number of terms in the document, producing a single attribute with a float value.

The results of including an attribute representing the ratio of link text to total text to

Table 5.18: Performance of classifiers using text extracted from resources and the number of outgoing links. The number in square brackets shows the percentage change in AUC compared to using text alone.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.9105 [0%]	0.9248 [-0.40%]	0.9188 [-0.54%]
Accuracy	0.8770	0.8810	0.8230
Recall	0.8779	0.8692	0.7994
Precision	0.8689	0.8793	0.8238
F-Measure	0.8724	0.8733	0.8074
kappa	0.7451	0.7469	0.6166
time (seconds)	5.4934	91.6085	170.0562

our tuned text classification models is shown in Table 5.19. All classifiers showed a drop in classification performance.

Outgoing link text

As with the textual content of the resources themselves, the textual content of outgoing links may help distinguish between educational and non-educational. For example, a resource advertising a book for purchase may have links saying "Buy now," whereas an educational resource may link to other resources with the text, "Learn more."

To investigate this feature we extracted the text content of links, excluding links that link to a fragment identifier within the resource itself, and created a word vector using the same method as the default vectorisation used for the text content of resources, as described in Section 5.2.1. The resulting term vector, which consisted of 1084 attributes, was treated independently of the term vector for text in the resource.

The results of including attributes based on link text to our tuned text classification models is shown in Table 5.20. The Random Forest and AdaBoost classifiers again drop in performance, but Naïve Bayes shows a small improvement in performance.

Table 5.19: Performance of classifiers using text extracted from resources and the ratio of normal text to link text. The number in square brackets shows the percentage change in AUC compared to using text alone.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.9102 [-0.03%]	0.9245 [-0.43%]	0.9202 [-0.39%]
Accuracy	0.8750	0.8820	0.8200
Recall	0.8758	0.8704	0.7951
Precision	0.8669	0.8800	0.8217
F-Measure	0.8703	0.8744	0.8037
kappa	0.7410	0.7491	0.6094
time (seconds)	5.5124	91.3059	170.1142

Table 5.20: Performance of classifiers using text extracted from resources and text of outgoing links. The number in square brackets shows the percentage change in AUC compared to using text alone.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.9117 [0.13%]	0.9278 [-0.08%]	0.9191 [-0.51%]
Accuracy	0.8760	0.8810	0.8250
Recall	0.8743	0.8692	0.8011
Precision	0.8683	0.8795	0.8267
F-Measure	0.8708	0.8735	0.8094
kappa	0.7417	0.7471	0.6207
time (seconds)	8.1649	108.5252	274.6863

5.3.2 Headings

Headings provide structure and organisation to web pages. Investigation of our development collection suggests that headings are used differently in resources judged educational.

HTML defines six levels of heading, with `H1` being the highest level and `H6` being the lowest level, and browsers by default typically render these headings differently from each other.

One difficulty of using header information to make inferences about the intended structure of a resource is that, while a link is a functional element, a heading is intended to have semantic significance. It is possible that other less semantically appropriate markup could be used, possibly in conjunction with cascading style sheets, to achieve a similar visual result as using a heading element. In addition, headings could be used in resources for reasons other than to provide structure.

Heading count

The educational resources in our development collection are often divided into sections using headings to delineate their structure. We therefore hypothesise that educational resources will have a higher ratio of headings to text than resources that were not judged to be educational.

We produce a single attribute by counting all occurrences of heading elements `H1` through to `H6`.

The results of including an attribute based on the number of headings in resources to our tuned text classification models is shown in Table 5.21. The addition of this attribute makes no difference to the performance of the Naïve Bayes classifier, while Random Forest and AdaBoost deteriorate.

Ratio of heading text to overall text

As with links, the number of headings in a document may be correlated with the length of the document, and it may be informative to normalise this data to take into account document length.

A single float-valued attribute is produced using the same method as was used to calculate the ratio of link text to overall text, using the text that occurs in `H1` through to `H6` elements.

Table 5.21: Performance of classifiers using text extracted from resources and the number of headings. The number in square brackets shows the percentage change in AUC compared to using text alone.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.9105 [0%]	0.9270 [-0.16%]	0.9190 [-0.52%]
Accuracy	0.8780	0.8840	0.8240
Recall	0.8787	0.8726	0.8003
Precision	0.8700	0.8823	0.8254
F-Measure	0.8734	0.8766	0.8084
kappa	0.7471	0.7535	0.6187
time (seconds)	5.4034	89.7834	168.7859

The results of including an attribute based on the ratio of heading text to total text to our tuned text classification models is shown in Table 5.22. Adding an attribute based upon the ratio of heading text to all text shows the similar changes in performance as was seen when adding heading count.

Tables 5.23 and 5.24 show the effect of using all the further features developed in this section, in addition to and instead of resource text respectively. Both the Naïve Bayes and AdaBoost classifiers show a small loss of performance when all further features are added, and the magnitude of improvement of the Random Forest classifier is lessened when compared to adding the further features individually. As expected, not using the text in resources leads to a substantial drop in classification performance.

5.3.3 Performance of further attributes

The additional attributes developed in this section affect the performance of each of our classifiers differently. When added individually to our tuned models based upon the text in resources, all attributes result in consistent degradation of performance for the AdaBoost and Random Forest classifiers. The Naïve Bayes classifier showed no change with most attributes, a small improvement when the text in outgoing links was used to create additional attributes,

Table 5.22: Performance of classifiers using text extracted from resources and the ratio of normal text to heading text. The number in square brackets shows the percentage change in AUC compared to using text alone.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.9105 [0%]	0.9265 [-0.22%]	0.9193 [-0.49%]
Accuracy	0.8770	0.8830	0.8240
Recall	0.8779	0.8713	0.8012
Precision	0.8689	0.8816	0.8241
F-Measure	0.8724	0.8755	0.8089
kappa	0.7451	0.7512	0.6194
time (seconds)	5.4006	89.7661	168.5918

Table 5.23: Performance of classifiers using text extracted from resources and all additional features described in this section. The number in square brackets shows the percentage change in AUC compared to using text alone.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.9103 [-0.02%]	0.7814 [-15.84%]	0.7730 [-16.32%]
Accuracy	0.8700.0.7340	0.7150.	
Recall	0.8648.0.7006	0.6568.	
Precision	0.8632.0.7277	0.7329.	
F-Measure	0.8637.0.7061	0.6569.	
kappa	0.7274.0.4178	0.3432.	
time (seconds)	11.5355	207.3606	434.8838

Table 5.24: Performance of classifiers using only additional features described in this section. The number in square brackets shows the percentage change in AUC compared to using text alone.

Measure	Naïve Bayes	AdaBoost	Random Forest
AUC	0.6187 [-32.05%]	0.7802 [-15.97%]	0.7156 [-22.54%]
Accuracy	0.6430.0.7210	0.6780.	
Recall	0.5714.0.6830	0.6293.	
Precision	0.6284.0.7134	0.6644.	
F-Measure	0.5518.0.6881	0.6298.	
kappa	0.1600.0.3841	0.2763.	
time (seconds)	2.1452	15.5413	240.4936

and a small drop in performance when considering the ratio of link text to the overall text in a resource. Thus the AdaBoost classifier remains the best performing classifier, with an AUC score of 0.9285.

Combining the further features has a detrimental effect on the classification performance of all the classifiers. Removing resource text as a feature leads to much worse performance.

5.4 Summary

In this chapter we described our investigation of features that might help differentiate between resources that are likely to support learning.

We found that reasonable classification performance can be achieved using text extracted from resources in conjunction with Naïve Bayes, AdaBoost, and Random Forest classifiers, which each show a classification performance improvement of about 20% over the OneR baseline. The AdaBoost algorithm's best performance was achieved when vectorising using the binary presence or absence of terms in resources, whereas Naïve Bayes and Random Forest performed best when word counts and normalisation were used.

We also found that, when added individually to a classification model using resource text, attributes developed from the links and headings found in a resource rarely improved

classification over simply using the text, and often caused degradation in the performance. Not using resource text as a model feature caused substantially poorer performance.

The results reported in this chapter were based on an analysis of a small development collection of resources, with judgments made by the author. In the following chapter, we describe the construction of a larger collection of independently judged resources. We use that collection, along with the classification models described above, to validate the findings of this chapter.

Chapter 6

Validating the Effectiveness of Machine Learning Techniques for Filtering Educational Resources

In Chapter 4 we described a user experiment investigating the construction of collections to evaluate systems that filter educational resources. We went on in Chapter 5 to develop and tune machine learning classification models to build such a filtering system. This chapter draws the previous two chapters together, describing a further user experiment to construct a collection of resources that we use to validate our machine learning models.

Specifically, we continue to explore methods to effectively filter educational resources by investigating the effectiveness of the techniques presented in Chapter 5 when applied to a larger, independent validation collection. We expect that, as was found during the development of classification models, the performance of the tuned Naïve Bayes, AdaBoost, and Random Forest classifiers using a term vector developed from resource text will be better than the OneR baseline. Further, we hypothesise that AdaBoost will outperform the Naïve Bayes and Random Forest classifiers. We also expect that attributes derived from hyperlinks and headings described in Section 5.3 will not provide any improvement over text alone.

While building our validation collection, to gain insight into the judging process, we collect other data about the judges and the judging process. This includes judges' expertise in the subject areas of the resources, the ease with which judgments were made, and how

confident they are that their judgments are correct.

Finally, when a judge decides that a resource is educational, we ask them to specify the educational depth of the resource on a 6-point scale.

This chapter is organised as follows. In Section 6.1 we describe our construction of a collection of resources to be used for validating our filtering system, including aspects of the judging process. In Section 6.2 we use the validation collection to test our hypotheses about classifier performance. We summarise the chapter in Section 6.3.

6.1 Constructing a validation collection

Following on from the approach described in Chapter 4, we again asked people to judge whether resources were likely to support learning. We begin this section by describing how these judgments were collected, including how this differed from our previous approach.

Our primary goal was to build a collection of judged resources that could be used to evaluate the effectiveness of our system for filtering search results for educational material, in an attempt to allow us to draw general conclusions about the effectiveness of our system. The collection was constructed after the development and tuning of our system, described in Chapter 5, to avoid any possibility that the design of our system would be biased towards correctly classifying resources in this validation collection.

In addition to building a large, independent validation collection, we also asked participants to answer questions about their experience of the judging process and the system for collecting judgments.

6.1.1 User task design

Though our previous work showed that judges are able to categorise resources as educational in the absence of context, they found the task difficult. To make the judging task simpler, in this task we provided the following context to participants.

> *In this task, you will be presented with resources from the Web and asked to judge whether or not each resource is likely to support the learning of a student in high school.*

Queries were sourced from the curriculum for a learner in years 9 or 10 in the school system in the state of Victoria, Australia. This corresponds with Level 6 of the Victorian curriculum as at the time of writing, as described in the Victorian Essential Learning Standards (VELS). The VELS documentation,[1] presents curriculum subject areas that a learner in years 9 or 10 will cover. These subject areas are presented in sections, and the headings for each section, for example *movement and physical activity* and *historical knowledge and understanding*, were used as queries. This resulted in 41 queries, which are listed in Appendix G.4.

In Chapter 4, we showed that a reasonable distribution of educational material can be expected in the top ten results for many queries. As such, the 41 queries were submitted to the Yahoo! Search API, with results limited to HTML documents without framesets that did not come from the VELS website. The first ten results conforming to these limitations for each query were included in the judgment pool, producing a collection of 410 resources.

To recruit judges, a recruitment email, which is in Appendix G.1, was sent to all students enrolled in one undergraduate degree in the School of Computer Science and Information Technology at RMIT University. However, this method did not result in the recruitment of sufficient judges, so additional judges were recruited through requests for volunteers in lectures and laboratory sessions.

To ensure that potential judges were familiar with the context being presented for the judging process, eligibility was restricted to students who had completed the Victorian Certificate of Education, the high school certificate offered in Victoria. This restriction was relaxed to completion of high school in Australia to allow for the recruitment of sufficient judges.

Prior to their participation, judges were presented with a plain language statement, which is in Appendix G.2, describing the research. They were also asked to sign a form consenting to the data gathered from their judgment session being used in the research, as set out in the plain language statement.

In Chapter 4 we showed that there is significant agreement between judges on whether resources are educational, and concluded that, when limited resources are available for judging, it is preferable to assess more resources than to have multiple judges assess each resource. Therefore, in these assessments, a single judge assessed each resource. Judges completed the

[1] http://vels.vcaa.vic.edu.au/downloads/vels_standards/velsrevlvl6.pdf, accessed on 10 April 2010

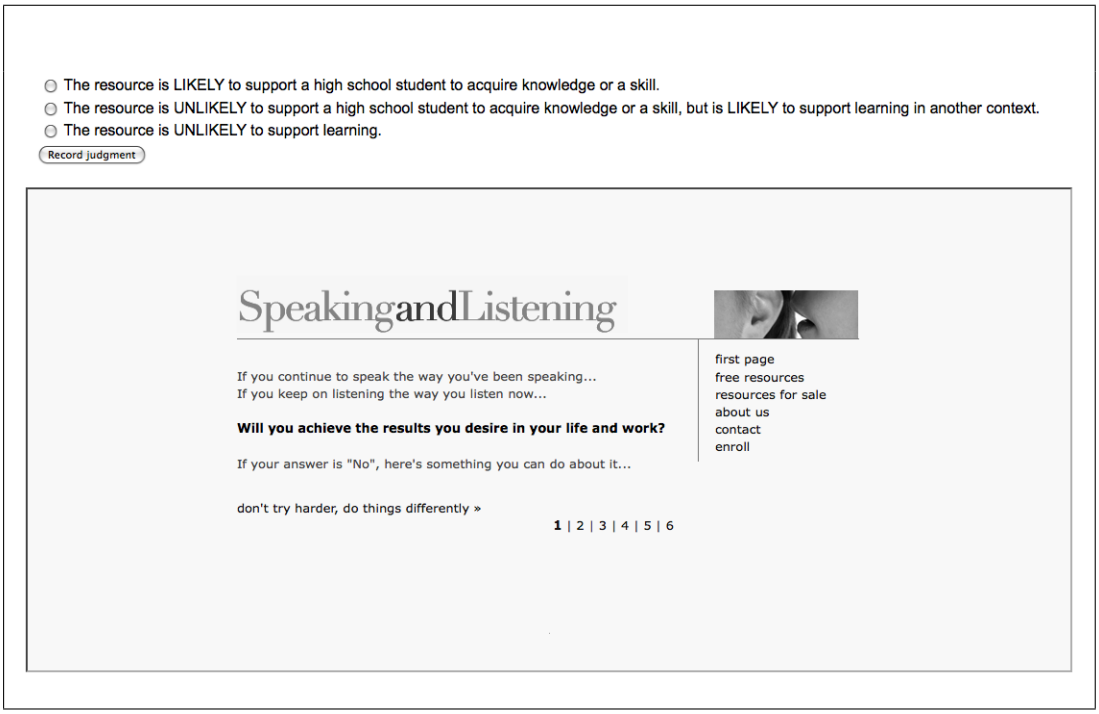


Figure 6.1: The judgment interface used in the construction of the validation collection.

judgments individually, either alone or in small groups, in a laboratory setting under the supervision of the author.

Our work in Chapter 4 showed that raters were more likely to judge that a given resource was educational if the query used to retrieve that resource was displayed, possibly because they were judging whether the resource was relevant to the displayed query rather than whether it was educational. In this task we therefore display only the context as described above and the query was not displayed. The judgment interface was similar to that described in Section 4.2.2. Resources were again presented sequentially via a web interface, shown in Figure 6.1, with the resource to be judged being randomly selected from the pool of unjudged resources. Judges were asked to classify resources according to the following choices:

- The resource is LIKELY to support a high school student to acquire knowledge or a skill.
- The resource is UNLIKELY to support a high school student to acquire knowledge or a skill, but is LIKELY to support learning in another context.
- The resource is UNLIKELY to support learning.

An HTML `iframe` element was used to embed single page resources, with links and form submission within the resources disabled. Raters again evaluated the resources without reference to other web pages. To avoid issues such as frame-busting, where Javascript is used to prevent a page from being viewed in an `iframe` by reloading the framed page as the top page, resources were pre-processed to remove HTML script elements.

To attempt to combat possible fatigue effects [Kelly, 2009], after every five resources participants were presented with a suggestion that they should take a short break.

6.1.2 Judgments

We recruited 21 judges, with 20 judges judging 20 resources each and one judge judging the remaining 10. Of the 410 resources, 169 were judged likely to support learning for a high school student, 108 were judged likely to support a learner in other contexts, and 133 were judged unlikely to support learning. As we are interested in all educational resources, there are 277 resources likely to support learning, independent of the context. There were a total of 53 432 terms in the collection, with an average of 656 terms per resource. On average, resources contained 137 hyperlinks and 10 headings elements.

To identify if any participant demonstrated extreme bias, judging resources as a particular category much more often than other judges, we calculate the percentage of judgments that each participant made at each level. Figure 6.2 shows a boxplot of these judgments. The *sample range* is the interval containing all the data, and the *interquartile range* is the middle half of that data. An *outlier* is a data point that is outside one and a half times the interquartile range below the lowest or above the highest quartile [Rosenkrantz, 2009]. Our data has no outliers at any of the three levels of our scale. Given our findings in Chapter 4 that judges are interchangeable, this is expected.

After each judgment, judges were presented with statements about the judgment they had just made, to which they could respond on a four-point scale, *Strongly disagree*, *Disagree*, *Agree*, and *Strongly agree*. Figure 6.3 shows the frequency of responses for each, broken down by how they judged the resource. Participants reported that most of the judgments were made confidently and easily, and that they were generally not expert in the topic covered in the resource. The responses to these questions did not appear to be strongly affected by how the participant had judged the current resource, though where resources were judged

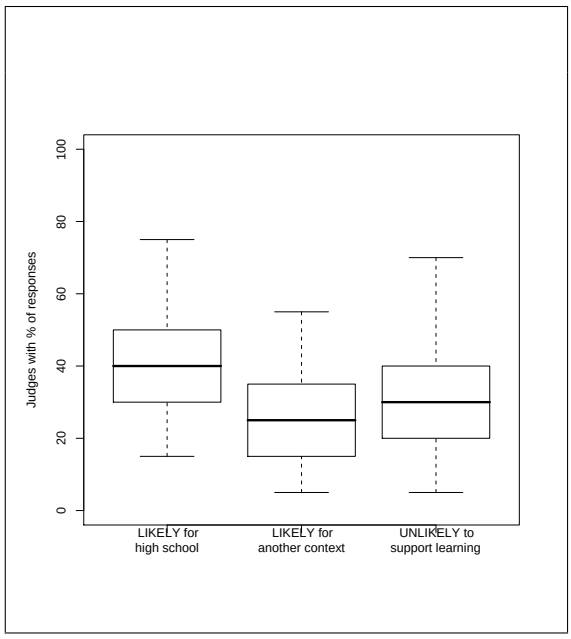


Figure 6.2: Percentage judgments that judges chose likely to support a high school student to acquire knowledge or a skill, likely to support learning in another context, or unlikely to support learning.

unlikely to support learning, there was a slight increase in the ease of judgment and how confident judges were that the judgment was correct.

Additionally, if the resource was judged as likely to support learning, either for learners in high school or in other contexts, participants were asked to estimate the educational depth of the resource. This estimate was collected on a 6-point scale developed from the cognitive dimension of Bloom's taxonomy [Bloom, 1956]. Specifically, participants were asked to respond to the following.

This resource helps learners ...

1. ... recall data or information.
2. ... explain ideas or concepts.
3. ... apply ideas or information in a new situation.
4. ... analyse the relationship between ideas or concepts.
5. ... develop a framework linking ideas or concepts.
6. ... make judgments about the value of ideas.

Figure 6.4 summarises the responses to this question. Resources judged likely to support learning were commonly assessed to be at the shallower end of the scale, with 55.6% at the shallowest two levels. Resources judged to be likely to support learners in high school were most common at all but the shallowest educational depth, where they comprised 32%. At other levels of depth, the percentage varied between 62.1% at the deepest level, to 79.3% at the fourth level. We revisit the effect of the educational depth of resources when discussing failure analysis in Subsection 6.2.1.

Comments made on the judgments indicate that participants were taking into account hyperlinks, page structure, and formatting when making judgments about the educational value of resources. In addition to structural clues, many comments made reference to the readability of resources, suggesting that participants were less likely to judge a resource as educational if it was not well written.

After participants completed judging all resources assigned to them, they were presented with statements about the judgment system and process, to which they could respond on

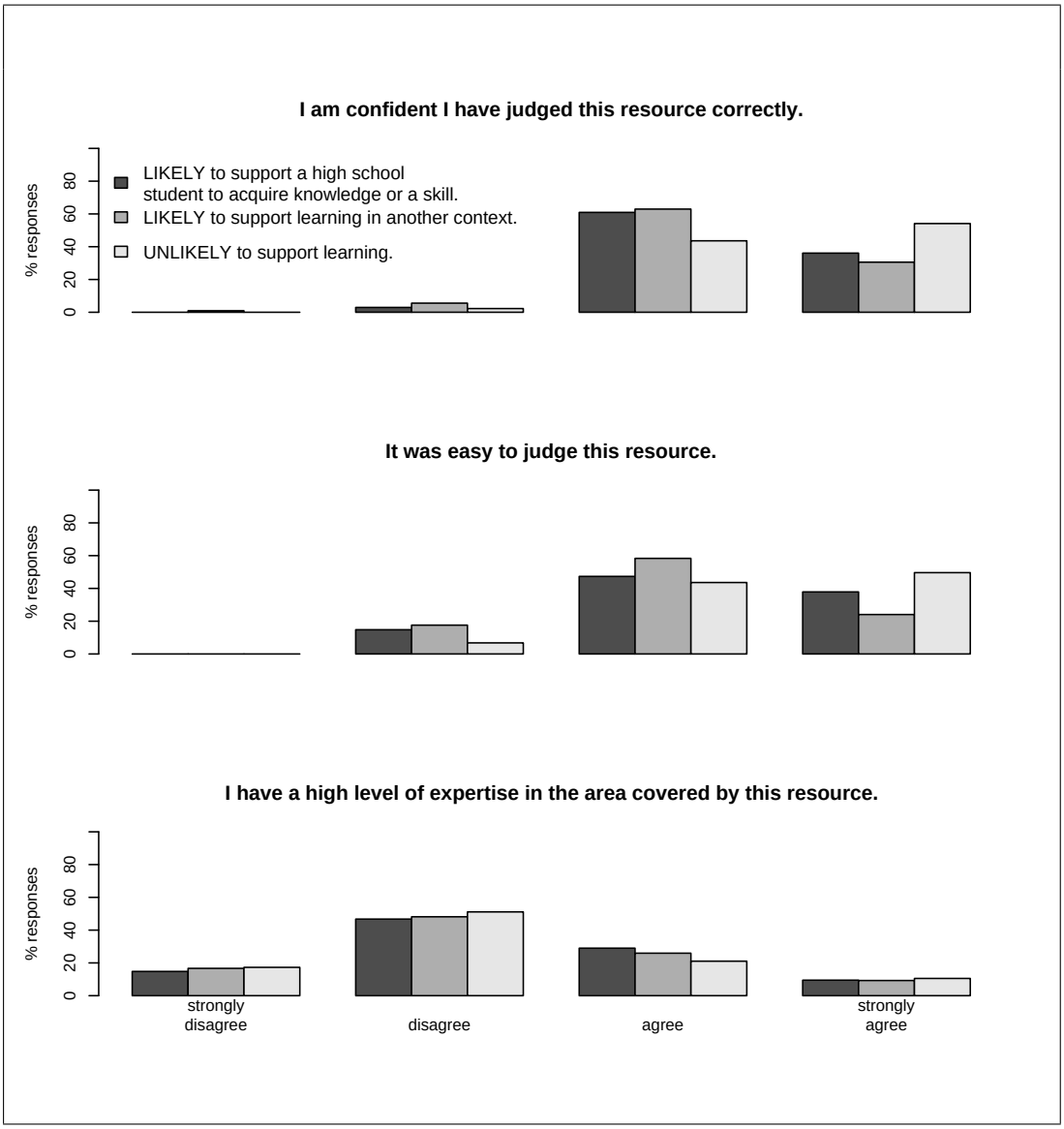


Figure 6.3: Responses to questions posed after each judgment, separated by how the resource was judged.

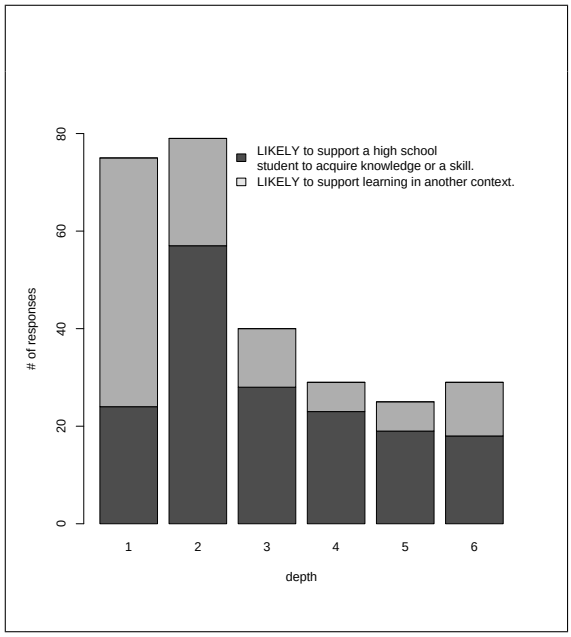


Figure 6.4: Educational depth of resources judged likely to support learning for high school students or in other contexts.

the same four-point scale as used after each judgment. The frequencies of responses for each question are shown in Figure 6.5.

All but two participants agreed that a search engine that returned educational resources would be useful and that they enjoyed the judging task. All judges agreed that the judging interface was easy to use.

Though all but two participants agreed that what they had to do was clear from the instructions, seven participants said they found the task confusing. Some confusion was apparent from requests for clarification made during the session and by written comments provided, which indicated that not all participants clearly understood that the task required them to judge each resource on its merits alone, and not other resources that may have been linked to or described. Two participants disagreed with the statement, "I consistenly used the same criteria to judge resources," though unfortunately neither elaborated on the inconsistency in their comments.

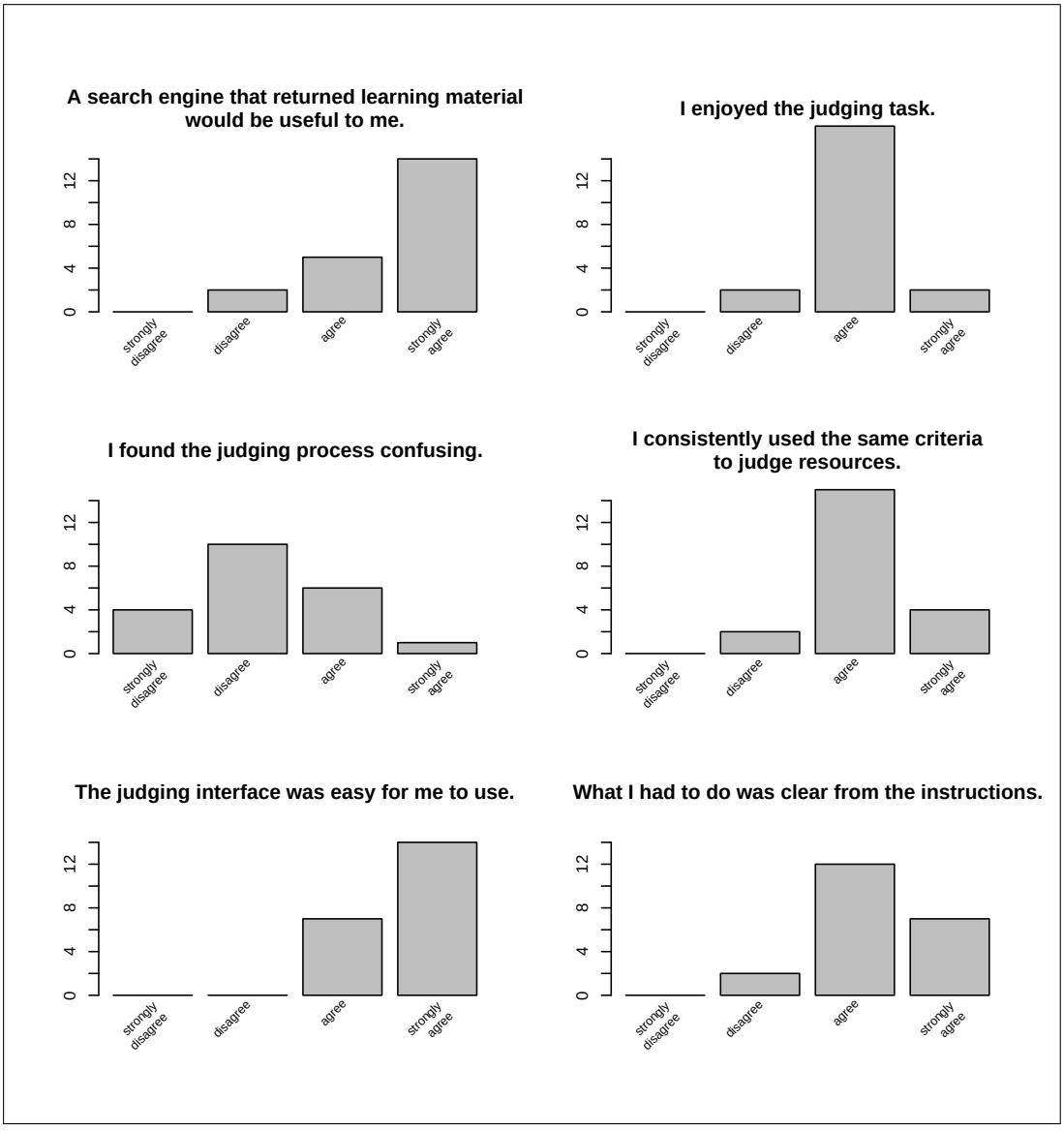


Figure 6.5: Responses to questions posed after all judgments have been performed.

6.2 Classifying resources as educational

Using the classifiers developed in the previous chapter, we classified the collection of judged resources. The Naïve Bayes classifier was constructed using normalised counts of terms occurring in resources, with a vector length of 2000 terms plus ties, the AdaBoost classifier was constructed using binary occurrence of terms and a vector length of 4000, and the Random Forest Classifier was constructed using normalised counts of terms and a vector length of 500 terms. In each case, the OneR classifier was induced from the same data used to build the classifier to which it was being compared.

Resources were judged as either likely to support learning in the context of learning in a high school, likely to support learning in some other unspecified context, or unlikely to support learning. As we are primarily interested in the ability of a system to differentiate educational resources regardless of context, we consider those resources judged likely to support learning together. This is therefore a two-class problem.

Results of classification using each model and OneR with identically vectorised input are shown in Table 6.1. As expected, all classifiers performed better than the OneR baseline. In each case, the observed difference was significant ($p < 0.01$). Although we expected the AdaBoost classifier to be the most effective, the low value of κ it achieved suggests it performs only marginally better than chance. The Random Forest gave the best performance, with an AUC of 0.7457, significantly outperforming the Naïve Bayes classifier ($p = 0.02$) and the AdaBoost classifier ($p < 0.01$). One possible explanation for this poorer than expected performance is that the tuning performed on the classifiers during development may have been dependent on the collection, and AdaBoost may have performed more effectively on the validation collection if the resource text had been transformed into a term vector of different length than the 4000 terms used here.

Table 6.2 shows the results of classification using the attributes developed based upon links and headings as well as text extracted from resources. As expected, adding these features does not improve the performance of classifiers over text alone, though each classifier still performs significantly better than the OneR classifier using the same input. The Naïve Bayes classifier performs best, with an AUC of 0.6808.

Table 6.1: Performance of classifiers on the validation collection as a two-class problem (education for high school students and others versus non-educational) based on text extracted from resources.

	Normalised word frequency 2000 terms	
Measure	OneR	Naïve Bayes
AUC	0.5054	0.6877 ($p < 0.01$)
Accuracy	0.6188	0.6105
Recall	0.5054	0.6191
Precision	0.5084	0.6045
F-Measure	0.4912	0.5954
kappa	0.0124	0.2113
time (seconds)	14.9688	22.8398
	Binary occurrence 4000 terms	
	OneR	AdaBoost
AUC	0.5243	0.6177 ($p < 0.01$)
Accuracy	0.6556	0.6595
Recall	0.5243	0.5067
Precision	0.5504	0.4962
F-Measure	0.4992	0.4489
kappa	0.0579	0.0162
time (seconds)	30.2319	346.2889
	Normalised word frequency 500 terms	
	OneR	Random Forest
AUC	0.5284	0.7457 ($p < 0.01$)
Accuracy	0.6261	0.7237
Recall	0.5284	0.6126
Precision	0.5379	0.7002
F-Measure	0.5237	0.6149
kappa	0.0627	0.2631
time (seconds)	3.3696	423.7767

Table 6.2: Performance of classifiers on the validation collection as a two-class problem (education for high school students and others versus non-educational) based on text extracted from resources and all additional features.

	Normalised word frequency 2000 terms	
Measure	OneR	Naïve Bayes
AUC	0.5045	0.6808 ($p < 0.01$)
Accuracy	0.6178.0.5934	
Recall	0.5045.0.6158	
Precision	0.5072.0.6018	
F-Measure	0.4902.0.5837	
kappa	0.0104.0.1998	
time (seconds)	24.3733	37.7216
	Binary occurrence 4000 terms	
	OneR	AdaBoost
AUC	0.4992	0.5397 ($p = 0.04$)
Accuracy	0.6210.0.6578	
Recall	0.4992.0.4956	
Precision	0.4982.0.4692	
F-Measure	0.4787.0.4247	
kappa	-0.0019.-0.0113	
time (seconds)	41.8584	502.3653
	Normalised word frequency 500 terms	
	OneR	Random Forest
AUC	0.5109	0.6053 ($p = 0.02$)
Accuracy	0.6159.0.6668	
Recall	0.5109.0.4945	
Precision	0.5149.0.3941	
F-Measure	0.5015.0.4036	
kappa	0.0242.-0.0147	
time (seconds)	11.4332	933.9641

6.2.1 Failure analysis

To gain insight into how classification might be improved, it is useful to examine resources that were incorrectly classified. Most classification errors are accounted for by resources that were were never correctly classified in the 10 runs of 10-fold cross-validation, with Naïve Bayes always misclassifying 105 resources, AdaBoost always misclassifying 72 resources, and Random Forest always misclassifying 96 resources. The resources that were misclassified by a classifier on every run represent 66% of the classification errors.

There were 25 resources that were always misclassified by all classifiers, and each of those was a resource that had been judged unlikely to support learning that was classified as educational. Judgment data for these resources are shown in Table 6.3, including comments made by the judges about their judgment of those particular resources.

There does not appear to be a query effect in the misclassification. Resources retrieved in response to 19 queries are represented in the always misclassified resources, and no query is represented more than twice. Similarly, the judge does not appear to influence the misclassification. Of the 21 judges who participated in the task, 14 judged a resource that was always misclassified, mostly with one or two judgments, although one judge has three judgments and another has five.

With an average of 956.76 unique terms, these misclassified resources were substantially longer than other resources that the judges considered unlikely to support learning, which averaged 466.80 unique terms. However, they were also longer than resources that were judged likely to support learning, which averaged 748.05 unique terms, thus it is not sufficient to simply take into account resource length.

Of the 25 resources, judges had made comments after judging on 16, and six of those comments reveal that the judge considered the resource to be poorly or confusingly formatted. This suggests penalising poorly formatted resources may lead to an improvement in classification, however, detection of this would not be trivial.

Judgment error is a phenomenon observed in IR system evaluation, where resources are judged incorrectly, either through misunderstandings of the task, fatigue, boredom, bias, or various other reasons [Carterette and Soboroff, 2010]. This suggests the possibility of judgment error for some of these resources, and inspection suggests that this may have been the case for a subset of the always misclassified resources. However, judgments in

general include a subjective component, so it is difficult to objectively identify "errors" in the judgment data.

Table 6.3: Judgment data for the 25 resources that were misclassified by all classifiers in all 10 runs of 10-fold cross-validation.

Resource	Query	User	Comment
361	37	2	Most facts or concepts were hidden in large amounts of text.
40	4	2	The odd way of formatting the page (changing the style every few paragraphs) would just distract the student and facts wouldn't sink in.
86	9	4	If anything, this would most likely distrupt the grammer of a student in high school. In the later stages of high school, a student should be learning words that they can use, rather than words that have no meaning when used in an essay. Just think what would happen when half of the students who did "VCE English" started to use slang in their persuasive essay. "This article was lol to a user. though some people would roflmao instead ttfn, ta ta for now" Kind of stupid
307	31	4	It looks like a bunch of repeated text. No sane person would read beyond the third scroll
117	12	5	Looked at on its own the resource is not much more than a link to other resources that would likely be helpful. At the beginning it seems like it will be a learning aide in its own right, but that was only the "The Basics" at the top
262	27	6	this page is poorly laid out. lots of irrelevant info
61	7	10	
136	14	10	
205	21	10	
208	21	14	
177	18	14	Did not provide information on the topic so much as information on the way the topic was taught.
330	33	14	Site was selling a book, and high school students are quite unlikely to wait potentially weeks for the book to ship, on the assumption it was even relevant.
29	3	14	Content was mainly composed of corporate buzzwords and irrelevant information. As a highschool student I would have found this almost incomprehensible, not to mention frustrating.
21	3	14	While the blurb of the book is tantalising, it doesn't provide much useful information.
167	17	15	This didn't appear to have any useful information. The organization was unclear, the text was redundant and the columns were too thin to read easily

Continued on next page

Table 6.3 – continued from previous page

Resource	Query	User	Comment
159	16	16	This was a blog containing an opinion backed with some information cited from books, rather than a researched discussion.
106	11	16	
200	20	17	
289	29	18	this page itself doesn't teach me anything. It's just talking about what it aims to teach
400	40	18	
193	20	20	The test said many thing but nothing of real factual interest.
348	35	21	
134	14	21	Layout is really bad.....
174	18	22	
63	7	23	This was a hard resource to judge, it didn't read well and essentially was just a massive slab of text. It provided statistics which could aid students if they needed information on this very specific topic.

Figure 6.6 shows the average number of times resources were misclassified, out of the 10 runs of 10-fold cross-validation, as a function of the educational depth the judge assigned to them. The zero point on this scale represents resources that were judged to be unlikely to support learning. It is clear that the Random Forest and AdaBoost classifiers perform better on resources that were judged likely to support learning, though none of the classifiers exhibits a clear misclassification pattern in relation to the depth of resources judged likely to support learning.

6.3 Summary

In this chapter we investigated the effectiveness of the techniques for classifying educational resources developed in Chapter 5 when applied to a larger, independent validation collection. This validation collection was developed by asking judges to examine resources and classify them as being likely to support a high school student to acquire knowledge or a skill, likely to support learning in another context, or unlikely to support learning. The 410 Resources judged were retrieved from the Web in response to queries extracted from a Victorian high school curriculum document.

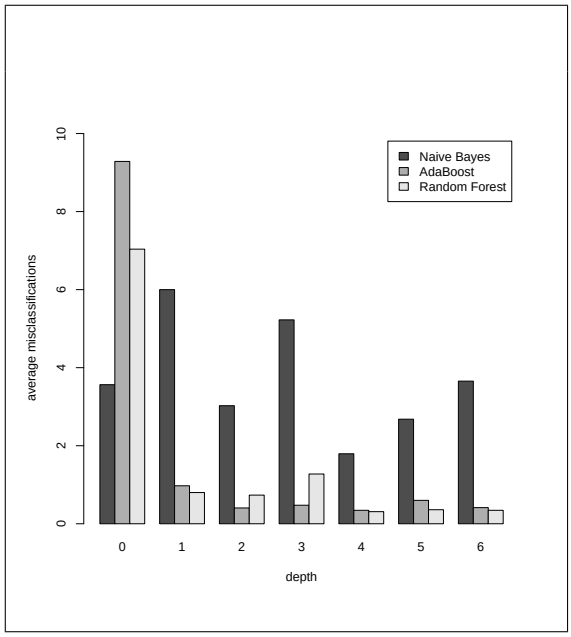


Figure 6.6: Average misclassification errors by the educational depth of the resource.

The judgments did not show that any judge was biased towards a particular judgment value, and judges did not appear more confident or find the judging task easier when they felt the resource was likely to support learning compared with unlikely to support learning.

Overall, judges said they enjoyed the task, and that they would find a system that returned educational resources useful. Although judges stated that the instructions were clear and they consistently used the same criteria to assess resources, a third of the judges said the process was confusing. Comments made at the end of the judging process suggest there was some confusion as to whether judgments should be made based only on the resources themselves or on resources to which they refer or link.

As expected from our preliminary investigation in Chapter 5, the performance of tuned Naïve Bayes, AdaBoost, and Random Forest classifiers built using a term vector of text extracted from the resources was significantly better than a OneR baseline. The addition of attributes derived from hyperlinks and headings did not improve classification performance.

Given the superior performance of the AdaBoost classifier observed in the previous chapter, we also hypothesised that AdaBoost would outperform the Naïve Bayes and Random Forest classifiers on the validation collection, however that was not the case. The Random

Forest classifier performed significantly better when built with the term vector of extracted text, while the Naïve Bayes classifier performed best when the additional features were added. It is possible that the tuning performed on the classifiers during development was dependent on the collection, and that a term vector of different length may have been more effective.

Comments from the judges indicate that they were taking into account aspects of page structure and hyperlinks when making judgments as to whether a resource was educational. This suggests that the attributes we extracted from these features, which harmed classification performance, did not capture the information that judges were using.

A possible confounding factor that we have not taken into account is the query used to retrieve resources. Our goal is to create a system that is capable of differentiating between resources that are educational and those that are not, independent of the query used to retrieve them. However, as we use multiple results for each query, when performing cross validation both the training and test sets will very likely contain resources retrieved with the same query. Future work could take into account this possible query effect by segregating resources by the query used to retrieve them, ensuring that the training and test sets do not both contain resources retrieved with the same query.

While the collection described in this chapter was larger than the collection used to develop the classifiers, it was still relatively small when compared to collections used in similar domains, such as the TREC collections. A better indication of the appropriate classifier to choose for future work might be provided by using a larger sample size. The development of a larger ground truth would be valuable future work.

In the following chapter we summarise the thesis, describing its contributions and describe other possible future work.

Chapter 7

Conclusions

Learning has a fundamental place in human experience. Given the proliferation of digital resources, finding appropriate resources to support that learning is not trivial. In this thesis we extend the understanding of the requirements of teachers and institutions in relation to the management and reuse of resources to support learning, propose a method for the evaluation of systems that filter resources returned by search engines for educational resources, and evaluate the performance of examples of such systems.

In Section 1.1, we presented the following research questions:

- How are educational resources currently used and managed?
- How should systems that filter educational resources be evaluated?
- What methods can be used to effectively filter educational resources?

We reiterate our main contributions and summarise our responses to those questions in Sections 7.1, 7.2, and 7.3. We then outline possible future research directions in Section 7.4, and present concluding remarks in Section 7.5.

7.1 Management and reuse of educational resources

In Chapter 3 we described our exploration of the management and reuse of educational resources, using focus groups, a survey, and two sets of interviews to investigate how educational resources are currently used and managed. This qualitative work allowed us to draw the following conclusions.

- When searching for resources to support learning, many educators would prefer to use a public search engine to search the Web rather than access an institutionally-managed system.
- There is disagreement about the quality of educational resources that institutions should be manage. Some people feel strongly that all resources submitted to institutional systems for the management of educational resources should receive quality review. Other people are in favour of allowing more informal sharing, for example of works in progress.
- Incentives should be provided to encourage both the contribution of resources and the reuse of resources contributed by others. These incentives should be comparable with those that encourage research activity, which is perceived to be more highly regarded than teaching.

Data collection was constrained to educators and experts in Australia, and while we have no reason to believe it to be the case, results may differ overseas. Surveys across institutions and internationally should be considered in future work in this area. Additionally, the use of a scale with a neutral mid-point may be beneficial.

7.2 Evaluation of educational resource filters

In Chapter 4 we examined the evaluation of systems that filter educational resources when retrieved from collections of heterogeneous resources, such as the Web. This contrasts with previous research into retrieval from homogeneous collections managed by institutions that ideally hold only educational resources. We presented a user experiment in which 8 participants judged each of 20 resources retrieved from the web by a search engine as likely or unlikely to support learning. In Chapter 6 we used the methodology outlined in Chapter 4 for the development of appropriate test collections to construct a larger, independent collection of web resources. In our second user experiment, 21 judges were asked to examine a total of 410 resources and assess whether they were likely to support the learning of a student in high school, likely to support learning in another context, or unlikely to support learning. We asked judges questions to investigate the judging process. Through this work, we provide the following contributions.

- We proposed an evaluation methodology for systems that filter educational resources. The proposed methodology is based upon the Cranfield method, which is widely used in the evaluation of information retrieval systems.

- By showing that there is a high level of agreement between judges when judging whether resources are likely to support learning, we established that it is sufficient to use a single assessment for each resource.

- We showed that judges more often decide a resource is unlikely to support learning when the query used to retrieve the resource is shown during the judgment of the resource. Based on this finding and on comments made during the judging process, we believe that this is because the query distracts judges, causing them to judge topical relevance rather than whether a resource is likely to support learning.

- Judges said that a search engine that returned educational resources would be personally useful to them.

The level of inter-rater agreement demonstrated in our user experiment is higher than that observed in the work used to justify the use of a single assessor in information retrieval evaluation. However, in that work Voorhees [1998] shows that system performance is stable, despite the variability of judgments. Demonstration of system stability should be examined in the future, when multiple systems for filtering educational resources have been implemented by other parties.

There was evidence to suggest that there was some confusion among judges involved in the construction of the larger validation collection as to whether the judging task required them to make judgments based only on the resources themselves or on resources to which they refer or link. In future work, this should be further clarified in the instructions given to participants.

7.3 Filtering educational resources

In Chapter 5 we explored methods for filtering web resources to differentiate those that are educational from those that are not educational. Using a development collection of 100 resources that had been judged by the author of this thesis, we evaluated a variety of machine

learning classification algorithms according to their ability to correctly classify resources based on attributes extracted from the text of resources, and further investigated whether attributes derived from hyperlinks and headings in resources might be useful in classification. In Chapter 6 we described the use of the independent labelled validation collection of 410 web resources to assess the performance of the classification models developed in Chapter 5. These models were constructed using text extracted from the resources and the Naïve Bayes, AdaBoost, and Random Forest classification algorithms. In examining methods for filtering educational resources, we provide the following contributions.

- Text extracted from web resources provides useful attributes for classifying resources as educational or not educational.

- We demonstrated that classification models produced using extracted text and the AdaBoost, Naïve Bayes, and Random Forest induction algorithms perform well for the filtering of educational resources. The classification algorithm used influences what method of text extraction is optimal. The best performance of AdaBoost was observed when using the binary absence or presence of terms in the text as attributes, and retaining the 4000 most frequent terms. Naïve Bayes and Random Forest perform best when the number of occurrences of terms, normalised by the length of the resources, are used as attributes, with the 2000 and 500 most frequent terms respectively.

- Attributes that we derived from hyperlink and heading features did not improve classification performance. These attributes were the number of internal links, the number of outgoing links, the ratio of link text to overall text, text extracted from links, the number of headings, and the ratio of heading text to overall text.

- The classification model constructed using the Random Forest classification algorithm significantly outperformed those built using Naïve Bayes and AdaBoost on our larger, independent validation collection.

- Though our earlier findings suggested that attributes derived from hyperlinks and headings do not improve classification effectiveness, comments made by judges indicate they were taking into account aspects of page structure and hyperlinks when making judgments as to whether a resource was educational. This suggests that the particular

attributes we extracted from these features did not capture the information judges were using, and that other aspects of these features may still be beneficial.

The larger validation collection used was relatively small when compared to collections used in similar domains, such as the TREC collections. Further work based on larger collections of resources labelled as educational or not educational should be undertaken to verify which particular classification algorithm is optimal.

7.4 Future work

In this section we describe possible future research that would extend the work presented in this thesis.

7.4.1 Multi-page resources

The resources retrieved from the Web and assessed by judges in the user experiments in this work were single HTML pages. Given the hyperlinked nature of the Web, it is likely that many pages are not designed to stand alone, but are part of an interlinked cluster of pages. Thus, a page that is judged unlikely to support learning on its own, may be judged likely to support learning if it was part of a larger resource. This is related to the concept of granularity as used in reference to reusable learning objects. The detection of what might be called *resource boundaries*, possibly using techniques such as similarity measurement [Lin, 1998] or clustering [Berkhin, 2006], would be a valuable extension to our work.

Considering multi-page resources exposes a further set of interesting problems. For example, which page in a multi-page resource should be considered an "entry point" and therefore presented to the user? How should ranked search results of such resources be presented? How should summaries of multiple pages be constructed?

7.4.2 Improving text extraction

In extracting text from HTML resources, we simply discarded `script` and `style` elements and removed HTML tags. Simply stripping HTML tags means that all other content is given the same weight, whether it occurs in the main content of the page, as part of the site-wide navigation, or in a footer.

Observation shows that many web pages contain substantial amounts of text that is unrelated to the content of the page, such as navigation and advertising. Finn et al. [2001] used this observation to develop an algorithm that extracts text from the body of a resource and ignores the extraneous content. They argue that the most important text in a resource tends to occur within an area of low tag density. HTML pages have a greater tag density at the beginning and end of the document. As we read the raw HTML from the start of the document, we see a steady rise in the number of HTML tags, followed by a plateau, and ending with a further rise. Text extracted only from the plateau may be more representative of the purpose of the resource, and therefore allow for improved classification performance.

Structural or semantic cues in the resources may also be useful. For example, more weight could be placed on terms that appear in headings, or in the opening or concluding sentences or paragraphs.

7.4.3 Readability measures

Comments made on the judgments during the construction of our validation collection indicate that, in addition to other clues, judges used the readability of resources to assess whether a resource was likely to support learning. This suggests that people are less likely to judge a resource as educational if it is poorly written.

Measures based on features of grammar or vocabulary, such as the Flesch-Kincaid readability test[1] and the Gunning Fog Index,[2] have been developed to quantify the readability of text in traditional formats such as books and articles, and used in fields such as language acquisition [Uitdenbogerd, 2006]. Statistical language models have also been found to be effective for assessing the readability of web content [Collins-Thompson and Callan, 2005]. These readability measures could be used to give more weight to more readable resources. In conjunction with user profiles, readability measures could also be used to ensure that resources are of an appropriate reading level for users.

[1] http://en.wikipedia.org/wiki/Flesch-Kincaid_readability_test
[2] http://en.wikipedia.org/wiki/Gunning_fog_index

7.4.4 Non-HTML resources

In this work, we considered only HTML resources, but there are many more types of indexable resources on the Web. There are two other kinds of resources, those that are published as standalone resources, such as PDFs or binary word processing documents, and those that can be part of composite resources and embedded in HTML resources, such as images, video or audio, and they should be considered in different ways. We have shown that text extracted from HTML resources is a useful feature for classifying items as educational or not educational, and this can be extended to standalone resources, from which text can be extracted. In the work presented in this thesis, we ignored embedded resources, however, text or other features could also be extracted from these to be used as inputs to classification algorithms. Future work could investigate whether the performance of the techniques described in this thesis are effective for embedded and standalone non-HTML resources.

7.4.5 Facet retrieval

When searching for educational resources in a homogeneous collection, a searcher is unlikely to be satisfied if they are presented with a resource where only a deeply buried part is relevant to their information need. Indeed, if the searcher is scanning a list of ranked results and making a judgment as to whether to examine a learning object more closely, they are unlikely to invest the time to search the resource for the relevant part if a summary or title does not look immediately relevant. It is even less likely that a user will delve deeply into a resource whose front page does not appear relevant, even if they do decide to review further.

This issue is reminiscent of work done in the field of structured information retrieval. Where a relevant structured document is divided into sections, it is not clear whether it is best to return the entire document or attempt to return a suitably sized subsection of the document.

7.4.6 User evaluation

In this thesis, we conducted qualitative research to investigate how educational resources are managed, including how people prefer to find resources. We then showed that a high level of agreement is possible when judging whether retrieved resources are educational, independent of the context in which the resource will be used. There is potential for future research to

investigate the how satisfied users are when searching for educational resources using search results that have been filtered using the techniques described in this thesis.

7.5 Final remarks

This thesis has contributed to helping people find appropriate digital educational resources to support our fundamental human need to learn, a pursuit in which we spend much of our time, both informally in our day-to-day lives and in the formal context of education. Possible practical applications of the work described in this thesis include the provision of a service similar to that offered by Google Scholar[3] that returns educational resources instead of academic papers or for educational institutions to offer filtered search of the Web for their students and staff.

Our work provides a basis for future research in a new area of research within information retrieval that investigates the retrieval of educational resources from the Web.

[3] http://scholar.google.com.au/

Appendix A

Characteristics of homogeneous educational resource collections

To understand the nature of existing LORs, we looked at two collections of learning objects. Rather than attempting to define a learning object as an abstract concept, we examined the types of resources being stored for reuse.

Collection 1 consists of 26 536 files used as part of several business related subjects. These files have not been developed to comply with any learning object standards. While descriptive metadata has not been added to the resources, there is a certain amount of metadata associated with them, such as the name, size and modification time of the files. The files are currently being reused in the same context over time, though it is not clear if they are being used in different contexts.

Collection 2 consists of 4.3GB of compressed material comprising almost 1 000 SCORM-compliant composite learning objects. This collection is managed by an organisation that oversees the development and maintenance of significant numbers of learning objects targeted at vocational education and training (VET) across Australia. In total there are 464 307 unique files.

The breakdown of file types for each collection is shown in Table A.1. Files have been grouped according to their file extensions, and similar file types have been merged. In both collections image and HTML files account for a large proportion of the total files, and there are significant numbers of Flash files, support files (XML Schema documents, JavaScript,

Table A.1: Breakdown of collections by file type.

	Collection 1		Collection 2	
File type	Number	Percentage	Number	Percentage
image	254 448	54.8%	8 777	33.1%
HTML	124 261	26.8%	6 072	22.9%
Flash	37 503	8.1%	3 389	12.8%
support	18 892	4.1%	3 225	12.2%
document	12 799	2.8%	3 055	11.5%
other	16 404	3.5%	2 018	7.6%

Cascading Style Sheets and Macromedia notes) and documents (Microsoft Word, RichText, PDF and PostScript).

Examination of the links in Collection 1 reveals that many files have been combined for presentation. As there are many HTML files, which generally include hyperlinks, this is not surprising. There are 29 872 link anchors and 17 961 links to images in the HTML. Of the anchors, 5 456 link to the target `top` within the document. Of the links to images, 3 216 link to files with "icon" in the file name and 234 of them are called `rmitlog.gif`. While these are being reused, it seems likely that these images add only presentational value rather than educational value.

Figure A.1 shows the percentages of different file types in Collection 1 by course. The courses are sorted by the average modification time of the files within the course. The course with the average oldest files appears on the leftmost side of the graph.

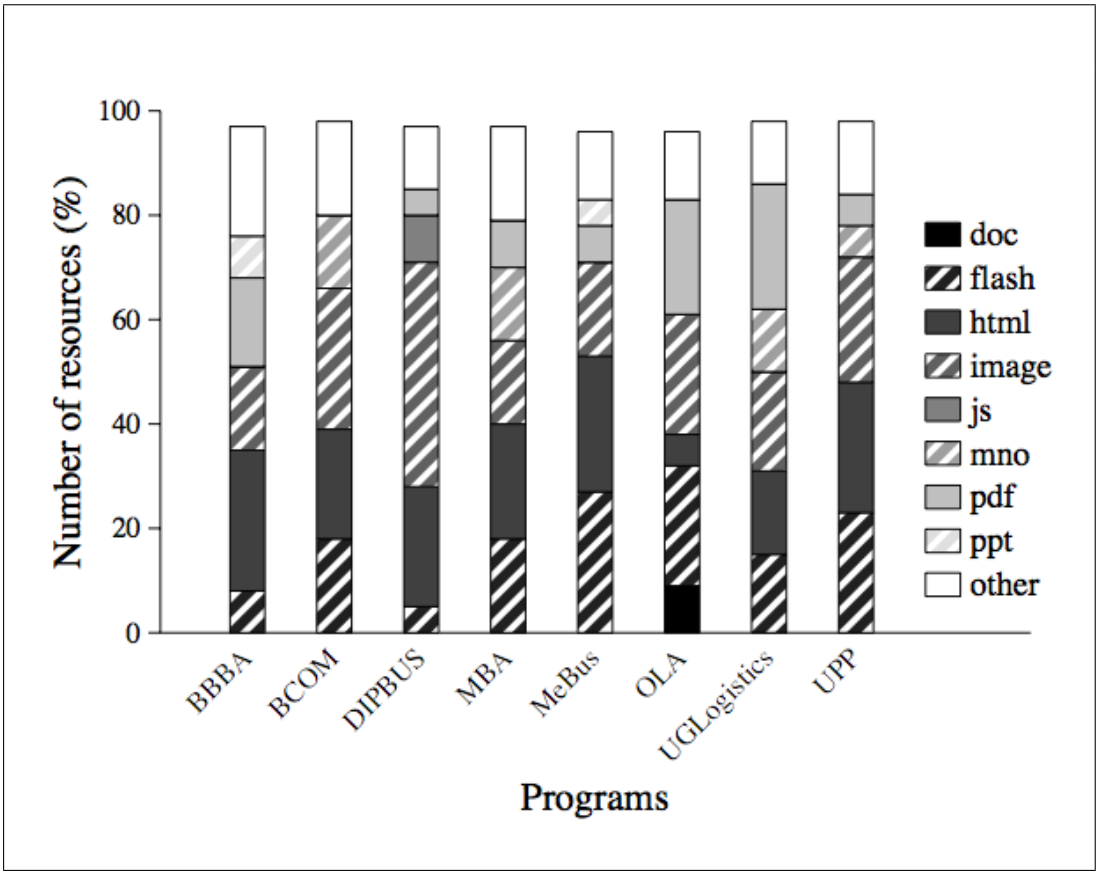


Figure A.1: File types by course

Appendix B

Focus groups

Section 3.2 describes focus groups exploring the management and reuse of educational resources. This appendix includes supplementary resources about the focus groups.

B.1 Emails to participate in focus groups

Focus group participants were emailed guiding questions before attending their sessions. This email differed depending on the group that the participant had been invited to.

B.1.1 "Aware" focus groups

> Thank you for agreeing to participate in a focus group regarding reusable learning objects (RLOs) in RMIT at [date and location].
>
> The focus group you are attending will be discussing issues around the development, use, reuse and storage of learning objects.
>
> The questions we will be asking you to address during the discussion will include the following;
>
> - Who do you know who is undertaking interesting work with learning objects, both within RMIT and outside, and what is the nature of that work?
> - How can the activities within RMIT be supported?

- What would stifle the activities of those within RMIT now and in the future?
- What would be the advantages and disadvantages of systems that centralised versus decentralised the storage of learning objects and the control over them?
- What would be the most productive way that RMIT could progress in the field of learning objects?
- What do you think a learning object is?
- What requirements would you have of a system to manage your use of learning objects?

You are also welcome to raise other issues that you think are relevant.

Please find attached a Plain Language Statement regarding the work of this project as required by the RMIT ethics committee. A signed hard copy of this form will be given to you at the focus group.

We will also be asking you to sign a consent form as required by the RMIT ethics committee at the focus group.

B.1.2 "Naïve" focus groups

Thank you for agreeing to participate in a focus group regarding reusable learning objects (RLOs) in RMIT at [date and location].

The focus group you are attending will be discussing issues around the development, use, reuse and storage of learning objects.

The questions we will be asking you to address during the discussion will include the following;

- Have you used, or are you using, material produced by someone else in your teaching or development?
- Are you aware of other people/groups who teach/develop in similar discipline areas to yourself?

- Are there topics/concepts/processes that you teach/develop which are in common with other teaching groups?
- Would you consider directly using material produced by those other groups in your own teaching/development? Why or why not?
- Would you consider revising material produced by those other groups for use in your teaching/development? Why or why not?
- Would you consider repurposing material produced by those other groups for use in your teaching/development? Why or why not?
- How would you react to the other people/groups using/revising/repurposing material that you had created?
- What do you think a Learning Object is/includes?
- What support would you need to help you deal with LOs?
- What requirements would you have of an RLO system?

You are also welcome to raise other issues that you think are relevant.

Please find attached a Plain Language Statement regarding the work of this project as required by the RMIT ethics committee. A signed hard copy of this form will be given to you at the focus group.

We will also be asking you to sign a consent form as required by the RMIT ethics committee at the focus group.

B.2 Focus group plain language statement

Participants in the focus groups were presented with a plain language statement before being involved in our research. The plain language statement describes the research, including the context and goals of the work, and a description of the process. The plain language statement is shown on the following pages.

RMIT UNIVERSITY

Educational Media Group
GPO Box 2476V
Melbourne 3001
Manager:
Amgad Louka – 9925 9621
amgad.louka@rmit.edu.au
Reusable Learning Objects Project – Stage 1
Project Officers:
Henric Beiers – 9925 3213
henric.beiers@rmit.edu.au
Michael Harris – 9925 9676
miharris@cs.rmit.edu.au

Reusable Learning Objects Project - Stage 1
Plain Language Statement – Focus Groups

10th June 2004

To Whom It May Concern,

You are invited to participate in a project investigating the use, reuse and storage of learning objects at RMIT.

Background
The Reusable Learning Objects Project is being run under the auspices of the Educational Technology Subcommittee of the RMIT Teaching & Learning Strategy Committee. A reference group chaired by Amgad Louka, Manager - Educational Media Group, manages it.

The project is intended to support and effectuate some objectives within the RMIT T&L Strategy 2004 – 2006. In particular:

- 'Improve the quality of teaching – Identify the effective use of technology to support student learning'
- 'Ensure the viability of our educational profile – develop sustainable curriculum and teaching models.'
- 'Enhance organizational capability – Ensuring a robust and reliable infrastructure to support eLearning'

The RLO Project will involve many stages that aim to culminate in the implementation, ongoing maintenance, effective management and adoption of RLOs as part of curriculum design and development, research and other business activities at RMIT.

The Educational Media Group is funding stage 1 of the project. Stage 1 will:

- Investigate current universal issues and practices in the development of a RLO approach in education;
- Interview and survey staff involved with the production and use of learning objects to capture their understanding of, and practices at RMIT;
- Identify RMIT initiatives and projects where the storage of RLOs and interoperability is a consideration (eg. DLS, Integrated Library Management System and eReserve, content management in Schools).
- Investigate standards that would enable interoperability with open archives and other institutional repositories.
- Identify barriers to the use of RLOs at RMIT

Any complaints about your participation in this project may be directed to:
The Secretary, RMIT Human Research Ethics Committee, University Secretariat,
RMIT, GPO Box 2476V, Melbourne, 3001.
The telephone number is (03) 9925 1745.
Details of the complaints procedure are available also from the above address

Figure B.1: Focus group plain language statement

- Document and report investigations, and
- Provide criteria/business requirements to inform the development and implementation of a RLO approach in teaching and learning at RMIT.

Process
As part of the project we are running a series of focus groups in order to gather the views of staff at RMIT regarding the use of learning objects.

A member of the project reference group, or one of the project team, has identified you as someone who may be interested in participating in these focus groups and contributing to the discussion. Participation in the interviews is voluntary and participants can withdraw from the process at any time they wish.

The focus groups will run for 1½ hours. There will be between 5 and 10 people in each of the focus groups. They will be tape-recorded and then transcribed to allow the project officers to undertake analysis of the data collected.

Tapes and transcripts will be stored in the Learning Objects Project offices of the Educational Media Group in a locked cupboard for a period of five years after which time they will be destroyed. Access will be restricted to staff of the Educational Media Group and staff for the time being working on the Learning Object Project

The data collected from the focus groups will contribute to three possible outputs.
i) The issues identified will form the basis of an online questionnaire for other RMIT staff to respond to.
ii) The issues will be also be reported directly in a report to be produced for submission to the project Reference Group, and then to the Ed. Tech. Subcommittee, by 3rd September 2004.
iii) A journal paper may be published based upon the work of the project.

When the focus groups are reported on, in either the final report or a published paper, there will be no identification of the individual members of the groups. We are more interested in the issues raised and discussion points.

If you have any questions regarding this project please contact any of the project staff. Our contact details are in the header to this letter. You may also contact the Associate Dean Research, Professor Robert Brooks, RMIT Business Portfolio, post: GPO Box 2476V, Melbourne 3001; email: robert.brooks@rmit.edu.au; phone: 9925 5593.

Amgad Louka	Henric Beiers	Michael Harris

Any complaints about your participation in this project may be directed to:
The Secretary, RMIT Human Research Ethics Committee, University Secretariat,
RMIT, GPO Box 2476V, Melbourne, 3001.
The telephone number is (03) 9925 1745.
Details of the complaints procedure are available also from the above address

Figure B.1: Focus group plain language statement (continued)

Appendix C

Expert interviews

Section 3.3 describes interviews with domain experts exploring the management and reuse of educational resources. This appendix includes supplementary resources about the expert interviews.

C.1 Expert interview plain language statement

Participants in the expert interviews were presented with a plain language statement before being involved in our research. The plain language statement describes the research, including the context and goals of the work, and a description of the process. The plain language statement is shown on the following pages.

RMIT UNIVERSITY

Educational Media Group
GPO Box 2476V
Melbourne 3001
Manager:
Amgad Louka – 9925 9621
amgad.louka@rmit.edu.au
Reusable Learning Objects Project – Stage 1
Project Officers:
Henric Beiers – 9925 3213
henric.beiers@rmit.edu.au
Michael Harris – 9925 9676
miharris@cs.rmit.edu.au

Reusable Learning Objects Project - Stage 1
Plain Language Statement – Expert Interviews

10th June 2004

To Whom It May Concern,

You are invited to participate in a project investigating the use, reuse and storage of learning objects at RMIT.

Background
The Reusable Learning Objects Project is being run under the auspices of the Educational Technology Subcommittee of the RMIT Teaching & Learning Strategy Committee. A reference group chaired by Amgad Louka, Manager - Educational Media Group, manages it.

The project is intended to support and effectuate some objectives within the RMIT T&L Strategy 2004 – 2006. In particular:

- 'Improve the quality of teaching – Identify the effective use of technology to support student learning'
- 'Ensure the viability of our educational profile – develop sustainable curriculum and teaching models.'
- 'Enhance organizational capability – Ensuring a robust and reliable infrastructure to support eLearning'

The RLO Project will involve many stages that aim to culminate in the implementation, ongoing maintenance, effective management and adoption of RLOs as part of curriculum design and development, research and other business activities at RMIT.

The Educational Media Group is funding stage 1 of the project. Stage 1 will:

- Investigate current universal issues and practices in the development of a RLO approach in education;
- Interview and survey staff involved with the production and use of learning objects to capture their understanding of, and practices at RMIT;
- Identify RMIT initiatives and projects where the storage of RLOs and interoperability is a consideration (eg. DLS, Integrated Library Management System and eReserve, content management in Schools).
- Investigate standards that would enable interoperability with open archives and other institutional repositories.
- Identify barriers to the use of RLOs at RMIT

Any complaints about your participation in this project may be directed to:
The Secretary, RMIT Human Research Ethics Committee, University Secretariat,
RMIT, GPO Box 2476V, Melbourne, 3001.
The telephone number is (03) 9925 1745.
Details of the complaints procedure are available also from the above address

Figure C.1: Expert interview plain language statement

- Document and report investigations, and
- Provide criteria/business requirements to inform the development and implementation of a RLO approach in teaching and learning at RMIT.

Process
As part of the project we are interviewing selected people in order to gather their views regarding the use of learning objects. In particular we are interested in interviewing people who have a significant interest or expertise in the area.

A member of the project reference group, or one of the project team, has identified you as someone who has such an interest or expertise and may also be interested in participating in an interview. Participation in the interviews is voluntary and participants can withdraw from the process at any time they wish.

We anticipate that interviews will run for no more than 1 hour. They will be tape-recorded and then transcribed to allow the project officers to undertake analysis of the data collected.

Tapes and transcripts will be stored in the Learning Objects Project offices of the Educational Media Group in a locked cupboard for a period of five years after which time they will be destroyed. Access will be restricted to staff of the Educational Media Group and staff for the time being working on the Learning Object Project

The data collected from these interviews will contribute to two possible outputs.
i) They will be reported upon directly in a report to be produced for submission to the project Reference Group, and then to the Ed. Tech. Subcommittee, by 3rd September 2004.
ii) A journal paper may be published based upon the work of the project.

When the interviews are reported on, in either the final report or a published paper, it is possible that there will be identification of individual interviewees. However, no such identification will be made without the express permission of the interviewee. All interviewees whose views are identified will be given a chance to confirm or correct the substance of their reported view before publication.

If you have any questions regarding this project please contact any of the project staff. Our contact details are in the header to this letter. You may also contact the Associate Dean Research, Professor Robert Brooks, RMIT Business Portfolio, post: GPO Box 2476V, Melbourne 3001; email: robert.brooks@rmit.edu.au; phone: 9925 5593.

Amgad Louka	Henric Beiers	Michael Harris

Any complaints about your participation in this project may be directed to:
The Secretary, RMIT Human Research Ethics Committee, University Secretariat,
RMIT, GPO Box 2476V, Melbourne, 3001.
The telephone number is (03) 9925 1745.
Details of the complaints procedure are available also from the above address

Figure C.1: Expert interview plain language statement (continued)

Appendix D

Survey

Section 3.4 describes a survey of RMIT staff to test the extent to which issues raised in the focus groups and interviews were important to a wider population of staff. This appendix includes supplementary resources about the survey, and presents the survey questions and detailed results.

D.1 Invitation to participate in survey

> The Teaching and Learning Portfolio has a project to investigate the management of digital resources used in teaching and learning, including their reuse and sharing across the university. We need feedback about how staff feel this reuse and sharing would affect them, and written a short survey to help us discover your thoughts. Please complete the survey at http://www.rmit.edu.au/library/lor/survey.

D.2 Survey plain language statement

Respondents to the survey were presented with a plain language statement before being involved in our research. The plain language statement describes the research, including the context and goals of the work, and a description of the process. The plain language statement is shown on the following pages.

RMIT UNIVERSITY

Educational Media Group
GPO Box 2476V
Melbourne 3001
Manager:
Amgad Louka – 9925 9621
amgad.louka@rmit.edu.au
Reusable Learning Objects Project – Stage 1
Project Officers:
Henric Beiers – 9925 3213
henric.beiers@rmit.edu.au
Michael Harris – 9925 9676
miharris@cs.rmit.edu.au

Reusable Learning Objects Project - Stage 1
Plain Language Statement – Survey

10th June 2004

To Whom It May Concern,

You are invited to participate in a project investigating the use, reuse and storage of learning objects at RMIT.

Background
The Reusable Learning Objects Project is being run under the auspices of the Educational Technology Subcommittee of the RMIT Teaching & Learning Strategy Committee. A reference group chaired by Amgad Louka, Manager - Educational Media Group, manages it.

The project is intended to support and effectuate some objectives within the RMIT T&L Strategy 2004 – 2006. In particular:

- 'Improve the quality of teaching – Identify the effective use of technology to support student learning'
- 'Ensure the viability of our educational profile – develop sustainable curriculum and teaching models.'
- 'Enhance organizational capability – Ensuring a robust and reliable infrastructure to support eLearning'

The RLO Project will involve many stages that aim to culminate in the implementation, ongoing maintenance, effective management and adoption of RLOs as part of curriculum design and development, research and other business activities at RMIT.

The Educational Media Group is funding stage 1 of the project. Stage 1 will:

- Investigate current universal issues and practices in the development of a RLO approach in education;
- Interview and survey staff involved with the production and use of learning objects to capture their understanding of, and practices at RMIT;
- Identify RMIT initiatives and projects where the storage of RLOs and interoperability is a consideration (eg. DLS, Integrated Library Management System and eReserve, content management in Schools).
- Investigate standards that would enable interoperability with open archives and other institutional repositories.
- Identify barriers to the use of RLOs at RMIT

Any complaints about your participation in this project may be directed to:
The Secretary, RMIT Human Research Ethics Committee, University Secretariat,
RMIT, GPO Box 2476V, Melbourne, 3001.
The telephone number is (03) 9925 1745.
Details of the complaints procedure are available also from the above address

Figure D.1: Survey plain language statement

- Document and report investigations, and
- Provide criteria/business requirements to inform the development and implementation of a RLO approach in teaching and learning at RMIT.

Process
As part of the project we are administering a survey in order to gather the views of staff at RMIT regarding the use of learning objects.

Participation in the survey is voluntary and anonymous. No identifying information is gathered as part of the survey process. Participants can withdraw from the process at any time they wish.

The data collected from this survey will contribute to two possible outputs.
i) The issues will be also be reported directly in a report to be produced for submission to the project Reference Group, and then to the Ed. Tech. Subcommittee, by 3rd September 2004.
ii) A journal paper may be published based upon the work of the project.

If you have any questions regarding this project, or would like to offer further individual input to the project, please contact any of the project staff. Our contact details are in the header to this letter. You may also contact the Associate Dean Research, Professor Robert Brooks, RMIT Business Portfolio, post: GPO Box 2476V, Melbourne 3001; email: robert.brooks@rmit.edu.au; phone: 9925 5593.

Amgad Louka	Henric Beiers	Michael Harris

Any complaints about your participation in this project may be directed to:
The Secretary, RMIT Human Research Ethics Committee, University Secretariat,
RMIT, GPO Box 2476V, Melbourne, 3001.
The telephone number is (03) 9925 1745.
Details of the complaints procedure are available also from the above address

Figure D.1: Survey plain language statement (continued)

D.3 Survey questions

A. Background Information

1. In what school, department or group are you employed?
2. How long have you been teaching in the post-secondary sector?
3. In which areas do you teach? (Select any that apply)
   - Higher Ed
   - VET
   - RMIT Training
   - Library
   - ITS
   - Other, please specify
4. In which modes do you teach? (Select any that apply)
   - Face-to-face
   - Online
   - Mixed mode
   - Distance
   - Off shore
   - Condensed
   - Other, please specify
5. Which of the following do you use in your teaching? (Select any that apply)
   - Blackboard
   - WebLearn
   - Other DLS tools
   - Library E Reserve
   - Other digital Library tools
   - Non DLS web sites
   - ANTA Toolboxes

- AEShareNet
- Other digital resources, please specify

B. When using resources created by other RMIT staff:
How important would it be for you to:
(Unimportant, Somewhat Important, Important, Very Important, Vital)

1. Know who created the resource?
2. Know who else had used the resource?
3. Know how others had used the resource?
4. Know if the resource had been changed?

Any other comments regarding questions 1–4:

How important would it be for you to:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

5. Be free to use the resource as is without restriction?
6. Be free to change the resource without restriction?

Any other comments regarding questions 5–6:

How important would it be for you to:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

7. Know that the resource had undergone a quality review?
8. Have access to a quality review regarding the resource?
9. Have the opportunity to participate in a quality review of the resource?
10. Have the opportunity to annotate the resource for the benefit of others?

Any other comments regarding questions 7–10:

Any other comments regarding the conditions under which you would, or would not, use resources created by other RMIT staff?

C. When contributing resources for use by other RMIT staff
How important would it be for you to:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

1. Be acknowledged as the creator of the resource in the storage system?
2. Be acknowledged as the creator of the resource in its subsequent use?
3. Be acknowledged as the creator of the resource if it was subsequently changed?

Any other comments regarding questions 1–3:

How important would it be for you to:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

4. Know who uses the resource?
5. Set general conditions on who can use the resource?
6. Control on a case-by-case basis who can use the resource?

Any other comments regarding questions 4–6:

How important would it be for you to:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

7. Know how the resource is used?
8. Set general conditions on how the resource can be used?
9. Control on a case-by-case basis how the resource can be used?

Any other comments regarding questions 7–9:

How important would it be for you to:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

10. Know of changes made to the resource?
11. Set general conditions on how the resource can be changed?
12. Control on a case-by-case basis how the resource can be changed?

Any other comments regarding questions 10–12:

How important would it be for you to:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

13. Be personally financially recompensed for the use of the resource?

14. Be personally rewarded through your workplan, promotion, awards or other mechanism for the use of the resource?
15. Have your group/school/portfolio financially recompensed for the use of the resource?

Any other comments regarding questions 13–15:

How important would it be for you to:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

16. Be free to share a resource without it undergoing a quality review?
17. Know if there was a quality review of the resource?
18. Have input into a quality review of the resource?

Any other comments regarding questions 16–18:

Any other comments regarding the conditions under which you would, or would not, contribute resources for use by other RMIT staff?

D. Using a computerised system for the reuse of resources
How important would it be for you that the system:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

1. Is able to search on key words?
2. Is able to refine a previous search?
3. Has a "find more like this" function?
4. Has a fast and efficient search capability?
5. Has an Advanced Search feature?

Any other comments regarding questions 1–5:

How important would it be for you that the system:
(Unimportant, Somewhat Important, Important,Very Important, Vital)

6. Has a simple user interface?
7. Has a consistent look and feel?

Any other comments regarding questions 6–7:

Any other comments regarding requirements of a system that facilitated the reuse of resources?

When considering the answers to all of these questions, what kind of resources were you thinking about? Please give an example.

D.4 Survey results details

Presented in the following pages are the results of the survey. Each graph shows the frequency of responses on the scale *Unimportant*, *Somewhat Important*, *Important*, *Very Important*, and *Vital* for a single question. Responses are divided by the area in which the respondent worked: higher education; vocational education and training; both higher education and vocational education and training; and other areas.

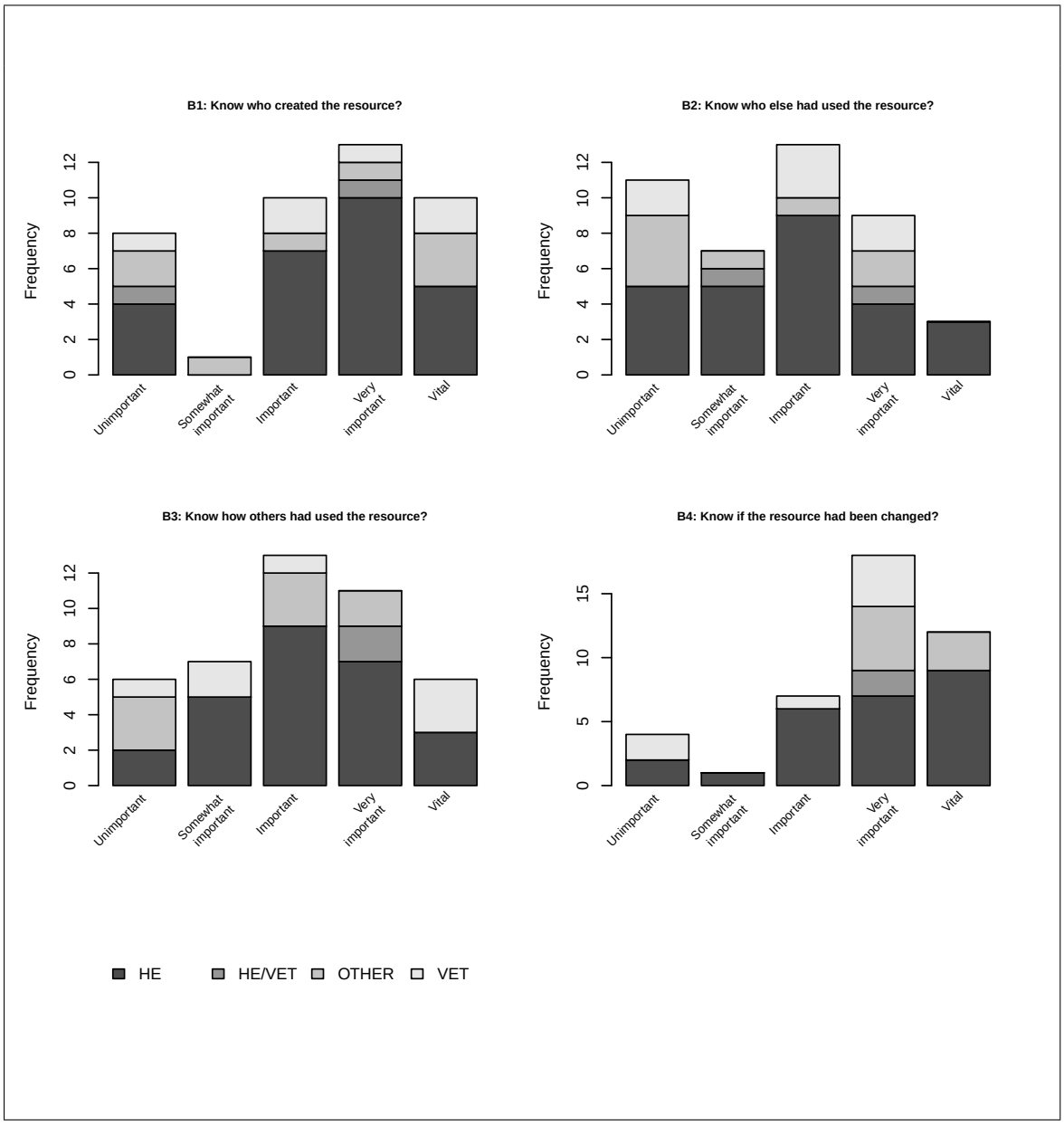


Figure D.2: Survey result details

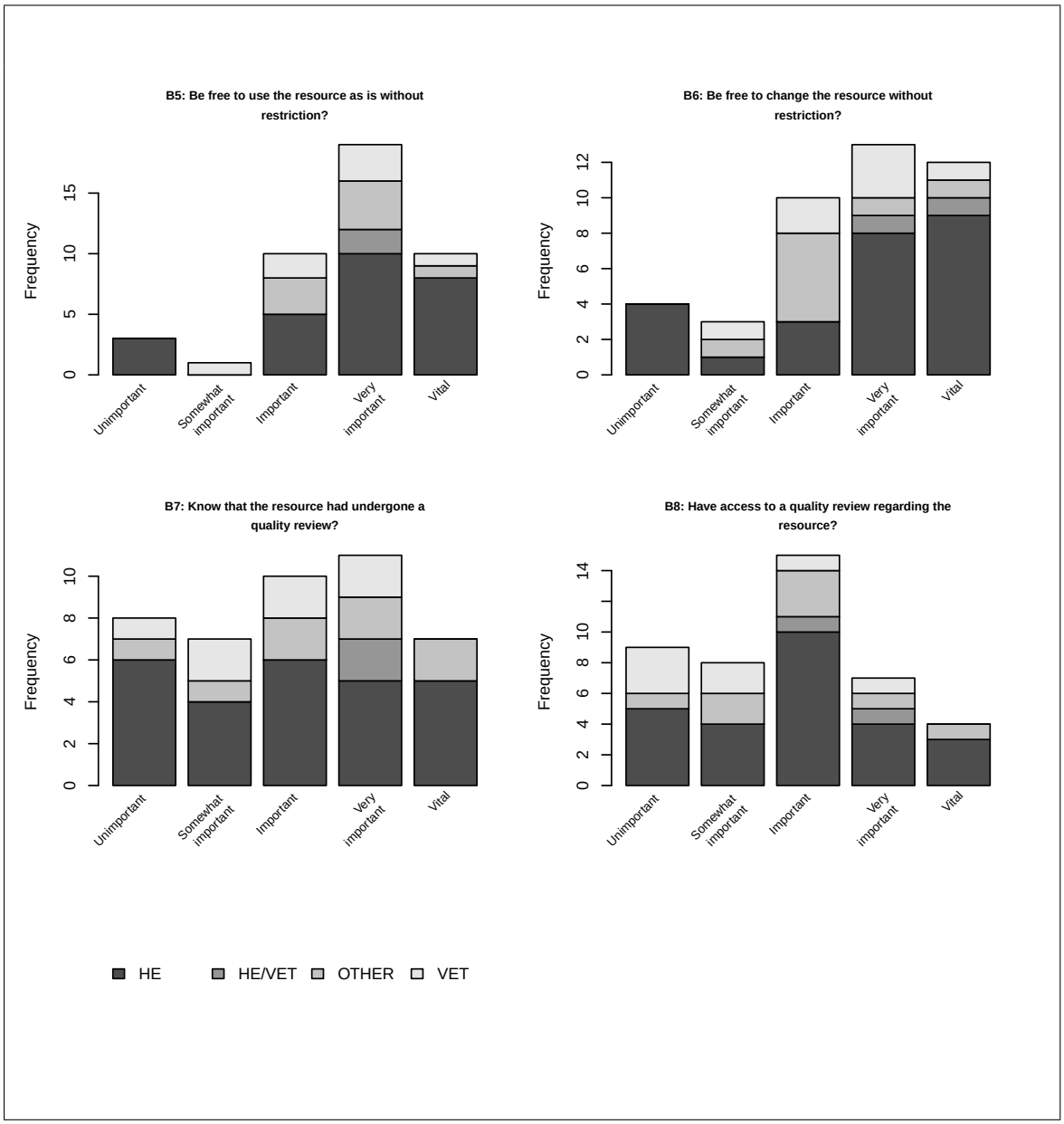


Figure D.2: Survey result details (continued)

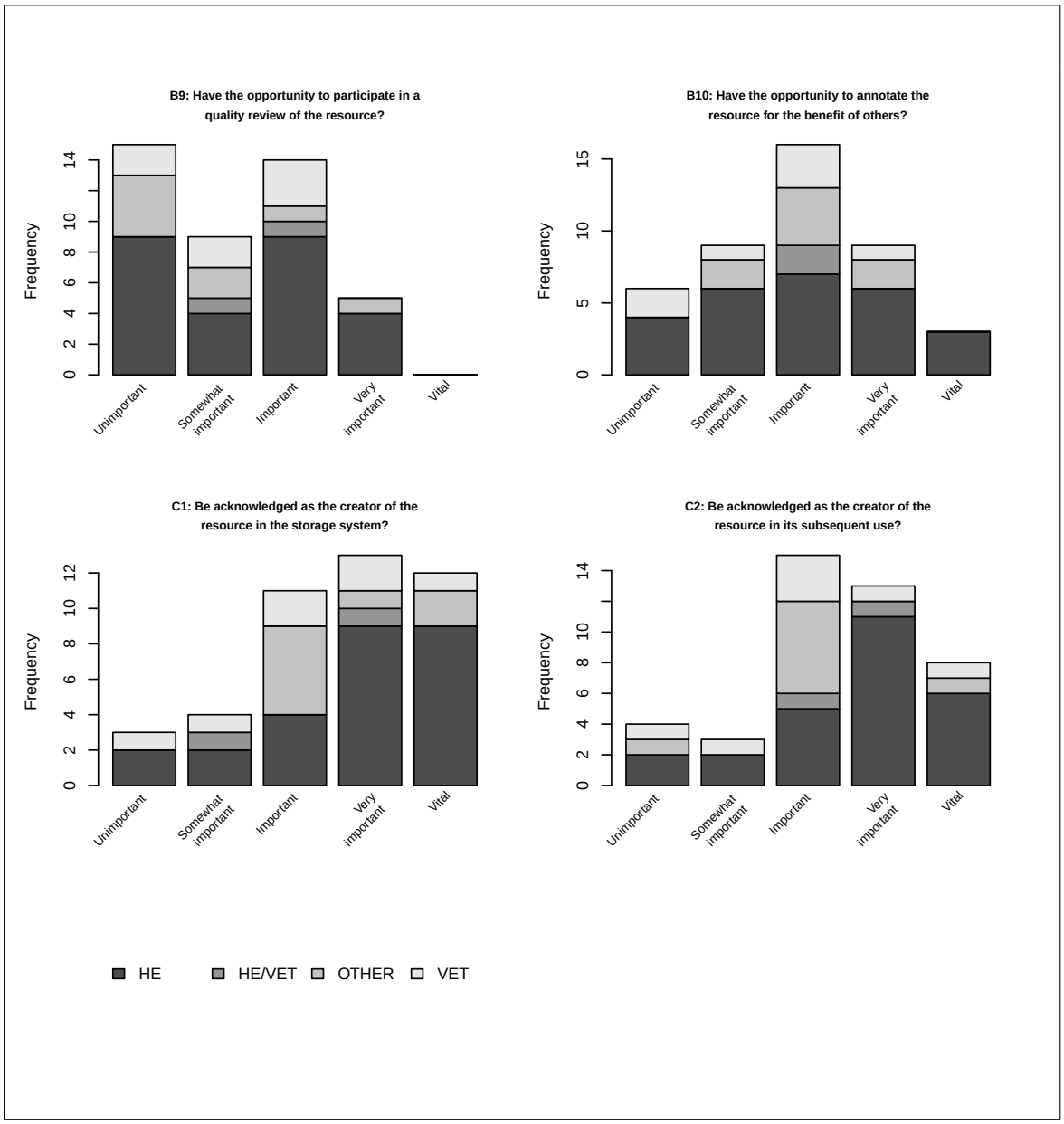


Figure D.2: Survey result details (continued)

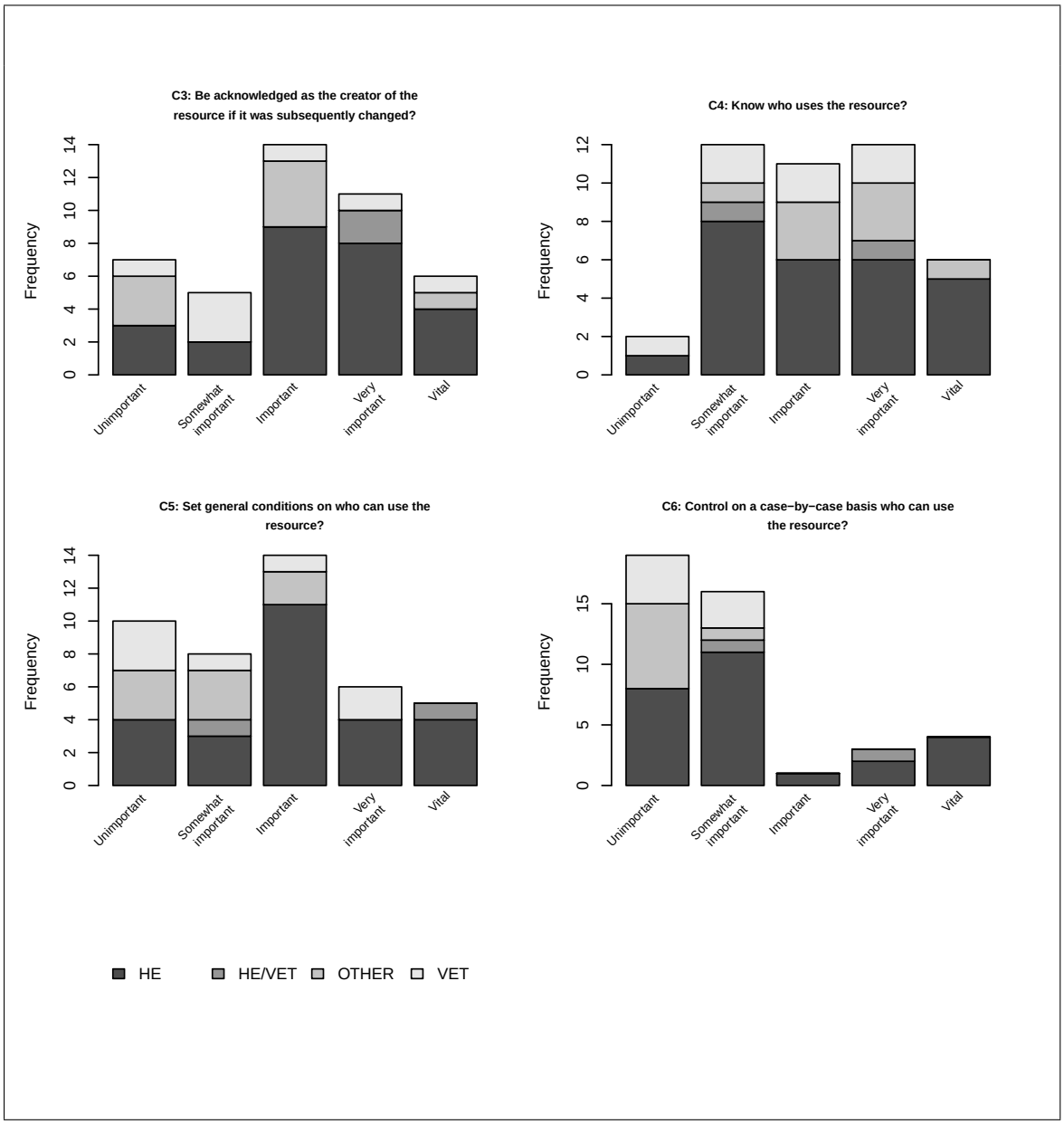


Figure D.2: Survey result details (continued)

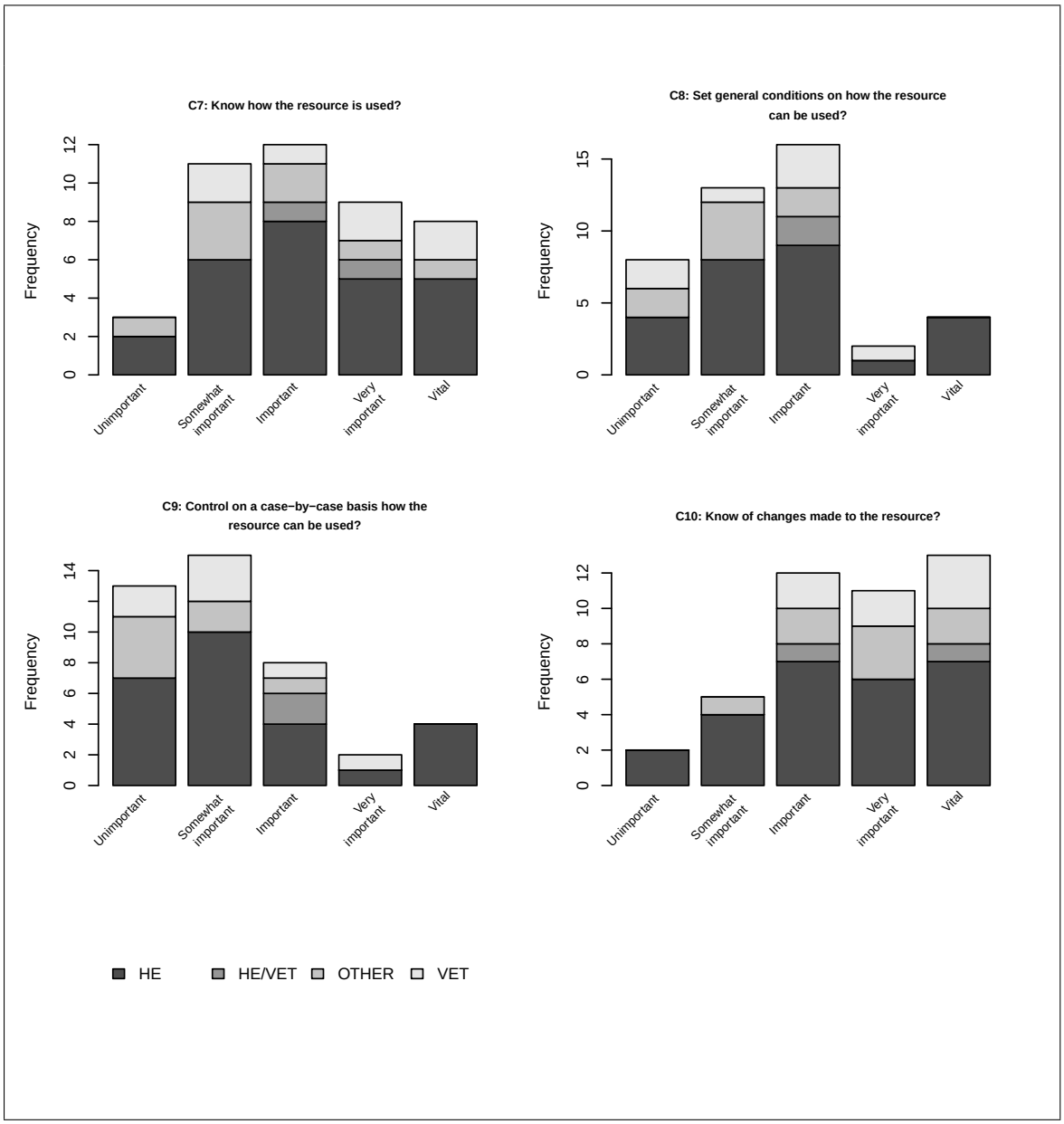


Figure D.2: Survey result details (continued)

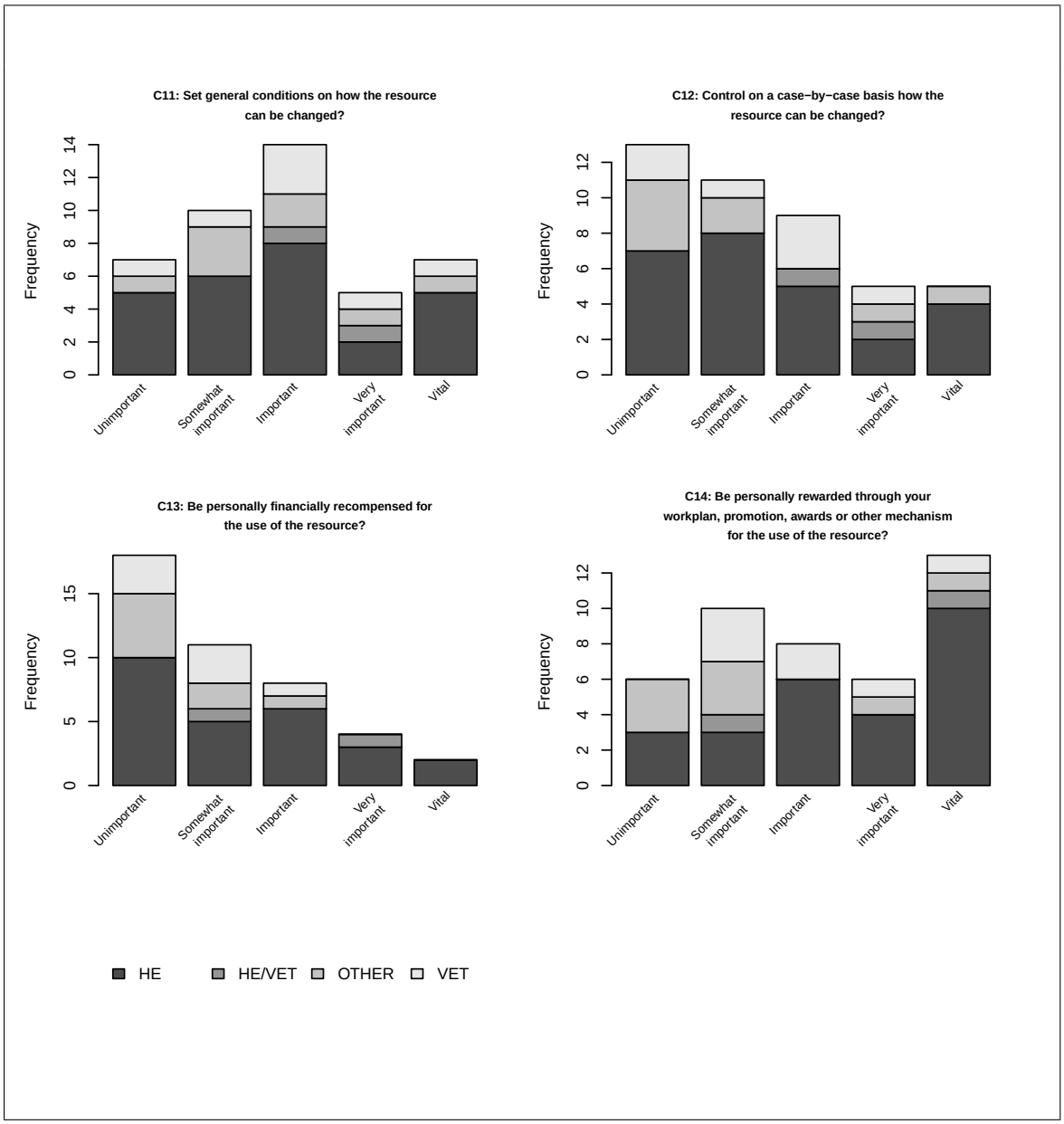


Figure D.2: Survey result details (continued)

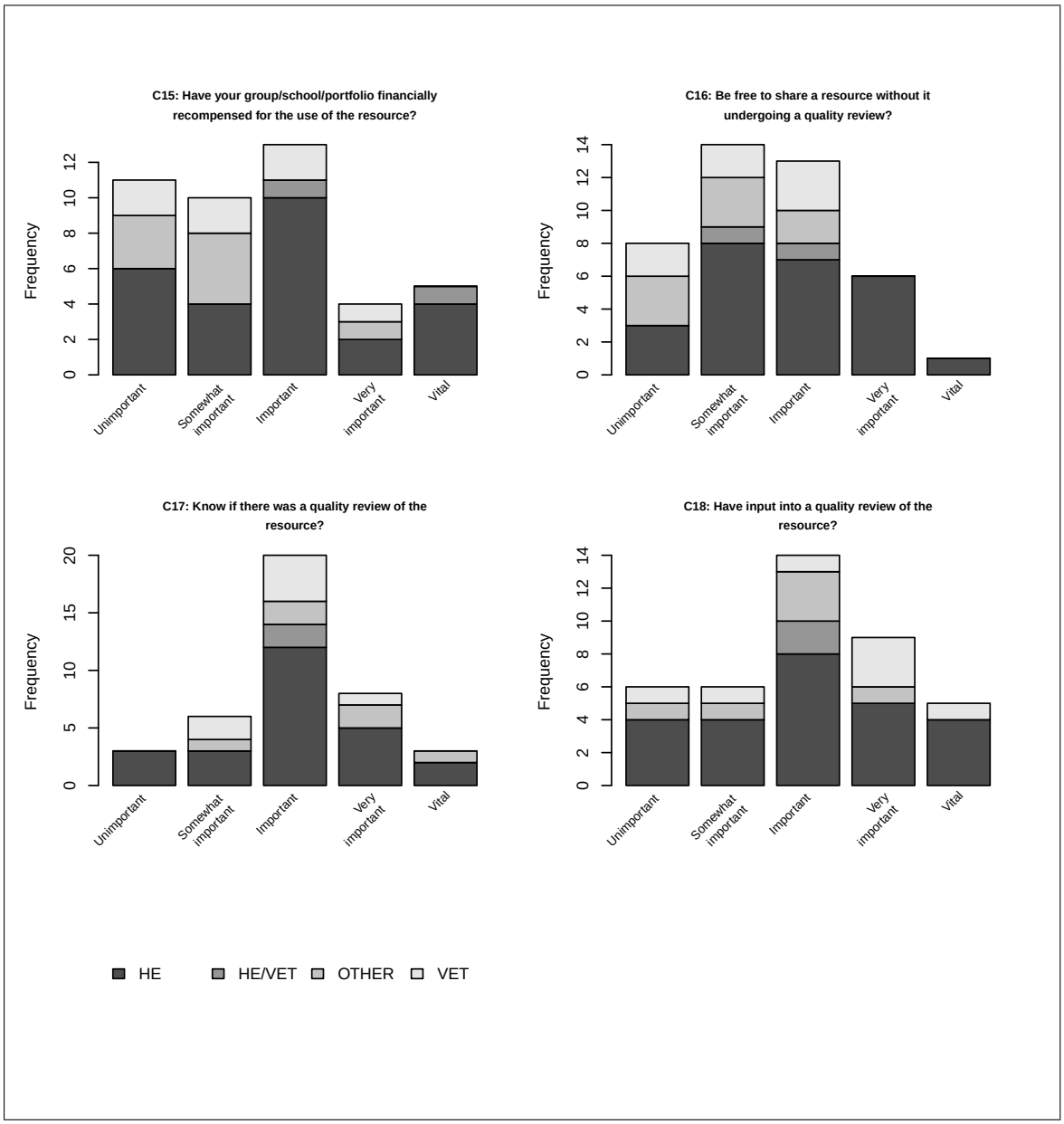


Figure D.2: Survey result details (continued)

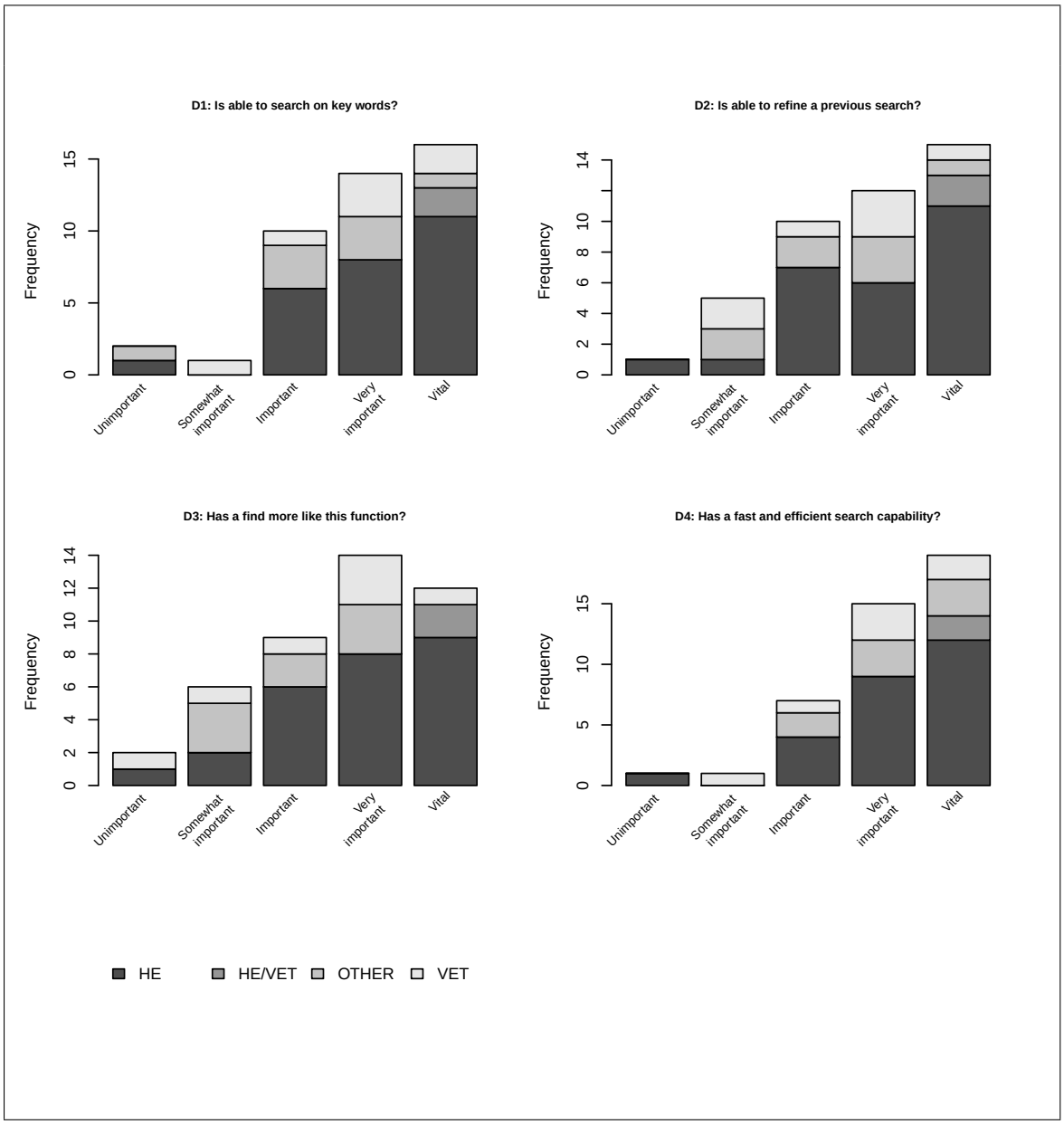


Figure D.2: Survey result details (continued)

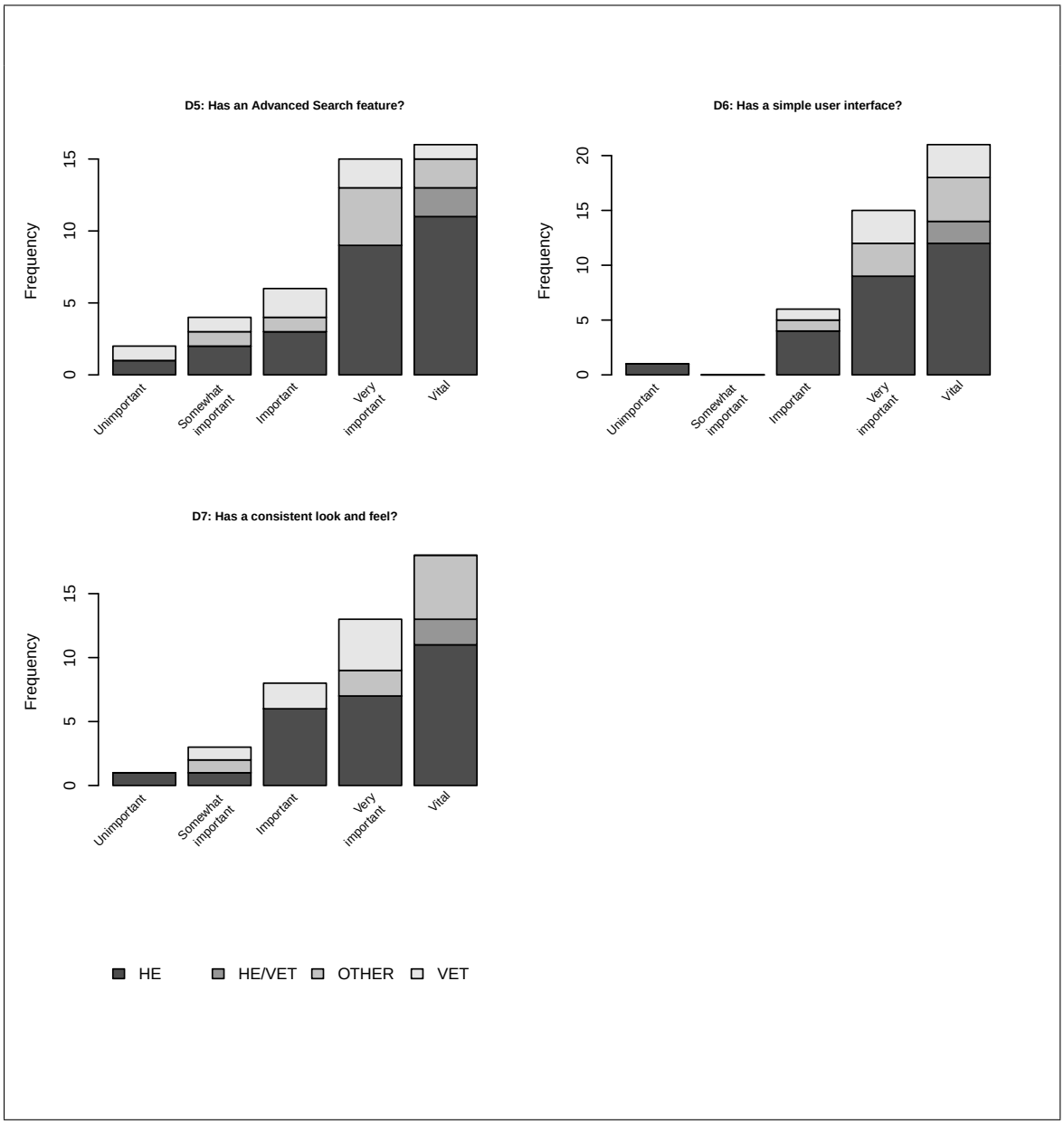


Figure D.2: Survey result details (continued)

Appendix E

Reuse interviews

E.1 Reuse interviews invitation email

The following email was sent to potential interviewees to invite them to participate in our interviews focussing on the successful reuse of educational resources.

> Dear [potential interviewee's name],
>
> [referer's name] suggested I contact you regarding some research that I'm doing about the reuse of digital resources for teaching and learning. I'm currently interviewing a number of people about their work practices, and I was hoping you would agree to participate. The interview would take no more than an hour, and would be run at a time and place convenient to you, sometime in December or early next year.
>
> The research is part of the Reusable Learning Objects (RLO) Project, which is being run through Educational Technology Sub-Committee of the Teaching & Learning Strategy Committee. The Project has been approved by the RMIT University Human Research Ethics Committee. Please see the attached Plain Language Statement and Interview Guide for complete details.
>
> I look forward to hearing from you.
>
> cheers, Michael

E.2 Reuse interview schedule

An interview schedule was included with the email presented in Appendix E.1. The interview schedule, shown in Figure E.1, described what would happen during the interview session and a guide of how long aspects of the interview would take. It also included some questions used to guide the interview.

Interview Schedule

This document sets out the schedule of interview for Stage 2 of the Reusable Learning Objects (RLO) Project.

Pre-interview

- Deliver plain language statement and informed consent form.

Interview preparation – 10 minutes

- Introduce myself.
- Give a brief overview of the Project and its background.
- Give an overview of the current research.
- Ensure the participant has read and understood the plain language statement and signed the informed consent form.

Interview questions – 40 minutes

1. What sorts of RLOs did you reuse?
   - 1.1. What characteristics do they share?
   - 1.2. What file types and granularity were they?
2. How did you reuse these resources?
   - 2.1. In what context were the RLOs reused? Were they reused in the same offering (meaning location) of a course at different times, in the same course but in different offerings, in completely different courses?
   - 2.2. What sort of contextualisation of the RLOs did you do? Was the original resource changed? Was it simply rebadged for a different offering? How was the surrounding material structured to incorporate the RLOs?
   - 2.3. What difficulties did you face when attempting to reuse? What assisted in your efforts to reuse?
   - 2.4. What support, both human and technical did you receive? What extra assistance would have helped?
3. How would you like to reuse resources?
   - 3.1. What sort of resources would you like to find? What information needs do you have?
   - 3.2. What sort of granularity RLO would you like to use?
   - 3.3. What sort of contextualisation would you like to or be willing to perform?
   - 3.4. How would you like to find RLOs?
   - 3.5. What sort of queries might you typically use to find a learning object?

Closing – 10 minutes

- Allow the interviewee to add any further comments.
- Reiterate what will happen to the data, especially that the interviewee will have the opportunity to verify any comments prior to any publication.
- Thank the interviewee.

Post-interview

- Return a copy of the signed consent form to the interviewee.

Figure E.1: Schedule for interviews with people who have successfully reused educational resources.

E.3 Reuse interview plain language statement

Participants in the reuse interviews were presented with a plain language statement before being involved in our research. The plain language statement describes the research, including the context and goals of the work, and a description of the process. The plain language statement is shown on the following pages.

RMIT UNIVERSITY

Educational Media Group
GPO Box 2476V
Melbourne 3001

www.rmit.edu.au

INVITATION TO PARTICIPATE IN A RESEARCH PROJECT

Project Information Statement

Project Title: Reusable Learning Objects Project - Stage 2

Investigators:

Amgad Louka (Manager, Educational Media Group, RMIT University, 9925 9621, amgad.louka@rmit.edu.au)
Michael C. Harris (PhD Candidate, School of CS & IT, 9925 9676, miharris@cs.rmit.edu.au)
Dr James A. Thom (Research Supervisor, Senior Lecturer, School of CS & IT, 9925 2992, jat@cs.rmit.edu.au)

You are invited to participate in a research project being conducted by RMIT University. This information sheet describes the project in straightforward language, or 'plain English'. Please read this sheet carefully and be confident that you understand its contents before deciding whether to participate. If you have any questions about the project, please ask one of the investigators.

Who is involved in this research project? Why is it being conducted?

The Reusable Learning Objects (RLO) Project is run through Educational Technology Sub-Committee of the Teaching & Learning Strategy Committee. A reference group chaired by the Educational Media Group Manager, Amgad Louka, manages the Project, which is intended to support objectives stated within the RMIT Teaching and Learning Strategy 2003-2006 and the University's draft e-learning vision. Michael C. Harris is a PhD candidate, supervised by Dr James A. Thom, examining RLO use.

Why have you been approached?

We are interviewing selected people to investigate the reuse of RLOs in practice. A member of a group involved in the production, use, reuse, or management of RLOs has identified you as someone who has been involved in reuse. Participation is voluntary and confidential and there will be no consequences if you decide not to participate.

What is the project about? What are the questions being addressed?

The RLO Project aims to examine issues around the adoption, implementation, ongoing maintenance and effective management of RLOs as part of curriculum design and development, research and other business activities at RMIT.

In 2004, two project officers completed Stage 1 of the Project in which they:

- Investigated current issues and practices in an RLO approach in education;
- Interviewed and surveyed staff involved with production and use of RLOs to investigate practices at RMIT;
- Identified RMIT initiatives and projects in which management of RLOs is a consideration;
- Identified barriers to the use of RLOs at RMIT;
- Produced a report with recommendations to inform the development and implementation of a RLO approach in teaching and learning at RMIT.

The current research is being conducted as part of Stage 2 of the RLO Project and has been approved by the RMIT University Human Research Ethics Committee. Research questions being addressed are:

1. How do people reuse RLOs in practice?
2. How would people like to be able to reuse RLOs?

If I agree to participate, what will I be required to do?

You will be asked to discuss and reflect upon your work practices and experiences. You will be provided with an interview schedule, including the questions to be covered, before the interview. We anticipate interviews will run for less than 1 hour and will be recorded and transcribed for analysis.

Figure E.2: Reuse interviews plain language statement

What are the risks associated with participation?

There are no risks associated with your participation outside your normal day-to-day work activities.

What are the benefits associated with participation?

The Project aims to improve the University's understanding of an RLO approach to teaching and learning, as outlined above, with the aim of providing increased return on investment and improvements in efficiency. These improvements would in turn be of direct benefit to you.

What will happen to the information I provide?

After transcription, audio tapes of interviews will be stored in the offices of the Educational Media Group in secure storage. Transcripts of interviews will be stored in an encrypted format independently of identifying information on a password-protected computer. Except for the purposes of transcription, access to tapes and transcripts will be restricted to staff of the Educational Media Group, the RLO Project reference group, and Michael C. Harris and his research supervisors. As per RMIT's data retention guidelines, tapes and transcripts will be retained for a period of five years following any relevant publication, after which time they will be destroyed.

The data collected from these interviews will contribute to three possible outputs.

1. They will be reported upon directly in a report to be produced for submission to the Project Reference Group, and then to the Ed. Tech. Subcommittee.
2. A journal or conference paper may be published based upon this work.
3. They will inform the PhD research being conducted by Michael C. Harris.

Non-identifying quotes may be reproduced in these outputs. Interviewees will have the opportunity to verify their statements prior to publication. Every reasonable attempt will be made to ensure your anonymity in any dissemination of the research results; however, given the nature of the research method and the information being collected, there is a slight possibility that you or someone else might be able to be identified by people other than the researchers on the basis of your responses.

What are my rights as a participant?

You have the right to:

- withdraw your participation at any time, without prejudice,
- have any unprocessed data withdrawn and destroyed, provided it can be reliably identified,
- have any questions answered at any time.

By participating in the interview, we assume you have given consent to using this information for our research. Participation is voluntary and participants can withdraw from the process at any time.

Who should I contact if I have any questions?

If you have any questions regarding this research, please contact Michael C. Harris, whose contact details provided above. If you are happy to participate in this project, please sign the attached consent form. Thank you for your time.

Amgad Louka	Michael C. Harris	Dr James A. Thom

Figure E.2: Reuse interviews plain language statement (continued)

Appendix F

Rater agreement task

Chapter 4 describes our user experiment examining rater agreement when judging whether or not resources are likely to support learning. This appendix presents the instructions provided to participants.

F.1 Agreement judgment instructions

Participants in the rater agreement task were asked to read instructions before completing the task. Introductory instructions are shown in Figure F.1, and instructions describing the judgment interfaces, with and without the query visible, are shown in Figure F.2.

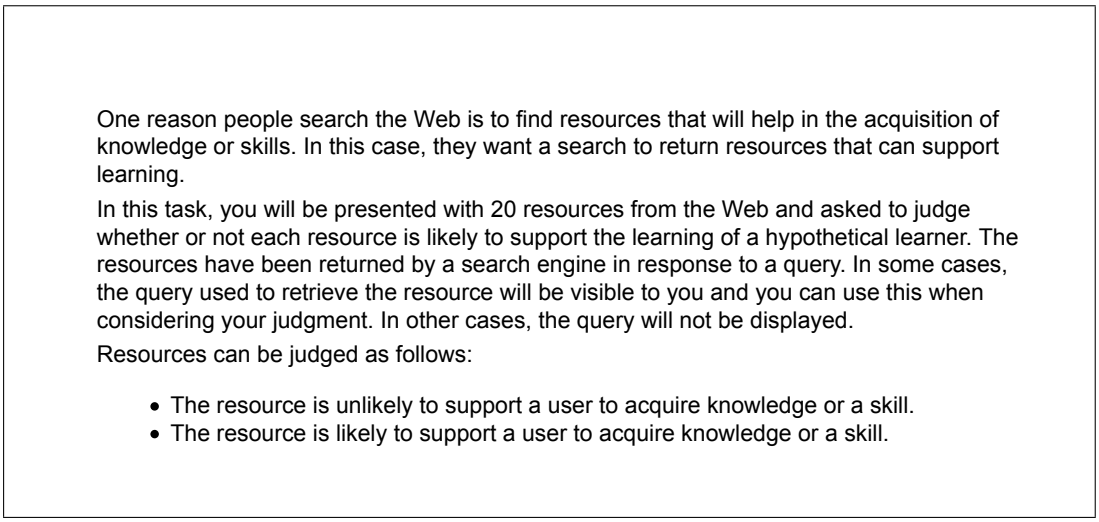
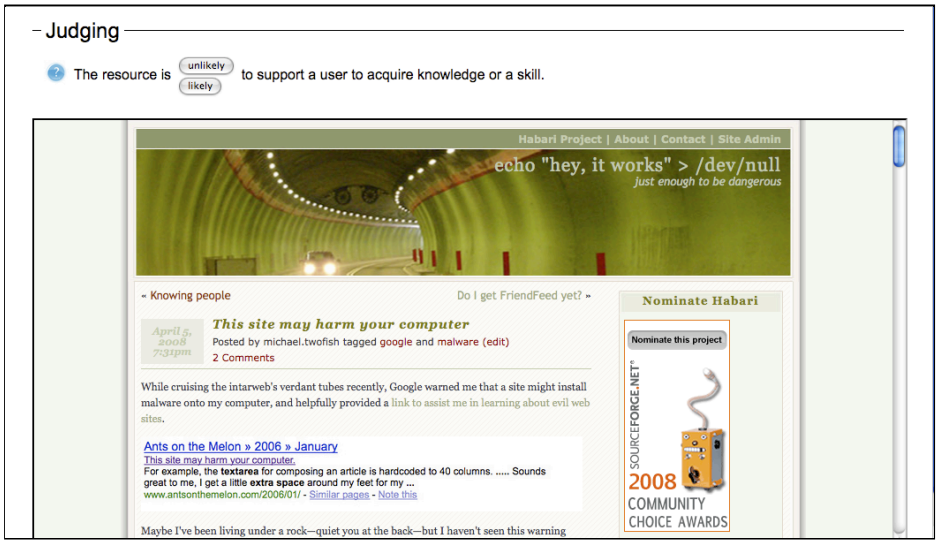


Figure F.1: Introductory instructions for judges in rater agreement experiment

Without the query visible

Below is a screenshot of the judgment interface without the query visible. The resource you are to judge is shown in a frame and you can move around it as you would a normal web page. However, links and forms in the resource have been disabled.



Your judgment will be recorded when you click 'unlikely' or 'likely'.

With the query visible

In the second case, the interface is identical except the query used to retrieve the resource is visible.

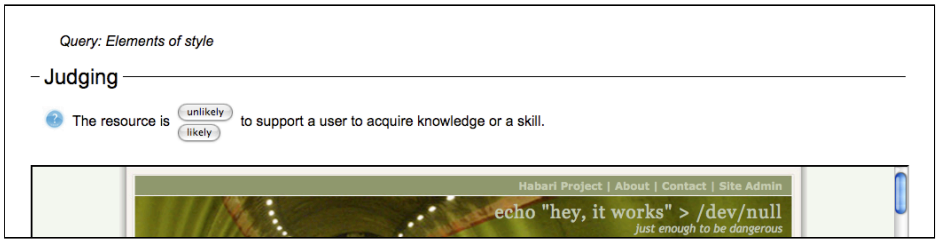


Figure F.2: Instructions describing the interfaces used by judges in rater agreement experiment

Appendix G

Validation collection construction

Chapter 6 describes the construction of a test collection for the validation of classification models developed in Chapter 5. This appendix presents the information provided to participants.

G.1 Collection construction recruitment email

Potential participants in the validation collection construction task were emailed and asked to be volunteer.

> Subject: I need your help for my research into finding learning resources
>
> Hello! I'm looking for volunteers to contribute one hour to a research project that aims to make it easier to find resources to help you learn.
>
> To be involved, you must be an undergraduate, have completed high school (VCE) in Victoria, and be able to attend a one hour session during the week starting 19 April.
>
> If you're willing to help, please respond to this email.
>
> cheers, Michael
>
> -- Michael C. Harris, PhD candidate, School of CS&IT, RMIT University

G.2 Validation collection construction plain language statement

Participants in the validation collection construction task were presented with a plain language statement before being involved in our research. The plain language statement describes the research, including the context and goals of the work, and a description of the process. The plain language statement is shown on the following pages.

RMIT UNIVERSITY

Educational Media Group
GPO Box 2476V
Melbourne 3001
Manager:
Amgad Louka – 9925 9621
amgad.louka@rmit.edu.au
Reusable Learning Objects Project – Stage 1
Project Officers:
Henric Beiers – 9925 3213
henric.beiers@rmit.edu.au
Michael Harris – 9925 9676
miharris@cs.rmit.edu.au

Reusable Learning Objects Project - Stage 1
Plain Language Statement – Focus Groups

10th June 2004

To Whom It May Concern,

You are invited to participate in a project investigating the use, reuse and storage of learning objects at RMIT.

Background
The Reusable Learning Objects Project is being run under the auspices of the Educational Technology Subcommittee of the RMIT Teaching & Learning Strategy Committee. A reference group chaired by Amgad Louka, Manager - Educational Media Group, manages it.

The project is intended to support and effectuate some objectives within the RMIT T&L Strategy 2004 – 2006. In particular:

- 'Improve the quality of teaching – Identify the effective use of technology to support student learning'
- 'Ensure the viability of our educational profile – develop sustainable curriculum and teaching models.'
- 'Enhance organizational capability – Ensuring a robust and reliable infrastructure to support eLearning'

The RLO Project will involve many stages that aim to culminate in the implementation, ongoing maintenance, effective management and adoption of RLOs as part of curriculum design and development, research and other business activities at RMIT.

The Educational Media Group is funding stage 1 of the project. Stage 1 will:

- Investigate current universal issues and practices in the development of a RLO approach in education;
- Interview and survey staff involved with the production and use of learning objects to capture their understanding of, and practices at RMIT;
- Identify RMIT initiatives and projects where the storage of RLOs and interoperability is a consideration (eg. DLS, Integrated Library Management System and eReserve, content management in Schools).
- Investigate standards that would enable interoperability with open archives and other institutional repositories.
- Identify barriers to the use of RLOs at RMIT

Any complaints about your participation in this project may be directed to:
The Secretary, RMIT Human Research Ethics Committee, University Secretariat,
RMIT, GPO Box 2476V, Melbourne, 3001.
The telephone number is (03) 9925 1745.
Details of the complaints procedure are available also from the above address

Figure G.1: Validation collection judgment plain language statement

- Document and report investigations, and
- Provide criteria/business requirements to inform the development and implementation of a RLO approach in teaching and learning at RMIT.

Process
As part of the project we are running a series of focus groups in order to gather the views of staff at RMIT regarding the use of learning objects.

A member of the project reference group, or one of the project team, has identified you as someone who may be interested in participating in these focus groups and contributing to the discussion. Participation in the interviews is voluntary and participants can withdraw from the process at any time they wish.

The focus groups will run for 1½ hours. There will be between 5 and 10 people in each of the focus groups. They will be tape-recorded and then transcribed to allow the project officers to undertake analysis of the data collected.

Tapes and transcripts will be stored in the Learning Objects Project offices of the Educational Media Group in a locked cupboard for a period of five years after which time they will be destroyed. Access will be restricted to staff of the Educational Media Group and staff for the time being working on the Learning Object Project

The data collected from the focus groups will contribute to three possible outputs.
i) The issues identified will form the basis of an online questionnaire for other RMIT staff to respond to.
ii) The issues will be also be reported directly in a report to be produced for submission to the project Reference Group, and then to the Ed. Tech. Subcommittee, by 3rd September 2004.
iii) A journal paper may be published based upon the work of the project.

When the focus groups are reported on, in either the final report or a published paper, there will be no identification of the individual members of the groups. We are more interested in the issues raised and discussion points.

If you have any questions regarding this project please contact any of the project staff. Our contact details are in the header to this letter. You may also contact the Associate Dean Research, Professor Robert Brooks, RMIT Business Portfolio, post: GPO Box 2476V, Melbourne 3001; email: robert.brooks@rmit.edu.au; phone: 9925 5593.

Amgad Louka Henric Beiers Michael Harris

Any complaints about your participation in this project may be directed to:
The Secretary, RMIT Human Research Ethics Committee, University Secretariat,
RMIT, GPO Box 2476V, Melbourne, 3001.
The telephone number is (03) 9925 1745.
Details of the complaints procedure are available also from the above address

Figure G.1: Validation collection judgment plain language statement (continued)

G.3 Validation collection construction instructions

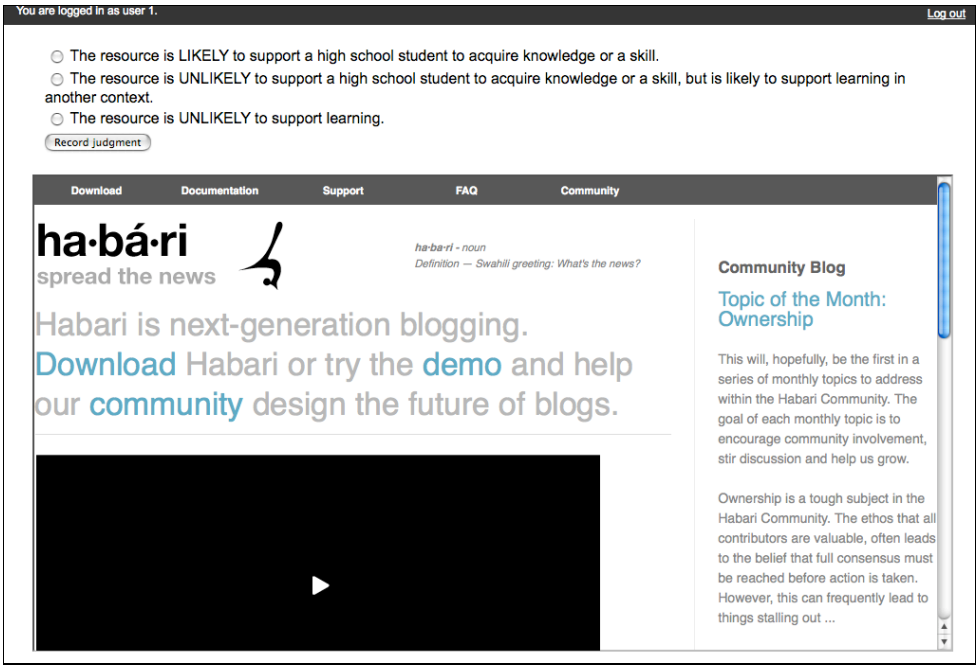
Participants in the validation collection construction task were asked to read instructions that introduced the task, described the process and presented the interface. Screenshots of the instructions are presented on the following pages.

In this task, you will be presented with resources from the Web and asked to judge whether or not each resource is likely to support the learning of a student in high school. The resources have been returned by a search engine in response to a query.

Resources can be judged as follows:

- The resource is LIKELY to support a high school student to acquire knowledge or a skill.
- The resource is UNLIKELY to support a high school student to acquire knowledge or a skill, but is LIKELY to support learning in another context.
- The resource is UNLIKELY to support learning.

Below is a screenshot of the judgment interface. The resource you are to judge is shown in a frame and you can move around it as you would a normal web page. However, links and forms in the resource have been disabled.



Your judgment will be recorded when you click 'Record judgment'.

After each judgment

Once you've judged the resource it will be hidden, and you will be asked some questions. If you'd like to see the resource again to help you answer the questions, click the button labelled 'Show the resource'. Below are all the questions that might be asked, with a short

Figure G.2: Instructions to judges in validation collection construction experiment

explanation.

You will be presented with some simple statements and asked whether you strongly disagree, disagree, agree, or strongly agree with them.

I am confident I have judged this resource correctly.
○ strongly disagree ○ disagree ○ agree ○ strongly agree

It was easy to judge this resource
○ strongly disagree ○ disagree ○ agree ○ strongly agree

I have a high level of expertise in the area covered by this resource
○ strongly disagree ○ disagree ○ agree ○ strongly agree

Some questions have a help icon next to them, like the one in this paragraph. Clicking the icon will reveal help from this page relating to that question.

If you have judged that the resource is educational, you will be asked to estimate the depth of the resource.

The depth is cumulative so, for example, choosing 3 implies the resource is also 1 and 2.

What is the deepest educational level that can describe this resource?
This resource helps learners ...
○ 1. ... recall data or information.
○ 2. ... explain ideas or concepts.
○ 3. ... apply ideas or information in a new situation.
○ 4. ... analyse the relationship between ideas or concepts
○ 5. ... develop a framework linking ideas or concepts.
○ 6. ... make judgments about the value of ideas.

You'll then have the opportunity to provide open comments about the process of judging the resource you've just judged.

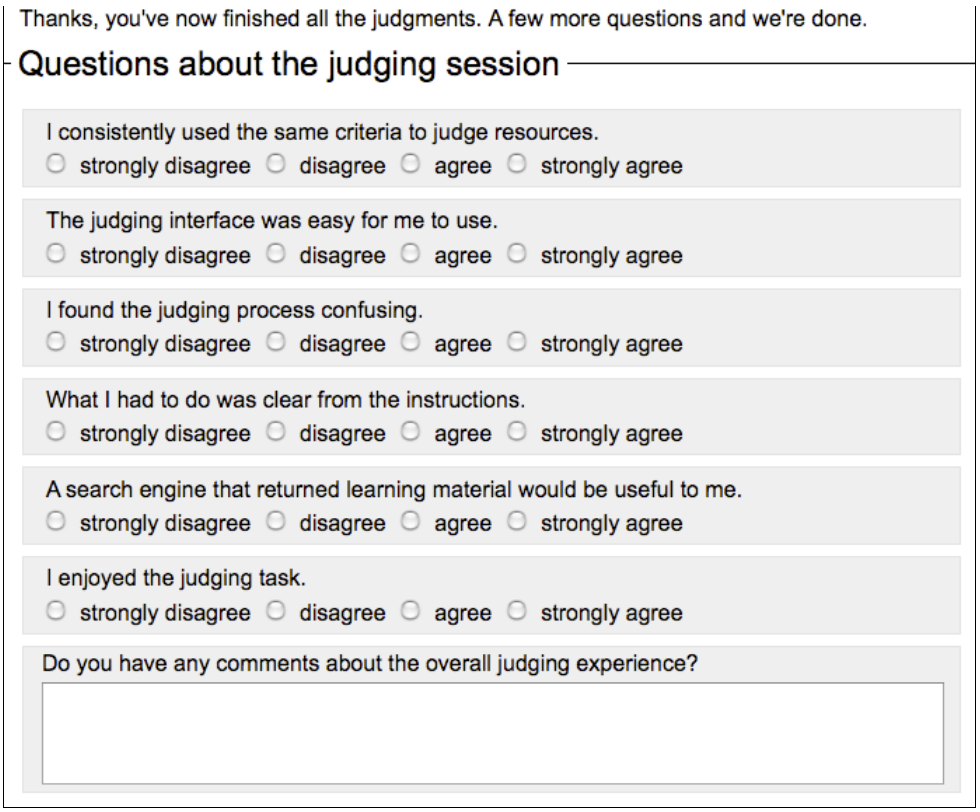
We are especially interested in your thoughts about how and why you made the judgment you did and answered the questions as you did. Comments about why you judged a resource as you did, descriptions of difficulties or issues with the judgment process, or clarification about your answers are welcome.

After all judgments

When you've finished judging all the resources assigned to you, you will be asked some general questions about the entire process. These will again be in the form of simple statements with which you can strongly disagree, disagree, agree, or strongly agree. These questions are the same for all participants and do not depend on how you judged resources.

You will also have the opportunity to provide open comments about the entire judging process, such as general difficulties making judgments or comments about the interface.

Figure G.2: Instructions to judges in validation collection construction experiment (continued)



To give you some experience with the judging process, the interface used for judging is presented in the following pages. The process and interface in this tour is the same as the actual judging, but your responses will not be recorded.

Once you've completed the tour, the judgment session will begin.

Figure G.2: Instructions to judges in validation collection construction experiment (continued)

G.4 Queries used to construct validation collection

- Movement and physical activity
- Health knowledge and promotion
- Building social relationships
- Working in teams
- The individual learner
- Managing personal learning
- Civic knowledge and understanding
- Community engagement
- Creating and making
- Exploring and responding
- Reading
- Writing
- Speaking and listening
- Humanities knowledge and understanding
- Humanities skills
- Economic knowledge and understanding
- Economic reasoning and interpretation
- Geographical knowledge and understanding
- Geospatial skills
- Historical knowledge and understanding
- Historical reasoning and interpretation
- Communicating in a language other than English
- Intercultural knowledge and language awareness
- Number
- Space
- Measurement, chance and data
- Working mathematically
- Structure
- Science knowledge and understanding
- Science at work
- Listening, viewing and responding
- Presenting
- Investigating and designing
- Producing
- Analysing and evaluating
- ICT for visualising thinking
- ICT for creating
- ICT for communicating
- Reasoning, processing and inquiry
- Creativity
- Reflection, evaluation and metacognition

Appendix H

Stop words used in text vectorisation

a	already	apart	because	brief
able	also	appear	become	but
about	although	appreciate	becomes	by
above	always	appropriate	becoming	c
according	am	are	been	came
accordingly	among	around	before	can
across	amongst	as	beforehand	cannot
actually	an	aside	behind	cant
after	and	ask	being	cause
afterwards	another	asking	believe	causes
again	any	associated	below	certain
against	anybody	at	beside	certainly
all	anyhow	available	besides	changes
allow	anyone	away	best	clearly
allows	anything	awfully	better	co
almost	anyway	b	between	com
alone	anyways	be	beyond	come
along	anywhere	became	both	comes

concerning
consequently
consider
considering
contain
containing
contains
corresponding
could
course
currently
d
definitely
described
despite
did
different
do
does
doing
done
down
downwards
during
e
each
edu
eg
eight
either
else
elsewhere
enough
entirely
especially
et
etc
even
ever
every
everybody
everyone
everything
everywhere
ex
exactly
example
except
f
far
few
fifth
first
five
followed
following
follows
for
former
formerly
forth
four
from
further
furthermore
g
get
gets
getting
given
gives
go
goes
going
gone
got
gotten
greetings
h
had
happens
hardly
has
have
having
he
hello
help
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
hi
him
himself
his
hither
hopefully
how
howbeit
however
i
ie
if
ignored
immediate
in
inasmuch
inc
indeed
indicate
indicated
indicates
inner
insofar
instead
into
inward
is
it
its
itself
j
just
k
keep
keeps
kept
know
knows
known
l
last
lately
later
latter
latterly
least
less
lest
let
like
liked
likely
little
ll
look
looking
looks
ltd

m	next	others	really	several
mainly	nine	otherwise	reasonably	shall
many	no	ought	regarding	she
may	nobody	our	regardless	should
maybe	non	ours	regards	since
me	none	ourselves	relatively	six
mean	noone	out	respectively	so
meanwhile	nor	outside	right	some
merely	normally	over	s	somebody
might	not	overall	said	somehow
more	nothing	own	same	someone
moreover	novel	p	saw	something
most	now	particular	say	sometime
mostly	nowhere	particularly	saying	sometimes
much	o	per	says	somewhat
must	obviously	perhaps	second	somewhere
my	of	placed	secondly	soon
myself	off	please	see	sorry
n	often	plus	seeing	specified
name	oh	possible	seem	specify
namely	ok	presumably	seemed	specifying
nd	okay	probably	seeming	still
near	old	provides	seems	sub
nearly	on	q	seen	such
necessary	once	que	self	sup
need	one	quite	selves	sure
needs	ones	qv	sensible	t
neither	only	r	sent	take
never	onto	rather	serious	taken
nevertheless	or	rd	seriously	tell
new	other	re	seven	tends

th
than
thank
thanks
thanx
that
thats
the
their
theirs
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
theres
thereupon
these
they
think
third
this
thorough
thoroughly
those
though
three
through
throughout
thru
thus
to
together
too
took
toward
towards
tried
tries
truly
try
trying
twice
two
u
un
under
unfortunately
unless
unlikely
until
unto
up
upon
us
use
used
useful
uses
using
usually
uucp
v
value
various
ve
very
via
viz
vs
w
want
wants
was
way
we
welcome
well
went
were
what
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
whoever
whole
whom
whose
why
will
willing
wish
with
within
without
wonder
would
would
x
y
yes
yet
you
your
yours
yourself
yourselves
z
zero

Appendix I

Raw classification results

Chapters 5 and 6 present summary measures of runs of 10 times 10-fold cross-validation. This appendix presents the unsummarised raw results, from which all measures can be calculated.

For all results, `a` represents educational and `b` represents not educational. The sum of the numbers in a row is the number of instances judged to be of that class by the human judges, and the sum of the numbers in a column is the number of instances that were predicted by the classifier to be of that class in the current run. Taking the first 10-fold cross-validation run of OneR as an example, we have the following result.

```
  a  b  <-- classified as
 54  7  | a
 17 22  | b
```

This shows that of the $54 + 7 = 61$ resources that judges rated as educational, 54 were correctly classified as educational and 7 incorrectly classified as not educational. Similarly, of the $17 + 22 = 39$ resources that judges rated as not educational, 22 were correctly classified and 17 were incorrectly classified.

I.1 Filtering Educational Resources

This section presents the unsummarised raw results of runs from Chapter 5.

I.1.1 Baseline

Raw baseline results, as described in Section 5.2.2.

ZeroR

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 61  0  |  61  0  |  61  0  |  61  0  |  61  0 |  a
 39  0  |  39  0  |  39  0  |  39  0  |  39  0 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 61  0  |  61  0  |  61  0  |  61  0  |  61  0 |  a
 39  0  |  39  0  |  39  0  |  39  0  |  39  0 |  b
```

OneR

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 54  7  |  56  5  |  53  8  |  55  6  |  54  7 |  a
 17 22  |  15 24  |  14 25  |  16 23  |  13 26 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 52  9  |  55  6  |  54  7  |  54  7  |  56  5 |  a
 17 22  |  15 24  |  13 26  |  15 24  |  15 24 |  b
```

I.1.2 Naïve Bayes

Raw baseline results, as described in Section 5.2.3.

Default

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 52  9  |  52  9  |  52  9  |  52  9  |  52  9 |  a
 11 28  |  12 27  |  11 28  |  11 28  |  11 28 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
```

```
51 10  |  52  9  |  52  9  |  52  9  |  52  9 |  a
11 28  |  11 28  |  11 28  |  11 28  |  10 29 |  b
```

With kernel function

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
52  9  |  52  9  |  52  9  |  52  9  |  52  9 |  a
10 29  |  10 29  |   9 30  |  10 29  |  10 29 |  b
```

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
52  9  |  52  9  |  52  9  |  52  9  |  52  9 |  a
11 28  |  10 29  |  10 29  |  11 28  |  10 29 |  b
```

With discretization

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
54  7  |  53  8  |  54  7  |  53  8  |  52  9 |  a
 8 31  |   6 33  |   6 33  |   7 32  |   7 32 |  b
```

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
54  7  |  53  8  |  53  8  |  53  8  |  54  7 |  a
 6 33  |   6 33  |   8 31  |   5 34  |  10 29 |  b
```

I.1.3 Rules

Raw JRip results, as described in Section 5.2.4.

Default

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
51 10  |  49 12  |  47 14  |  49 12  |  51 10 |  a
11 28  |  13 26  |  10 29  |  13 26  |   6 33 |  b
```

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
51 10  |  49 12  |  52  9  |  44 17  |  48 13 |  a
```

```
 10 29  |  11 28  |  11 28  |  11 28  |   6 33 |  b
```

Tuned subsets

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 47 14  |  49 12  |  51 10  |  49 12  |  50 11 |  a
  9 30  |   9 30  |  11 28  |  14 25  |   8 31 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 54  7  |  46 15  |  47 14  |  47 14  |  50 11 |  a
  6 33  |  11 28  |   5 34  |  12 27  |   9 30 |  b
```

Tuned runs

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 49 12  |  49 12  |  54  7  |  49 12  |  51 10 |  a
  8 31  |   5 34  |   9 30  |  11 28  |   8 31 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 51 10  |  48 13  |  47 14  |  48 13  |  51 10 |  a
  6 33  |   7 32  |   7 32  |   9 30  |   8 31 |  b
```

I.1.4 Trees

Raw J48 results, as described in Section 5.2.5.

Unpruned

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 51 10  |  49 12  |  52  9  |  51 10  |  54  7 |  a
 10 29  |  10 29  |  11 28  |  16 23  |  10 29 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 49 12  |  53  8  |  50 11  |  53  8  |  51 10 |  a
 12 27  |   9 30  |   9 30  |   9 30  |   8 31 |  b
```

Subtree raising

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 52  9  |  50 11  |  52  9  |  52  9  |  55  6 |  a = 0
  7 32  |   7 32  |   8 31  |  13 26  |   7 32 |  b = 1

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 49 12  |  53  8  |  51 10  |  53  8  |  51 10 |  a
  8 31  |   8 31  |   6 33  |   7 32  |   9 30 |  b
```

Subtree replacement

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 51 10  |  50 11  |  52  9  |  52  9  |  54  7 |  a
  7 32  |   7 32  |   8 31  |  13 26  |   7 32 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 49 12  |  53  8  |  50 11  |  53  8  |  51 10 |  a
  8 31  |   6 33  |   6 33  |   7 32  |   9 30 |  b
```

Reduced error

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 54  7  |  53  8  |  55  6  |  53  8  |  56  5 |  a
 10 29  |  13 26  |  14 25  |  16 23  |  11 28 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  57  4  |  55  6  |  55  6  |  57  4 |  a
 12 27  |  11 28  |  13 26  |  13 26  |  13 26 |  b
```

I.1.5 Support vector machines

Raw SMO results, as described in Section 5.2.6.

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
```

```
 53  8  |  52  9  |  50 11  |  52  9  |  53  8 |  a
  7 32  |  10 29  |  10 29  |  10 29  |   9 30 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  53  8  |  53  8  |  53  8  |  51 10 |  a
 12 27  |   9 30  |  11 28  |   8 31  |   9 30 |  b
```

I.1.6 Multilayer Perceptrons

Raw Multilayer Perceptron results, as described in Section 5.2.7.

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 56  5  |  56  5  |  53  8  |  55  6  |  56  5 |  a
  8 31  |  10 29  |   9 30  |  10 29  |   8 31 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  55  6  |  56  5  |  54  7  |  53  8 |  a
 10 29  |  10 29  |  10 29  |   9 30  |   9 30 |  b
```

I.1.7 Boosting

Raw SMO results, as described in Section 5.2.8.

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 57  4  |  56  5  |  55  6  |  55  6  |  59  2 |  a
 10 29  |   7 32  |   5 34  |   5 34  |   4 35 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  58  3  |  57  4  |  56  5  |  55  6 |  a
  9 30  |   7 32  |   7 32  |   7 32  |   6 33 |  b
```

I.1.8 Bagging

Raw SMO results, as described in Section 5.2.9.

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b   <-- classified as
56  5  | 55  6  | 55  6  | 56  5  | 55  6 |  a
11 28  |  9 30  | 10 29  | 11 28  |  9 30 |  b
```

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b   <-- classified as
56  5  | 55  6  | 55  6  | 55  6  | 56  5 |  a
11 28  |  8 31  | 10 29  | 10 29  | 10 29 |  b
```

I.1.9 Stopping

Raw results using stopping in term vector construction, as described in Section 5.2.11.

Naïve Bayes

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b    <-- classified as
56  5  | 56  5  | 56  5  | 54  7  | 54  7  |  a
16 23  | 16 23  | 18 21  | 14 25  | 14 25  |  b
```

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b   <-- classified as
54  7  | 52  9  | 56  5  | 56  5  | 54  7 |  a
19 20  | 14 25  | 15 24  | 14 25  | 15 24 |  b
```

AdaBoost

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b   <-- classified as
54  7  | 55  6  | 55  6  | 56  5  | 56  5 |  a
17 22  | 15 24  | 12 27  | 15 24  | 14 25 |  b
```

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b   <-- classified as
54  7  | 56  5  | 57  4  | 55  6  | 57  4 |  a
16 23  | 13 26  | 13 26  | 15 24  | 13 26 |  b
```

Random Forest

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b   <-- classified as
```

```
 57  4  |  57  4  |  57  4  |  57  4  |  57  4 |  a
 14 25  |  13 26  |  13 26  |  13 26  |  12 27 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 57  4  |  56  5  |  56  5  |  57  4  |  57  4 |  a
 13 26  |  12 27  |  12 27  |  11 28  |  12 27 |  b
```

I.1.10 Stemming

Raw results using stemming in term vector construction, as described in Section 5.2.11.

Naïve Bayes

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  53  8  |  53  8  |  53  8  |  53  8 |  a
  7 32  |   6 33  |   6 33  |   7 32  |   5 34 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  52  9  |  53  8  |  53  8  |  53  8 |  a
  7 32  |   5 34  |   8 31  |   5 34  |   6 33 |  b
```

AdaBoost

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  54  7  |  52  9  |  53  8  |  54  7 |  a
  9 30  |  13 26  |  12 27  |   9 30  |   8 31 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 52  9  |  57  4  |  55  6  |  53  8  |  56  5 |  a
 11 28  |   9 30  |   9 30  |  10 29  |  12 27 |  b
```

Random Forest

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 56  5  |  55  6  |  56  5  |  55  6  |  55  6 |  a
```

```
11 28  |  13 26  |   9 30  |  10 29  |   7 32 |  b
```

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
56  5  |  54  7  |  55  6  |  55  6  |  56  5 |  a
12 27  |   5 34  |   9 30  |   8 31  |  10 29 |  b
```

I.1.11 Word counts

Raw results using word counts in term vector construction, as described in Section 5.2.11.

Naïve Bayes with discretization

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
54  7  |  54  7  |  54  7  |  52  9  |  54  7 |  a
11 28  |   9 30  |   9 30  |   8 31  |   9 30 |  b
```

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
54  7  |  53  8  |  54  7  |  54  7  |  55  6 |  a
 8 31  |   9 30  |   8 31  |   9 30  |  10 29 |  b
```

Naïve Bayes with kernel function

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
50 11  |  47 14  |  48 13  |  46 15  |  48 13 |  a
17 22  |  16 23  |  17 22  |  17 22  |  16 23 |  b
```

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
50 11  |  48 13  |  47 14  |  48 13  |  49 12 |  a
17 22  |  16 23  |  17 22  |  16 23  |  16 23 |  b
```

AdaBoost

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
53  8  |  53  8  |  54  7  |  53  8  |  51 10 |  a
 9 30  |   9 30  |   7 32  |   8 31  |   9 30 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  53  8  |  53  8  |  53  8  |  54  7 |  a
  8 31  |   8 31  |   6 33  |   6 33  |   9 30 |  b
```

Random Forest

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  56  5  |  56  5  |  54  7  |  55  6 |  a
  9 30  |  11 28  |   8 31  |   8 31  |   9 30 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 54  7  |  54  7  |  55  6  |  55  6  |  56  5 |  a
  9 30  |   8 31  |   9 30  |   8 31  |  11 28 |  b
```

I.1.12 Normalise length

Naïve Bayes with discretization

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 51 10  |  50 11  |  52  9  |  51 10  |  51 10 |  a
  9 30  |   6 33  |   6 33  |   7 32  |   6 33 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 51 10  |  52  9  |  51 10  |  53  8  |  51 10 |  a
  9 30  |   5 34  |   5 34  |   6 33  |   7 32 |  b
```

Naïve Bayes with kernel function

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 46 15  |  50 11  |  48 13  |  44 17  |  47 14 |  a
  7 32  |   4 35  |   6 33  |   6 33  |   6 33 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
```

```
48 13  |  47 14  |  48 13  |  47 14  |  47 14 |  a
 8 31  |   3 36  |   5 34  |   6 33  |   7 32 |  b
```

AdaBoost

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
54  7  |  57  4  |  56  5  |  53  8  |  57  4 |  a
 9 30  |   8 31  |  10 29  |   9 30  |   7 32 |  b
```

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
55  6  |  55  6  |  58  3  |  54  7  |  56  5 |  a
 9 30  |  10 29  |   7 32  |   5 34  |   7 32 |  b
```

Random Forest

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
56  5  |  55  6  |  54  7  |  55  6  |  53  8 |   a
10 29  |  10 29  |   9 30  |  10 29  |   8 31 |   b
```

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
55  6  |  55  6  |  54  7  |  55  6  |  55  6 |  a
10 29  |   8 31  |   8 31  |  11 28  |  10 29 |  b
```

I.1.13 Word count and normalise length

Naïve Bayes with discretization

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
53  8  |  53  8  |  54  7  |  53  8  |  52  9 |  a
 3 36  |   4 35  |   4 35  |   4 35  |   4 35 |  b
```

```
 a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
52  9  |  52  9  |  53  8  |  52  9  |  53  8 |  a
 5 34  |   5 34  |   3 36  |   2 37  |   4 35 |  b
```

Naïve Bayes with kernel function

a	b	a	b	a	b	a	b	a	b	<-- classified as
46	15	49	12	47	14	48	13	47	14	a
10	29	9	30	9	30	8	31	9	30	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
47	14	45	16	44	17	49	12	49	12	a
8	31	10	29	8	31	8	31	11	28	b

AdaBoost

a	b	a	b	a	b	a	b	a	b	<-- classified as
51	10	55	6	52	9	55	6	54	7	a
8	31	8	31	10	29	9	30	5	34	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
50	11	52	9	51	10	50	11	53	8	a
12	27	11	28	10	29	11	28	12	27	b

Random Forest

a	b	a	b	a	b	a	b	a	b	<-- classified as
55	6	55	6	55	6	56	5	54	7	a
12	27	9	30	9	30	10	29	11	28	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
55	6	55	6	54	7	54	7	55	6	a
12	27	11	28	8	31	10	29	13	26	b

I.1.14 Term frequency and inverse document frequency

Raw results using $TF \cdot IDF$ in term vector construction, as described in Section 5.2.11.

Naïve Bayes

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 54  7  |  54  7  |  54  7  |  52  9  |  54  7 |  a
 11 28  |   9 30  |   9 30  |   8 31  |   9 30 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 54  7  |  53  8  |  54  7  |  54  7  |  55  6 |  a
  8 31  |   9 30  |   8 31  |   9 30  |  10 29 |  b
```

AdaBoost

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  53  8  |  54  7  |  53  8  |  51 10 |  a
  9 30  |   9 30  |   7 32  |   8 31  |   9 30 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  53  8  |  53  8  |  53  8  |  54  7 |  a
  8 31  |   8 31  |   6 33  |   6 33  |   9 30 |  b
```

Random Forest

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  55  6  |  55  6  |  54  7  |  55  6 |  a
 10 29  |  11 28  |  10 29  |  10 29  |   8 31 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 56  5  |  55  6  |  55  6  |  54  7  |  54  7 |  a
  9 30  |   8 31  |   9 30  |   9 30  |  10 29 |  b
```

I.1.15 Number of internal links

Raw results using the number of external links as a feature, as described in Section 5.3.1.

Naïve Bayes

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  54  7  |  54  7  |  53  8  |  53  8 |  a
  5 34  |   5 34  |   5 34  |   5 34  |   4 35 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  53  8  |  53  8  |  53  8  |  54  7 |  a
  5 34  |   5 34  |   5 34  |   3 36  |   4 35 |  b
```

AdaBoost

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  56  5  |  55  6  |  57  4  |  57  4 |  a
  8 31  |   7 32  |   6 33  |   6 33  |   7 32 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  57  4  |  59  2  |  56  5  |  57  4 |  a
 10 29  |   6 33  |   7 32  |   6 33  |   7 32 |  b
```

Random Forest

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  55  6  |  55  6  |  56  5  |  55  6 |  a
 10 29  |   9 30  |  10 29  |  11 28  |  11 28 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 56  5  |  55  6  |  55  6  |  56  5  |  55  6 |  a
  9 30  |  10 29  |  11 28  |  11 28  |  12 27 |  b
```

I.1.16 Outgoing link count

Raw results using the number of external links as a feature, as described in Section 5.3.1.

Naïve Bayes

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  54  7  |  54  7  |  53  8  |  53  8 |  a
  5 34  |   5 34  |   5 34  |   5 34  |   4 35 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  53  8  |  53  8  |  53  8  |  54  7 |  a
  5 34  |   5 34  |   5 34  |   3 36  |   4 35 |  b
```

AdaBoost

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  56  5  |  55  6  |  57  4  |  57  4 |  a
  9 30  |   8 31  |   7 32  |   7 32  |   7 32 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  57  4  |  59  2  |  55  6  |  57  4 |  a
 10 29  |   6 33  |   7 32  |   6 33  |   5 34 |  b
```

Random Forest

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 56  5  |  55  6  |  55  6  |  55  6  |  56  5 |  a
 15 24  |  11 28  |  12 27  |  13 26  |  12 27 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  55  6  |  55  6  |  56  5  |  55  6 |  a
 12 27  |  11 28  |  12 27  |   9 30  |  13 26 |  b
```

I.1.17 Ratio of link text to overall text

Raw results using the number of external links as a feature, as described in Section 5.3.1.

Naïve Bayes

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 54  7  |  54  7  |  53  8  |  53  8  |  53  8 |  a
  5 34  |   5 34  |   5 34  |   6 33  |   4 35 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  52  9  |  53  8  |  53  8  |  54  7 |  a
  5 34  |   5 34  |   5 34  |   3 36  |   4 35 |  b
```

AdaBoost

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  56  5  |  55  6  |  56  5  |  57  4 |  a
  8 31  |   8 31  |   6 33  |   6 33  |   7 32 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  58  3  |  58  3  |  56  5  |  57  4 |  a
 11 28  |   6 33  |   7 32  |   6 33  |   6 33 |  b
```

Random Forest

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 56  5  |  55  6  |  55  6  |  56  5  |  56  5 |  a
 12 27  |  12 27  |  13 26  |  12 27  |  13 26 |  b
```

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  55  6  |  55  6  |  55  6  |  56  5 |  a
 12 27  |  12 27  |  12 27  |  13 26  |  13 26 |  b
```

I.1.18 Outgoing link text

Raw results using the number of external links as a feature, as described in Section 5.3.1.

Naïve Bayes

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 54  7  |  54  7  |  54  7  |  53  8  |  54  7 |  a
  5 34  |   5 34  |   5 34  |   5 34  |   4 35 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  53  8  |  53  8  |  53  8  |  55  6 |  a
  5 34  |   7 32  |   8 31  |   3 36  |   5 34 |  b
```

AdaBoost

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  56  5  |  55  6  |  57  4  |  57  4 |  a
  7 32  |   7 32  |   9 30  |   6 33  |   7 32 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  57  4  |  60  1  |  55  6  |  56  5 |  a
  8 31  |   6 33  |   7 32  |   8 31  |   7 32 |  b
```

Random Forest

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 56  5  |  55  6  |  56  5  |  55  6  |  55  6 |  a
 13 26  |  10 29  |  11 28  |  13 26  |  13 26 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  55  6  |  57  4  |  55  6  |  56  5 |  a
 13 26  |  10 29  |  13 26  |  13 26  |  11 28 |  b
```

I.1.19 Heading count

Raw results using the number of external links as a feature, as described in Section 5.3.2.

Naïve Bayes

a	b	a	b	a	b	a	b	a	b	<-- classified as
53	8	54	7	54	7	53	8	53	8	a
5	34	5	34	5	34	5	34	4	35	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
54	7	53	8	53	8	53	8	54	7	a
5	34	5	34	5	34	3	36	4	35	b

AdaBoost

a	b	a	b	a	b	a	b	a	b	<-- classified as
55	6	56	5	55	6	57	4	57	4	a
8	31	7	32	6	33	6	33	7	32	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
55	6	57	4	59	2	56	5	57	4	a
10	29	6	33	7	32	6	33	7	32	b

Random Forest

a	b	a	b	a	b	a	b	a	b	<-- classified as
56	5	55	6	55	6	57	4	55	6	a
13	26	14	25	13	26	12	27	11	28	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
55	6	55	6	55	6	55	6	56	5	a
12	27	9	30	12	27	11	28	13	26	b

I.1.20 Ratio of heading text to overall text

Raw results using the number of external links as a feature, as described in Section 5.3.2.

Naïve Bayes

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  54  7  |  54  7  |  53  8  |  53  8 |  a
  5 34  |   5 34  |   5 34  |   5 34  |   4 35 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 53  8  |  53  8  |  53  8  |  53  8  |  54  7 |  a
  5 34  |   5 34  |   5 34  |   3 36  |   4 35 |  b
```

AdaBoost

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  56  5  |  55  6  |  57  4  |  57  4 |  a
  8 31  |   7 32  |   6 33  |   6 33  |   7 32 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  57  4  |  59  2  |  56  5  |  57  4 |  a
 10 29  |   6 33  |   8 31  |   6 33  |   7 32 |  b
```

Random Forest

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  55  6  |  55  6  |  56  5  |  56  5 |  a
 11 28  |  15 24  |  12 27  |  12 27  |  10 29 |  b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  55  6  |  55  6  |  55  6  |  55  6 |  a
 11 28  |  12 27  |  11 28  |  12 27  |  12 27 |  b
```

I.1.21 Resource text and all additional features

Naïve Bayes

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
```

```
 56  5  |  55  6  |  54  7  |  53  8  |  54  7 |   a
  8 31  |   5 34  |   6 33  |   6 33  |   4 35 |   b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  53  8  |  54  7  |  53  8  |  55  6 |  a
  5 34  |   8 31  |   8 31  |   6 33  |   6 33 |  b
```

AdaBoost

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 48 13  |  55  6  |  49 12  |  50 11  |  49 12 |   a
 19 20  |  20 19  |  22 17  |  16 23  |  14 25 |   b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 55  6  |  52  9  |  52  9  |  56  5  |  54  7 |  a
 17 22  |  19 20  |  17 22  |  18 21  |  14 25 |  b
```

Random Forest

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 54  7  |  57  4  |  56  5  |  59  2  |  55  6 |   a
 22 17  |  25 14  |  27 12  |  22 17  |  21 18 |   b

  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 56  5  |  56  5  |  57  4  |  56  5  |  56  5 |  a
 25 14  |  22 17  |  27 12  |  24 15  |  22 17 |  b
```

I.1.22 All additional features

Naïve Bayes

```
  a  b  |   a  b  |   a  b  |   a  b  |   a  b   <-- classified as
 56  5  |  57  4  |  54  7  |  53  8  |  54  7 |   a
 30  9  |  29 10  |  30  9  |  29 10  |  29 10 |   b
```

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b  <-- classified as
53  8  | 55  6  | 54  7  | 56  5  | 55  6 |  a
30  9  | 29 10  | 31  8  | 28 11  | 29 10 |  b
```

AdaBoost

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b  <-- classified as
52  9  | 52  9  | 51 10  | 49 12  | 55  6 |  a
19 20  | 16 23  | 24 15  | 19 20  | 20 19 |  b
```

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b  <-- classified as
53  8  | 53  8  | 52  9  | 50 11  | 55  6 |  a
19 20  | 17 22  | 18 21  | 20 19  | 19 20 |  b
```

Random Forest

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b  <-- classified as
52  9  | 53  8  | 52  9  | 51 10  | 54  7 |  a
23 16  | 24 15  | 23 16  | 22 17  | 23 16 |  b
```

```
 a  b  |  a  b  |  a  b  |  a  b  |  a  b  <-- classified as
51 10  | 51 10  | 51 10  | 51 10  | 53  8 |  a
23 16  | 22 17  | 23 16  | 22 17  | 26 13 |  b
```

I.2 Validating Effectiveness

This section presents the unsummarised raw results of runs from Chapter 6.

I.2.1 High school students and others versus non-educational based on resource text

Results of classification using each model and OneR with identically vectorised input, as shown in Table 6.1.

OneR with Naïve Bayes vectorisation

a	b	a	b	a	b	a	b	a	b	<-- classified as
223	54	235	42	240	37	228	49	231	46	a
109	24	106	27	109	24	111	22	106	27	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
230	47	223	54	237	40	218	59	229	48	a
111	22	102	31	116	17	113	20	104	29	b

Naïve Bayes

a	b	a	b	a	b	a	b	a	b	<-- classified as
169	108	164	113	165	112	163	114	168	109	a
45	88	55	78	46	87	45	88	47	86	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
166	111	160	117	166	111	162	115	164	113	a
46	87	50	83	42	91	49	84	49	84	b

OneR with AdaBoost vectorisation

a	b	a	b	a	b	a	b	a	b	<-- classified as
246	31	258	19	251	26	243	34	249	28	a
112	21	118	15	118	15	105	28	110	23	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
248	29	245	32	248	29	252	25	248	29	a
107	26	116	17	113	20	115	18	116	17	b

AdaBoost

a	b	a	b	a	b	a	b	a	b	<-- classified as
272	5	274	3	270	7	277	0	242	35	a

```
132   1  |  132   1  |  129   4  |  133   0  |  116  17 |   b

  a   b  |    a   b  |    a   b  |    a   b  |    a   b   <-- classified as
270   7  |  261  16  |  253  24  |  251  26  |  239  38 |   a
127   6  |  121  12  |  118  15  |  118  15  |  109  24 |   b
```

OneR with Random Forest vectorisation

```
  a   b  |    a   b  |    a   b  |    a   b  |    a   b   <-- classified as
227  50  |  231  46  |  214  63  |  235  42  |  219  58 |   a
 98  35  |  101  32  |   96  37  |   95  38  |  104  29 |   b

  a   b  |    a   b  |    a   b  |    a   b  |    a   b   <-- classified as
223  54  |  224  53  |  225  52  |  216  61  |  220  57 |   a
 97  36  |  102  31  |  101  32  |  105  28  |   98  35 |   b
```

Random Forest

```
  a   b  |    a   b  |    a   b  |    a   b  |    a   b   <-- classified as
257  20  |  257  20  |  258  19  |  257  20  |  257  20 |   a
 95  38  |   94  39  |   94  39  |   92  41  |   95  38 |   b

  a   b  |    a   b  |    a   b  |    a   b  |    a   b   <-- classified as
258  19  |  253  24  |  258  19  |  258  19  |  260  17 |   a
 94  39  |   92  41  |   93  40  |   94  39  |   93  40 |   b
```

I.2.2 High school students and others versus non-educational based on resource text and all additional features

Results of classification using each model and OneR with identically vectorised input and all additional features, as shown in Table 6.2.

OneR with Naïve Bayes vectorisation

```
  a   b  |    a   b  |    a   b  |    a   b  |    a   b   <-- classified as
```

```
 218  59  |  235  42  |  240  37  |  228  49  |  231  46 |   a
 109  24  |  106  27  |  109  24  |  111  22  |  106  27 |   b
```

```
   a   b |    a   b |    a   b |    a   b |    a   b   <-- classified as
 230  47  |  224  53  |  238  39  |  218  59  |  229  48 |   a
 112  21  |  102  31  |  116  17  |  113  20  |  104  29 |   b
```

Naïve Bayes

```
   a   b |    a   b |    a   b |    a   b |    a   b   <-- classified as
 162 115  |  150 127  |  152 125  |  150 127  |  158 119 |   a
  40  93  |   42  91  |   43  90  |   39  94  |   43  90 |   b
```

```
   a   b |    a   b |    a   b |    a   b |    a   b   <-- classified as
 150 127  |  151 126  |  150 127  |  150 127  |  156 121 |   a
  46  87  |   48  85  |   36  97  |   46  87  |   43  90 |   b
```

OneR with AdaBoost vectorisation

```
   a   b |    a   b |    a   b |    a   b |    a   b   <-- classified as
 235  42  |  231  46  |  234  43  |  235  42  |  235  42 |   a
 113  20  |  111  22  |  113  20  |  113  20  |  114  19 |   b
```

```
   a   b |    a   b |    a   b |    a   b |    a   b   <-- classified as
 230  47  |  240  37  |  237  40  |  239  38  |  227  50 |   a
 107  26  |  118  15  |  108  25  |  116  17  |  114  19 |   b
```

AdaBoost

```
   a   b |    a   b |    a   b |    a   b |    a   b   <-- classified as
 266  11  |  255  22  |  270   7  |  259  18  |  271   6 |   a
 128   5  |  125   8  |  130   3  |  124   9  |  130   3 |   b
```

```
   a   b |    a   b |    a   b |    a   b |    a   b   <-- classified as
```

272	5	270	7	269	8	256	21	264	13	a
132	1	132	1	130	3	127	6	127	6	b

OneR with Random Forest vectorisation

a	b	a	b	a	b	a	b	a	b	<-- classified as
227	50	220	57	227	50	228	49	229	48	a
104	29	104	29	111	22	111	22	105	28	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
232	45	216	61	214	63	228	49	222	55	a
100	33	98	35	103	30	103	30	109	24	b

Random Forest

a	b	a	b	a	b	a	b	a	b	<-- classified as
271	6	272	5	274	3	272	5	274	3	a
133	0	133	0	133	0	133	0	132	1	b

a	b	a	b	a	b	a	b	a	b	<-- classified as
274	3	274	3	273	4	272	5	273	4	a
132	1	132	1	132	1	133	0	132	1	b

Appendix J

Glossary

AUC Area Under (ROC) Curve

IDF Inverse Document Frequency

IR Information Retrieval

LOR Learning Object Repository

MLP Multilayer Perceptron

OER Open Educational Resource

RLO Reusable Learning Object

ROC Receiver Operator Characteristic

SCORM Sharable Content Object Reference Model

SMO Sequential Minimal Optimization

SVM Support Vector Machine

TAFE Tertiary and Further Education

TF Term Frequency

TREC Text REtrieval Conference

VCE Victorian Certificate of Education

VELS Victorian Essential Learning Standards

VET Vocational Education and Training

Bibliography

Hewlett Foundation education program: report to the board. The William and Flora Hewlett Foundation, 2009. http://www.hewlett.org/uploads/files/annual_report/2009_Education_Program_Report_to_the_Board.pdf Accessed: 16 September 2010.

SCORM 2004 4th edition content aggregation model (CAM) version 1.1. Angelo Panar (editor), Advanced Distributed Learning, 2009. http://www.adlnet.gov/Technologies/scorm/SCORMSDocuments/20044thEdition/Documentation.aspx Accessed: 20 August 2010.

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

A. Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, 1992.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, Reading, MA, USA, 1999.

M. J. Bates. Indexing and access for digital libraries and the Internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13):1185–1205, 1998.

N. J. Belkin and W. B. Croft. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, 35(12):29–38, December 1992.

A. Ben-David. Comparison of classification accuracy using Cohen's weighted kappa. *Expert Systems with Applications*, 34(2):825–832, 2008a.

A. Ben-David. About the relationship between ROC curves and Cohen's kappa. *Engineering Applications of Artificial Intelligence*, 21(6):874–882, 2008b.

P. Berkhin. A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping Multidimensional Data: Recent Advances in Clustering*, pages 25–71. Springer, Sydney, Australia, 2006.

A. N. Bissell. Permission granted: open licensing for educational resources. *Open Learning: The Journal of Open and Distance Learning*, 24(1):97–106, 2009.

B. S. Bloom, editor. *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. Longman, White Plains, NY, USA, 1956.

N. Boskic. Learning objects design: What do educators think about the quality and reusability of learning objects? In *Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies (ICALT'03)*, pages 306–307, Athens, Greece, 2003.

R. R. Bouckaert and E. Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In H. Dai, R. Srikant, and C. Zhang, editors, *Advances in Knowledge Discovery and Data Mining, Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 3–12, Sydney, Australia, 2004. Springer.

T. Boyle. Design principles for authoring dynamic, reusable learning objects. *Australian Journal of Educational Technology*, 19(1):46–58, 2003.

A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

M. Braschler and C. Peters. CLEF methodology and metrics. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Cross-Language Information Retrieval and Evaluation, Lecture Notes in Computer Science, Vol. 2406*, pages 394–404. Springer-Verlag, 2002.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

P. Brusilovsky and J. Vassileva. Course sequencing techniques for large-scale web-based education. *International Journal of Continuing Engineering Education and Lifelong Learning*, 13(1/2):75–94, 2003.

C. Buckley and E. Voorhees. Retrieval system evaluation. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, pages 53–75. MIT Press, Cambridge, MA, USA, 2005.

C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10(6):491–508, 2007.

R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing and Management*, 28(5):619–627, 1992.

L. M. Campbell. Engaging with the learning object economy. In A. Littlejohn, editor, *Reusing online resources: a sustainable approach to e-learning*, pages 35–45. Kogan Page, London, UK, 2003.

L. M. Campbell, A. Littlejohn, and C. Duncan. Share and share alike: Encouraging the reuse of academic resources through the Scottish electronic staff development library. *Association for Learning Technology Journal*, 9(2):28–38, 2001.

S. Carson. The unwalled garden: growth of the opencourseware consortium, 2001–2008. *Open Learning: The Journal of Open and Distance Learning*, 24(1):23–29, 2009.

B. Carterette and I. Soboroff. The effect of assessor error on IR system evaluation. In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, and J. Savoy, editors, *SIGIR '10: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–546, Geneva, Switzerland, 2010. ACM.

A. Charlesworth, N. Ferguson, S. Schmoller, N. Smith, and R. Tice. *Sharling eLearning Content - a synthesis and commentary.* Joint Information Systems Committee (JISC), UK, 2007. http://ie-repository.jisc.ac.uk/46/ Accessed: 6 July 2010.

C. M. Christensen, S. Aaron, and W. Clark. Disruption in education. *EDUCAUSE Review*, 38(1): 44–54, 2003.

P. Clark and T. Niblett. Induction in noisy domains. In I. Bratko and N. Lavrac, editors, *Progress in Machine Learning: Proceedings of the 2nd European Working Session on Learning*, pages 11–30, Bled, Yugoslavia, 1987. Sigma Press.

C. W. Cleverdon. The Cranfield tests on index languages devices. *Aslib Proceedings*, 19(6):173–194, 1967.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

W. W. Cohen. Fast effective rule induction. In A. Prieditis and S. J. Russell, editors, *Proceedings of the 12th International Conference on Machine Learning*, pages 115–123, Tahoe City, CA, USA, 1995. Morgan Kaufmann.

K. Collins-Thompson and J. Callan. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462, 2005.

G. V. Cormack. Email spam filtering: A systematic review. *Foundations and Trends Information Retrieval*, 1(4):335–455, 2008.

E. Counts. From Gertie to gigabytes: revealing the world with digital media. *International Journal of Instructional Media*, 33(1):23–31, 2006.

B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison Wesley, Boston, MA, USA, 2009.

C. W. Curtis. *Linear Algebra: An Introductory Approach*. Springer Science and Business Media, New York, NY, USA, fourth edition, 1984.

J. Dalziel. Reflections on the COLIS (collaborative online learning and information systems) demonstrator project and the 'learning object lifecycle'. In A. Williamson, C. Gunn, A. Young, and T. Clear, editors, *Winds of Change in the Sea of Learning: Proceedings of the 19th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education*, pages 159–166, Auckland, New Zealand, 2002. UNITEC Institute of Technology.

H. V. de Sompel, M. L. Nelson, C. Lagoze, and S. Warner. Resource harvesting within the oai-pmh framework. *D-Lib Magazine*, 12(12), 2004.

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In A. Prieditis and S. Russell, editors, *Proceedings of the 12th International Conference on Machine Learning*, pages 194–202. Morgan Kaufmann, 1995.

S. Downes. Learning objects: Resources for distance education worldwide. *International Review of Research in Open and Distance Learning*, July 2001.

S. Downes. Learning objects: Resources for learning worldwide. In R. McGreal, editor, *Online Education Using Learning Objects*, pages 21–31. Routledge/Falmer, London, UK, 2004.

S. Dumais, J. Platt, M. Sahami, and D. Heckerman. Inductive learning algorithms and representations for text categorization. In G. Gardarin, J. C. French, N. Pissinou, K. Makki, and L. Bouganim, editors, *CIKM '98: Proceedings of the 7th international conference on Information and knowledge management*, pages 148–155, Bethesda, MD, USA, 1998. ACM Press.

E. Duval. LearnRank: Towards a real quality measure for learning. In U.-D. Ehlers and J. M. Pawlowski, editors, *Handbook on Quality and Standardisation in E-Learning*, pages 457–463. Springer Berlin Heidelberg, 2006.

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

T. Fawcett. ROC graphs: Notes and practical considerations for researchers. Technical report, Intelligent Enterprise Technologies Laboratory, HP Laboratories, Palo Alto, CA, USA, 2004.

A. Finn, N. Kushmerick, and B. Smyth. Fact or fiction: Content classification for digital libraries. In A. Smeaton and J. Callan, editors, *Proceedings of the 2nd DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries*, Dublin, Ireland, 2001. http://www.ercim.eu/publication/ws-proceedings/DelNoe02/ Accessed: 17 November 2010.

R. A. Fisher. *Statistical methods for research workers.* Oliver & Boyd, Edinburgh, UK, 1950.

P. A. Flach. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Proceedings of the 20th International Conference on Machine Learning*, pages 194–201, Washington, DC, USA, 2003.

J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382, 1971.

A. Fontana and J. Frey. The interview: From structured questions to negotiated text. In N. Denzin and Y. Lincoln, editors, *Handbook of qualitative research*, pages 645–672. SAGE Publications, Thousand Oaks, CA, USA, second edition, 2000.

F. J. Fowler. *Survey research methods.* SAGE Publications, Thousand Oaks, CA, USA, third edition, 2002.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, pages 148–156, San Francisco, CA, USA, 1996. Morgan Kaufmann.

N. Friesen. Three objections to learning objects. In R. McGreal, editor, *Online Education Using Learning Objects*, pages 59–70. Routledge/Falmer, London, UK, 2004.

J. Futrelle, S.-S. Chen, and K. C. Chang. NBDL: a CIS framework for NSDL. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 124–125, 2001.

M. D. Gall, W. R. Borg, and J. P. Gall. *Educational research: an introduction.* Longman Publishing Group, White Plains, NY, USA, sixth edition, 1996.

A. S. Gibbons, J. Nelson, and R. Richards. The nature and origin of instructional objects. In D. A. Wiley, editor, *The Instructional Use of Learning Objects*. Agency for Instructional Technology, Bloomington, IN, USA, 2001.

S. Goldrei, J. Kay, and R. J. Kummerfeld. Exploiting user models to automate the harvesting of metadata for learning objects. In A. Cristea, R. Carro, and F. Garzotto, editors, *Proceedings of the 3rd International Workshop on Authoring of Adaptive and Adaptable Educational Hypermedia*, pages 19–26, Amsterdam, The Netherlands, 2005.

D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. John Wiley and Sons, Inc., New York, NY, USA, 1966.

J. R. Griffiths and P. Brophy. Student searching behavior and the web: use of academic resources and google. *Library Trends*, 53(4):539–554, 2005.

K. Gwet. *Statistical tables for inter-rater agreement*. STATAXIS Publishing Company, Gaithersburg, MD, USA, 2001.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18, November 2009.

M. J. Hannafin and J. R. Hill. Resource-based learning. In J. M. Spector, M. D. Merrill, J. van Merrienboer, and M. P. Driscoll, editors, *Handbook of Research on Educational Communications and Technology*, pages 525–536. Taylor and Francis, New York, NY, USA, third edition, 2007.

S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47:37–49, 1996.

D. Hawking and J. Zobel. Does topic metadata help with web search? *Journal of the American Society for Information Science and Technology*, 58(5):613–628, 2007.

A. J. Head and M. B. Eisenberg. How today's college students use Wikipedia for course-related research. *First Monday*, 15(3), 2010.

M. Hedstrom. Recordkeeping metadata: presenting the results of a working meeting. *Archival Science*, 1(3):243–251, 2001.

E. Heinrich and J. Chen. A framework for the multi-modal description of learning objects. In *Dublin Core Conference*, pages 32–37, 2001.

J. R. Hill and M. J. Hannafin. Teaching and learning in digital environments: The resurgence of resource-based learning. *Educational Technology Research and Development*, 49(3):37–52, 2001.

W. Hodgins. The future of learning objects. In D. A. Wiley, editor, *The Instructional Use of Learning Objects*. Agency for Instructional Technology, Bloomington, IN, USA, 2001.

R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–91, 1993.

P. Ingwersen. Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of Documentation*, 52(1):3–50, 1996.

K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, Athens, Greece, 2000. ACM.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

G. H. John and P. Langley. Estimating continuous distributions in Bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

C. Kayrooz and C. Trevitt. *Research in organisations and communities : tales from the real world*. Allen & Unwin, Crows Nest, NSW, Australia, 2004.

G. Kazai, N. Gövert, M. Lalmas, and N. Fuhr. The INEX evaluation initiative. In H. Blanken, T. Grabs, H.-J. Schek, R. Schenkel, and G. Weikum, editors, *Intelligent search on XML data, Lecture Notes in Computer Science, Vol. 2818*, pages 279–293. Springer-Verlag, 2003.

D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends Information Retrieval*, 3(1—2):1–224, 2009.

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5): 604–632, 1999.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1143. Morgan Kaufmann, 1995.

T. Koppi and N. Lavitt. Institutional use of learning objects three years on: Lessons learned and future directions. In D. Lassner and C. McNaught, editors, *EdMedia 2003 - World Conference on Educational Multimedia, Hypermedia & Telecommunications*, pages 644–648, Honolulu, HI, USA, 2003.

T. Koppi, L. Bogle, and M. Bogle. Learning objects, repositories, sharing and reusability. *Open Learning: The Journal of Open and Distance Learning*, 20(1):83–91, 2005.

H. L. Kundel and M. Polansky. Measurement of observer agreement. *Radiology*, 228(2):303–308, 2003.

Y. Kural, S. E. Robertson, and S. Jones. Clustering information retrieval search outputs. In *Proceedings of the 21st BCS IRSG Colloquium on Information Retrieval*, Workshops in Computing, Newcastle upon Tyne, UK, February 1999. BCS.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

P. D. Leedy and J. E. Ormrod. *Practical research: planning and design.* Prentice Hall, Upper Saddle River, NJ, USA, eighth edition, 2005.

D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, Madison, WI, USA, 1998.

A. Littlejohn. Staff development in the reuse of resources. In A. Littlejohn, editor, *Reusing online resources: a sustainable approach to e-learning.* Kogan Page, London, UK, 2003.

A. Littlejohn, I. Jung, and L. Broumley. A comparison of issues in the reuse of resources in schools and colleges. In A. Littlejohn, editor, *Reusing online resources: a sustainable approach to e-learning.* Kogan Page, London, UK, 2003.

T. W. Malone, K. R. Grant, F. A. Turbak, S. A. Brobst, and M. D. Cohen. Intelligent information-sharing systems. *Communications of the ACM*, 30(5):390–402, 1987.

K. Markey. Twenty-five years of end-user searching, part 1: Research findings. *Journal of the American Society for Information Science and Technology*, 58(8):1071–1081, 2007.

S. J. Mason and N. E. Graham. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128(584):2145–2166, 2002.

J. A. Maxwell. *Qualitative Research Design : An Interactive Approach (Applied Social Research Methods).* SAGE Publications, Thousand Oaks, CA, USA, 1994.

C. McNaught. Identifying the complexity of factors in the sharing and reuse of resources. In A. Littlejohn, editor, *Reusing online resources: a sustainable approach to e-learning*, pages 199–211. Kogan Page, London, UK, 2003.

D. M. Merrill. Instructional transaction theory (ITT): Instructional design based on knowledge objects. In C. M. Reigeluth, editor, *Instructional-Design Theories and Models: A New Paradigm of Instructional Theory*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2000.

M. B. Miles and M. A. Huberman. *Qualitative Data Analysis: An Expanded Sourcebook*. SAGE Publications, Thousand Oaks, CA, USA, second edition, 1994.

M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–281, Dublin, Ireland, 1994. Springer-Verlag NY, NY, USA.

C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003.

J. Najjar, S. Ternier, and E. Duval. The actual use of metadata in ARIADNE: An empirical analysis. In *Proceedings of the 3rd Annual Ariadne Conference*, pages 1–6, Katholieke Universiteit Leuven, Belgium, 2003.

J. C. Nesbit and K. Belfer. Collaborative evaluation of learning objects. In R. McGreal, editor, *Online Education Using Learning Objects*, pages 138–153. Routledge/Falmer, London, UK, 2004.

D. F. Noble. Digital diploma mills: The automation of higher education. *First Monday*, 1(3), 1998.

X. Ochoa and E. Duval. Use of contextualized attention metadata for ranking and recommending learning objects. In *CAMA 2006: Proceedings of the 1st international workshop on Contextualized attention metadata: collecting, managing and exploiting of rich usage information*, pages 9–16, Arlington, VA, USA, 2006. ACM.

X. Ochoa, C. K., M. Meire, and E. Duval. Frameworks for the automatic indexation of learning management systems content into learning object repositories. In *EdMedia 2005 - World Conference on Educational Multimedia, Hypermedia & Telecommunications*, pages 1407–1414, Montreal, Canada, 2005.

F. Oldenettel and M. Malachinski. The LEBONED metadata architecture. In *WWW2003 - Proceedings of the 12th Intenational World Wide Web Conference*, Budapest, Hungary, 2003.

R. Oliver and J. Herrington. Factors influencing quality online learning experiences. In G. Davies and E. Stacey, editors, *Quality Education @ a Distance*, pages 137–142. Kluwer Academic Publishers, London, UK, 2003.

R. Oliver, R. Wirski, L. Wait, and V. Blanksby. Learning designs and learning objects: where pedagogy meets technology. In C.-K. Looi, D. Jonassen, and M. Ikeda, editors, *Towards sustainable and scalable educational innovations informed by the learning sciences*, pages 330–337. IOS Press, Amsterdam, The Netherlands, 2005.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Stanford Digital Library Technologies Working Paper, Stanford University, CA, USA, 1998.

A. Peshkin. The goodness of qualitative research. *Educational Researcher*, 22(2):23–29, 1993.

R. A. Phillips, R. Gururajan, S. Rai, F. Sudweeks, M. Jones, D. Shiers, and R. O'Neil. Education researchers report interaction of IT systems & repositories project: Use and useability of learning objects within the COLIS demonstrator framework. 2003. http://eprints.usq.edu.au/3081/1/Phillips_Rai_Sudweeks_Gururajan_Jones_Shiers_O'Neil.pdf Accessed: 9 November 2010.

J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 185–208. MIT Press, Cambridge, MA, USA, 1998.

P. R. Polsani. Use and abuse of reusable learning objects. *Journal of Digital Information*, 3(4), 2003. http://journals.tdl.org/jodi/article/viewArticle/89/88 Accessed: 10 November 2010.

M. F. Porter. Snowball: A language for stemming algorithms, 2001. http://snowball.tartarus.org/texts/introduction.html Accessed: 10 August 2010.

D. M. W. Powers. Evaluation: From precision, recall and F-factor to ROC, informedness, markedness and correlation. Technical report, Flinders University, Adelaide, South Australia, 2007.

F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 43–48, Newport Beach, CA, USA, 1997. AAAI Press.

F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th International Conference on Machine Learning*, pages 445–453, Madison, WI, USA, 1998.

R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, USA, 1993.

S. E. Robertson. The parameter description of retrieval tests part I: The basic parameters. *Journal of Documentation*, 25(1):1–27, 1969.

S. E. Robertson. Ranking in principle. *Journal of Documentation*, 34(2):93–100, 1978.

C. Robson. *Real world research.* Blackwell Publishing Ltd, Oxford, UK, second edition, 2002.

D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW2004 - Proceedings of the 13th Intenational World Wide Web Conference*, pages 13–19, New York, NY, USA, 2004. ACM.

F. Rosenblatt. The perception: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. Reprinted in *Neurocomputing: foundations of research*, pages 89–114. MIT Press, Cambridge, MA, USA, 1988.

W. A. Rosenkrantz. *Introduction to Probability and Statistics for Science, Engineering, and Finance.* Taylor and Francis, Boca Raton, FL, USA, 2009.

M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, WI, USA, 1998. AAAI Technical Report WS-98-05.

S. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317–327, 1997.

M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends Information Retrieval*, 4(4):247–375, 2010.

T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):1915–1933, 2007.

C. Schaffer. A conservation law for generalization performance. In *Proceedings of the 11th Intenational Machine Learning Conference*, pages 259–265, Hingham, MA, USA, 1994. Kluwer Academic Publishers.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

K. Schlusmans, R. Koper, and W. Giesbertz. Work processes for the development of integrated e-learning courses. In W. Jochems, J. van Merrienboer, and R. Koper, editors, *Integrated eLearning*, pages 26–138. RoutledgeFalmer, London, UK, 2004.

F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1): 1–47, 2002.

B. Shapira, P. Shoval, and U. Hanani. Experimentation with an information filtering system that combines cognitive and sociological filtering integrated with user stereotypes. *Decision Support Systems*, 27(1-2):5–24, 1999.

J. Sim and C. C. Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 85(3):257–268, 2003.

B. Simon, Z. Miklós, W. Nejdl, M. Sintek, and J. Salvachua. Smart space for learning: A mediation infrastructure for learning services. In *WWW2003 - Proceedings of the 12th Intenational World Wide Web Conference*, Budapest, Hungary, 2003.

A. F. Smeaton. Indexing, browsing and searching digital video and digital audio information. In M. Agosti, F. Crestani, and G. Pasi, editors, *Proceedings of the 3rd European Summer-School on Lectures on Information Retrieval-Revised Lectures, Lecture Notes on Information Retrieval*, pages 93–110, Varenna, Italy, 2001.

M. S. Smith and C. M. Casserly. The promise of open educational resources. *Change*, 38(5):8–17, 2006.

M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009.

J. B. South and D. W. Monson. A university-wide system for creating, capturing, and delivering learning objects. In D. A. Wiley, editor, *The Instructional Use of Learning Objects*. Agency for Instructional Technology, Bloomington, IN, USA, 2001.

K. Sparck Jones and C. J. van Rijsbergen. Report on the need for and provision of an "ideal" test collection. Technical report, University Computer Laboratory, Cambridge, UK, 1975.

P. Taylor and S. Richardson. *Constructing a National Scheme for External Peer Review of ICT-based Teaching and Learning Resources*. Evaluations and Investigations Programme, Higher Education Division, DEST, Commonwealth of Australia, 2001.

S. Ternier, D. Massart, A. C. S. Guinea, S. Ceri, and E. Duval. Interoperability for searching learning object repositories: The prolearn query language. *D-Lib Magazine*, 14(1/2), 2008.

C. A. Twigg. *Who Owns Online Courses and Course Materials? Intellectual Property Policies for a New Learning Environment*. The Pew Learning and Technology Program, Center for Academic

Transformation Rensselaer Polytechnic Institute, Troy, NY, USA, 2000. http://www.thencat.org/Monographs/Whoowns.html Accessed: 30 June 2010.

J. Uebersax. Raw agreement indices, March 2008. http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm Accessed: 20 August 2010.

A. Uitdenbogerd. Web readability and computer-assisted language learning. In L. Cavedon and I. Zukerman, editors, *Proceedings of the Australasian Language Technology Workshop*, pages 99–106, Sydney, Australia, 2006.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 1979.

V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

E. M. Voorhees. On expanding query vectors with lexically related words. In *The Second Text Retrieval Conference (TREC 2)*, NIST Special Publication 500-215, pages 223–231, Department of Commerce, National Institute of Standards and Technology, 1993.

E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323, Melbourne, Australia, 1998. ACM.

E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.

E. M. Voorhees. Evaluation by highly relevant documents. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, New Orleans, LA, USA, 2001a. ACM.

E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, London, UK, 2001b. Springer-Verlag.

P. Wallis and J. A. Thom. Relevance judgements for assessing recall. *Information Processing and Management*, 32(3):273–286, 1996.

D. Wiley and S. Gurrell. A decade of development. *Open Learning: The Journal of Open and Distance Learning*, 24(1):11–21, 2009.

D. Wiley, S. Waters, B. Lambert, D. Dawson, M. Barclay, D. Wade, and L. Nelson. Overcoming the limitations of learning objects. *Journal of Educational Multimedia Hypermedia*, 13(4):507–521, 2004.

D. A. Wiley. *Learning object design and sequencing theory*. PhD thesis, Brigham Young University, Provo, UT, USA, 2000a.

D. A. Wiley. Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In D. A. Wiley, editor, *The Instructional Use of Learning Objects*. Agency for Instructional Technology, Bloomington, IN, USA, 2000b.

D. A. Wiley. The learning objects literature. In J. M. Spector, M. D. Merrill, J. van Merrienboer, and M. P. Driscoll, editors, *Handbook of Research on Educational Communications and Technology*, pages 345–353. Taylor and Francis, New York, NY, USA, third edition, 2007a.

D. A. Wiley. On the sustainability of open educational resource initiatives in higher education. Technical report, Organisation for Economic Co-operation and Development, 2007b. http://www.oecd.org/dataoecd/33/9/38645447.pdf Accessed: 17 February 2009.

D. A. Wiley, A. Gibbons, and M. M. Recker. A reformulation of learning object granularity. 2000. http://reusability.org/granularity.pdf Accessed: 10 November 2010.

I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, USA, second edition, June 2005.

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314, Melbourne, Australia, 1998. ACM.